THE APPLICATION OF MACHINE LEARNING, BIG DATA TECHNIQUES, AND CRIMINOLOGY TO THE ANALYSIS OF RACIST TWEETS.

by

Ed Day

Canterbury Christ Church University

Thesis submitted
for the degree of Doctor of Philosophy

2018

I, Ed Day, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Word Count: 85,994

Abstract

Racist tweets are ubiquitous on Twitter. This thesis aims to explore the creation of an automated system to identify tweets and tweeters, and at the same time gain a theoretical understanding of the tweets. To do this a mixed methods approach was employed: machine learning was utilised to identify racist tweets and tweeters, and grounded theory and other qualitative techniques were used to gain an understanding of the tweets' content.

84 million tweets that all contained racist words were collected from Twitter. 84,000 of these were hand annotated as racist or not.

The machine learning was performed in a Hadoop cluster, utilising Spark and Hive. To identify racist tweets, systematic comparison of seven different algorithms, and a large number of textual, user-derived and geographical features was performed. New features: time of day and day of week were also evaluated. The 84,000 hand annotated tweets were used as input to the machine learning supervised classification processes. It was found that the combination of support vector machines with hour of day as additional feature was optimal for accuracy (0.93) and AUPRC (0.86).

A qualitative exploration of tweets was also performed, including a grounded theory analysis.

A novel machine learning system to identify racist accounts was created using metrics from the racist tweets, concepts from the grounded theory and a combination of the two as feature inputs. All three sets of features gave accuracy of at least 0.82.

The ambiguity of the tweets meant they were difficult to classify, for both humans and machines, as to whether the tweeter's intentions were racist or not, the word 'nigga' being particularly problematic.

Grounded theory analysis of the tweets showed extremely narrow rhetoric that could be summarised in a single theoretical concept: the defence of the in-group.

Acknowledgements

Thanks to Abhaya Induruwa and Robin Bryant for their excellent and patient supervision and input.

Thanks to Sarah for all her help, and for putting up with this for all these years.

Thanks to Harry, Ruby and Mackie for being great kids.

Thanks to those who helped with the onerous task of reading thousands of racist tweets: Mike Hewitt, Annalise Ralph, Ant Murray, Jazz Budds and Gayle Lennox.

Contents

List of Figures

List of Tables

List of Abbreviations

ANN	Artificial Neural Network
AUPRC	Area under the Precision Recall Curve
AUROC	Area Under the Receiver Operating characteristics Curve
BOW	Bag of Words
CSEW	The Crime Survey for England and Wales
DT	Decision Tree
FN	False Negative
FP	False Positive
GBT	Gradient Boosted Tree
HDFS	Hadoop Distributed File System
MAMA	Measuring Anti-Muslim Attacks project
N5	Set of ngrams from length 1 to 5
NB	Naive Bayes
NLP	Natural Language Processing
PCA	Principal Component Analysis
RAT	Routine Activity Theory
RF	Random Forest

ROC	Receiver Operating Characteristics
SVM	Support Vector Machine
SA	Sentiment Analysis
TN	True Negative
TP	True Positive
WEKA	Waikato Environment for Knowledge Analysis

List of Code

Chapter 1

Introduction

This thesis aims to apply novel machine learning and big data techniques to the automated identification of racist tweets and tweeters. Racist tweets are problematic for a number of reasons: there are a form of cyber violence towards an individual or group, they can be unpleasant for nontargeted individuals to read, and they can be influential as tools for recruiting people to racist groups, or for providing solidarity within a group.

The racist language seen in such tweets is a subset of hateful language[1], which is a common problem on the internet (Waseem and Hovy, 2016). In the UK the balance of freedom of expression versus the prohibition of hateful language is weighted towards the proscribing of racist language and language that discriminates with respect to group membership based on beliefs or characteristics [2]. Thus, in the UK at least, racist language can be thought of as a cybercrime[3], that is a form of crime mediated by the internet. Cybercrime in England and Wales is extremely prevalent with an estimated 7.6 million offences in 2015 [4]. These offences cover a wide array of crimes. For a more focused view

[1] Hateful language often includes words that many would find offensive. The denigration of minorities, for example by using racial epithets, is ubiquitous in such language, as is the use of swear words. Any research into such language requires the discussion of offensive words and phrases, and therefore caution should be used when reading this thesis, since offensive terms are used uncensored.

[2] For a discussion on this see Chapter 3.

[3] This is discussed in Chapter 4.

[4] The Crime Survey for England and Wales (CSEW) is an annual self-report survey aimed at measure households' experience of crime victimisation. Around 50,000 addresses are randomly selected to complete

of online racism, the Home Office statistics for online hate crime can be examined. The Home Office monitors five strands of hate crime: race or ethnicity, religion or beliefs, sexual orientation, disability and transgender identity. It publishes an annual report: 'Hate Crime, England and Wales' that reports hate crime statistics for England and Wales, which was last published in October 2017 (O'Neill, 2017). This report is largely based on police recorded crime statistics.[5] The hate crime data that are recorded by the police are 'any criminal offence which is perceived, by the victim or any other person, to be motivated by hostility or prejudice towards someone based on a personal characteristic.' (ibid., p.2). In April 2015 it became a requirement for police forces to flag any cases in which it was thought 'that an offence was committed, in full or in part, through a computer, computer network or other computer-enabled device' (ibid., p.18). The 2016/17 report was the first to include statistical analysis of the number of online hate crimes.[6] From the data it was determined that there were a total of 1,067 online hate crimes, of which 671 (63%) were deemed to be part of the race hate crime strand. Despite race being by far the largest number of the online hate crimes, race had the lowest proportion of crimes flagged as online (79%), when all hate crime is considered (both online and offline). The majority (84%) of the online hate crimes were violence against the person, with 13% public order offences. Of the violence against the person offences 92% were classed as harassment, with only 4% classed as 'racially or religiously aggravated harassment'. These data are somewhat surprising, in that they suggest online hate crimes are rare, with only 671 in a year, making their prevalence at a comparable level with murder. This is in conflict with the high levels of racist tweets that can be seen on Twitter. One explanation for this is that there is a long journey between someone receiving a racist tweet and this act being

a survey (Kantar, 2015). Reports are produced from the CSEW data and the overall number of offences is often reported as a trend. For example in the 2015 CSEW data showed that in the year ending June 2015 there were less than 7 million offences, which was a significant reduction from the peak figure of 19 million which was in 1995 (Office for National Statistics, 2015).

Prior to 2015 cyber offences were not included in the data of the CSEW's forerunner, the British Crime Survey. From May to August 2015 a field trial was performed to look at the scale of cybercrime. This trial produced estimates of 5.1 million incidents of online fraud, and 2.5 million incidents of crime falling under the Computer Misuse Act (ibid.). If these 7.6 million cyber offences are included in the overall number of offences then this more than doubles, starkly illustrating the impact of cybercrime.

[5] The report does include crime survey data, but due to the scarcity of hate crime data within the crime survey, it only includes these data every three years, the last such inclusion being in 2014/15, prior to the inclusion of online offences in the The Crime Survey for England and Wales (CSEW, formerly the British Crime Survey).

[6] These online hate crime statistics are deemed to be 'experimental' due to the poor quality of the data, caused by, for example, the likelihood that forces are under-using the online flag. Only 23 out of the 44 forces in England and Wales provided data of sufficient quality to be used in the report.

counted in police recorded crime statistics.[7] Despite the police recorded figures being so low, the data from this research, and that of others, suggests that racism is undoubtedly an issue on Twitter.

There is some evidence that racism on Twitter peaks as a result of triggering events[8] such as the murder of Fusilier Lee Rigby in 2013[9] (Awan and Zempi, 2017; Williams and Burnap, 2015). It is relatively easy to identify vitriolic tweets using hashtags related to triggering events (Burnap et al., 2014), and this is an important task, however it is equally important and perhaps harder to identify and 'police' the everyday racist content encountered on Twitter. This research is focused on these 'mundane' levels of racism that occur daily on Twitter.

In studying racism on Twitter, itis first necessary to decide on what meaning can be given to the content of a tweet. The understanding of a tweet's meaning is a highly complex and difficult task, as there are many complexities to the understanding of language in any medium. These include, for example, 'sarcasm' which is relatively easily understood by humans but is a difficult construct for machine learning and Natural Language Processing (NLP) systems to handle (Riloff et al., 2013). Progress has been made in the automated understanding of tweets but, the existing solutions for the automated detection of hateful language still need improving (Pitsilis et al., 2018). Indeed Mondal

[7] The police recorded crime figures may be low due to one or both of under-reporting and under-recording such incidents. Under-reporting occurs when witnesses of an incident determine that it is not worth their time to report the incident to the police. This may be due to a number of factors including embarrassment, not wanting to 'get into trouble', not being aware of being a victim of crime, and believing that the police will not take any action. The last reason seems justified in an online environment (at least with respect to fraud) since the government's online fraud reporting tool suggests that no action will be taken. `actionfraud.police.uk` is the National Fraud Intelligence Bureau (NFIB)'s online reporting tool, and its website states, 'although the police cannot investigate every report individually, the information you provide will aid them'.

If a crime *is* reported then the police need to make a decision as to whether it will be *recorded* or not. The police make this decision based on whether the incident is a crime under the law, following rules set out in the National Crime Recording Standards and Home Office Counting Rules. This is by no means a simple decision and it may also be influenced by other factors such as political interest and intellectual bias (Bryant and Bryant, 2014).

[8] Indeed there is evidence that both on *and offline* crime peak as a result of triggering events (Burnap et al., 2014).

[9] Lee Rigby was fatally attacked by two men on a street in Woolwich, London on May 22, 2013. The two men later found guilty of his murder, Michael Adebolajo and Michael Adebowale, initially ran him down then attempted to behead him. He was targeted because of the Help for Heroes sweatshirt he was wearing. Help for Heroes is a charity that raises money in the UK for war veterans, and Adebolajo and Adebowale stated that the murder was a revenge attack, since they believed that Muslims were regularly being killed by British soldiers (McEnery et al., 2015).

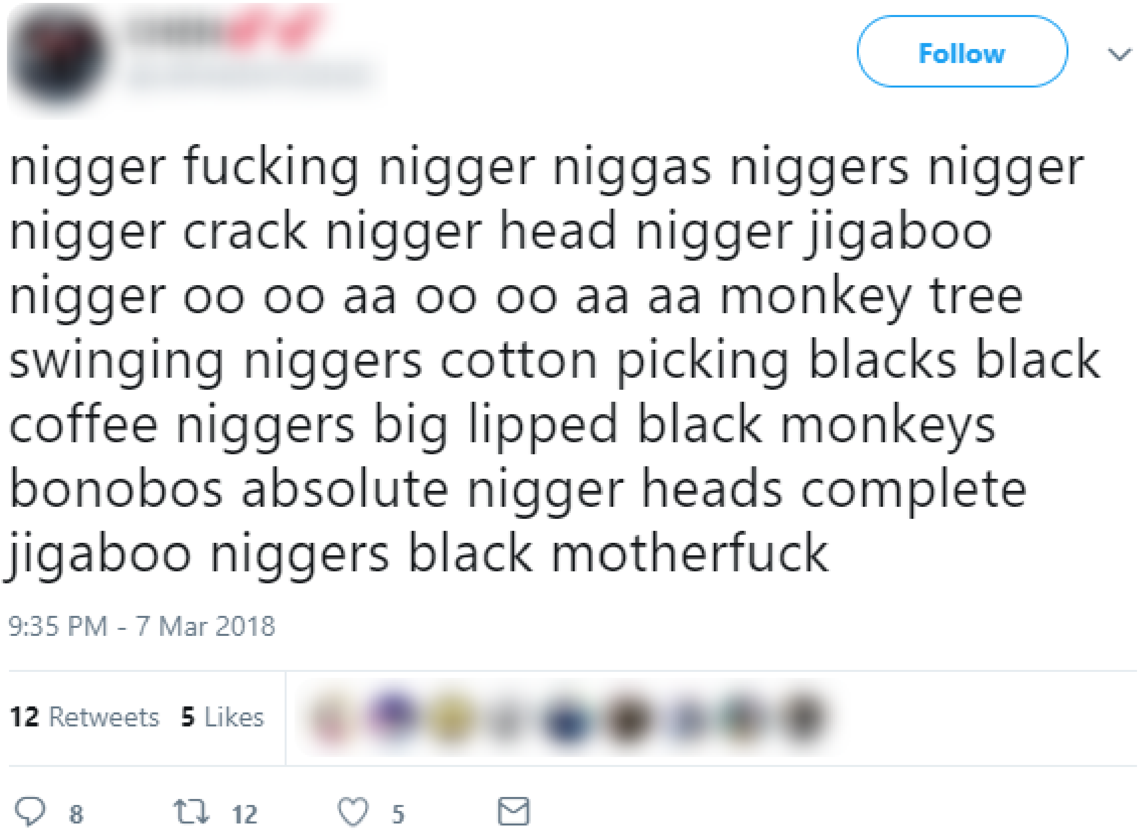
et al. (2017, p.86) note that 'it is safe to say that computational methods to detect hate speech currently are in a nascent stage.'

Latterly Twitter itself has been addressing the issue of racist language in its network. In March 2017 Twitter announced that it had introduced new algorithmic approaches to idenity hateful language, although it did not give any details of how these approaches work, other than suggesting they target user behaviour rather than keywords (Perez, 2017). Twitter's new filters have been criticised for poorly identifying racist tweets and for their false negatives, for example, the suspension of actress Rose McGowan's Twitter account (Lomas, 2017). Partly in response to the preceived paucity of these efforts, in September 2017 the European Commission announced that it would monitor the progress of internet entities in tackling illegal online content, and determine whether further measures were necessary, including additional regulation. At the beginning of March 2018 the European Commission recommended

> a set of operational measures - accompanied by the necessary safeguards - to be taken by companies and Member States to further step up this work before it determines whether it will be necessary to propose legislation. These recommendations apply to all forms of illegal content ranging from terrorist content, incitement to hatred and violence, child sexual abuse material, counterfeit products and copyright infringement (European Commission, 2018).

Partly as a consequence of its attempts to combat hateful language, Twitter was censured by the European Commission for deleting user data without informing users, and also changing its terms of service without informing them (dw.com, 2018).

This addition of hateful language filtering, is seen as a major change to the way Twitter is viewed, indeed it has been argued that Twitter 'no longer wants to be Twitter', that is Twitter wants to suppress the anarchic nature of its conversations (Feldman, 2017). As a consequence, in 2018 Twitter announced a request for research proposals into its 'conversational health' (Twitter, 2018). However at the time of writing, March 2018, it is still possible to see tweets on Twitter of the kind illustrated in Figure 1.1, which shows an extremely abusive tweet aimed at black people. There was a continuation of the conversation, when the original poster replied to a tweeter who took offence at his original post, with a further stream of antiblack invective. Another user replied to the conversation saying ‘jack and twitter isn’t going to like this’, presumably noting the fact that Twitter is now, supposedly, policing such rhetoric.



[redacted]

nigger fucking nigger niggas niggers nigger nigger crack nigger head nigger jigaboo nigger oo oo aa oo oo aa aa monkey tree swinging niggers cotton picking blacks black coffee niggers big lipped black monkeys bonobos absolute nigger heads complete jigaboo niggers black motherfuck

9:35 PM - 7 Mar 2018

12 Retweets 5 Likes

Figure 1.1: An example tweet from March 2018.

As well as being undesirable in itself, the presence of online hateful language might be related to offline crime. Whether there is a correlation between online racism and offline antisocial or even criminal behaviour, is unclear. Statistics from the MAMA (Measuring Anti-Muslim Attacks) project show a fluctuating pattern of anti-Muslim abuse with respect to its online/offline nature. MAMA is an independent UK government-funded service that allows for the reporting of anti-Muslim incidents via its portal, Tell MAMA. The MAMA project records the experiences of Muslims in the UK who were subjected to both online and physical world anti-Muslim hate, and analyses these data and disseminates their analyses via an annual report (Faith Matters, 2018). In the 2015 report Tell MAMA recorded 729 incidents in total and, of these, 548 were verified as anti-Muslim incidents.

Of these, the majority were online, totalling 402 out of the 548 verified incidents (Littler and Feldman, 2015). In the most recent report, detailing incidents in 2016, there were now more reports, 1,223, of which 953 were verified. Of these verified reports 642 were 'offline' and 311 were online, a significant annual reduction in both the total number of reports and the percentage of the total that were classified as *online* anti-Muslim incidents (Faith Matters, 2017).

Whether there is a correlation or not, there are influential voices in the UK suggesting that there is indeed such a link, and these commentators sometimes go further and argue that it is a causal relationship. For example, Commissioner of the Metropolitan Police, Cressida Dick argued that social media sites 'rev people up' and make street violence more likely (BBC, 2018).

1.1 Research Aim

Whether there is a correlation or even a causal link between online and off-line crime, racist tweets are indeed a problem on Twitter. They may be criminal acts or they may merely be offensive, but either way their identification is desirable and their ubiquity means an automated solution is needed to identify them.

The aim of this thesis is to assist in the problems of identifying racist tweets and tweeters. To do so a mixed methods approach, which applies techniques of both machine learning and grounded theory, was utilised.

Machine learning allows for the analysis of datasets, such as the Twitter data in this research, that are too large for humans to analyse feasibly. Using computing and statistics it enables the detection of patterns invisible to humans, and aids in the reduction of subjectivity in the analysis of data.

In addition to this quantitative approach a qualitative analysis of the tweets was performed using grounded theory, in which concepts and categories emerge only from the tweets' textual data (Strauss and Corbin, 1990). Grounded theory provides a richer notion of the meanings within the tweets, uncovering themes that are ignored by the numerical analysis.

To address these research areas a System for the Automated Detection of Racist Tweets and Accounts (SADRTA) was created. This system comprised:

- Data collection software – python programs to harvest tweets from Twitter and 'clean' the captured data, deleting and malformed records. The data was stored as text files.
- HiveQL scripts which took the resultant text files and converted them to Hive tables stored in a HortonWorks HDP 2.6 Hadoop cluster.
- A series of Spark machine learning routines written in Scala and runnable on the Hadoop cluster. These routines produced models based on features derived from the tweets' text and metadata and 84,000 labelled tweets as input. The models were used to predict the remaining unlabelled 84 million tweets.
- HiveQL scripts which summarised the predicted tweets creating metrics such as retweet count.
- A grounded theory analysis of a sample of the predicted racist tweets was performed.
- Spark Scala machine learning routines which took the output of 4 and 5 above and created models to predict racist accounts.

Using this mixed methods approach and SADRTA a series of research questions in this area were addressed. These questions arose from a review of the literature, and discussions with law enforcement and others. The questions are given in the next section.

1.2 Research Questions

The research questions were addressed by this thesis are the following:

- Question (Q1): is it possible to have an efficient, accurate and reliable automated racist tweet identifier?

  - Sub Question (Q1-1): how can reliability, efficiency and accuracy be measured?
  - Q1-2: what are the current approaches to automated racist tweet identification, and can they be improved upon?
  - Q1-3: what data do the current approaches to automated racist tweet identification use as input, and are there any other possibilities?

- Q2: if such an identifier can be created, how can this further the ability to understand and identify influential racist Twitter accounts?
  - Q2-1: how should influential be defined?
  - Q2-2: what are the current approaches to identifying influential accounts, and can these be applied to racist Twitter accounts?

- Q3: Can qualitative analysis, criminology and psychology be used to further the understanding of racist tweets?
  - Q3-1: what themes emerge from a qualitative analysis of racist tweets?
  - Q3-2: can criminological and psychological theories, such as Routine Activity Theory (RAT), be applied to racist accounts and tweeters?

These questions were explored using two datasets of tweets containing racist terms, the first comprised 84 million tweets collected over 79 days in the summer of 2016, the second contained 28 million tweets collected over 34 days in the summer of 2017. A sample of tweets was annotated and used as input to machine learning algorithms to identify racist tweets in the rest of the first dataset. These predicted racist tweets were used in a grounded theory analysis, the results of which were used in another machine learning algorithm to predict which accounts seem to be producing racist tweets, and thus could be thought of as 'racist accounts'. These results were compared with those of a third machine learning algorithm which aim to predict racist accounts based on metrics of account activity.

1.3 Structure of the Thesis

The next chapter, Chapter 2, contains a review of other work in the field of the automated detection of racist language.

Chapter 3 presents a discussion of racism on Twitter. An introduction to Twitter is given followed by various conceptualisations of ‘race’, ‘racism’ and ‘ethnicity’ and how they can be used to identify and understand the content of the tweets. In addition an analysis of what is meant by ‘racist speech’ is given. Finally the implications these ideas have on the methodology of this research are discussed, included in this is the difficulty of determining the meaning and intent of a racist tweet, and how the word ‘nigga’ is particularly problematic.

Chapter 4 introduces the key theoretical perspectives that underpin the research. A criminological theoretical framework will be used. Such a framework must first consider what is meant by cybercrime and this is discussed in the next section. Then a brief summary of criminological theory is given followed by a discussion of the differences between situational and dispositional criminology. Within criminology, a particularly useful perspective is that of Routine Activity Theory (RAT), which is discussed in detail along with the similar Lifestyle Exposure Theory (LET). Then RAT’s applicability to cybercrime is discussed. Psychological perspectives are also discussed as they can shed light on why people may sent racist tweets.

Chapter 5 explores the use of mixed methods, that applying a combination of quantitative (ML) and qualitative (grounded theory) methods to answer the research questions posed. First a theoretical discussion of mixed methods is given. Then the use of machine learning is explored, including an explanation of the difference between supervised and unsupervised learning. Then the application of machine learning to textual data is discussed, including a section on how the data is preprocessed prior to input into an ML algorithm, along with the techniques employed to evaluate the output of these classifiers. Finally the qualitative analysis of the tweets is discussed, including various techniques, the main one being that of grounded theory.

Chapter 6 provides a discussion of the relationship between big data and criminology and gives further details on the theory of machine learning. There are sections on *features*, that is the dimensions of the data used as input to prediction algorithms. Text, user, geographical and temporal features are discussed. Then Spark extraction and transformation methods are explored. This is followed by section on the different algorithms utilised in this research.

Chapter 7 presents results for the machine learning procedures and qualitative analysis. For the machine learning results first the effects of varying oversampling fraction of negatives are given. This is followed by a discussion of text, user, geographical and temporal features and their efficacy as input to the machine learning algorithms. Features are further considered in both discussing whether text should be treated as Ngrams or BOW and whether hour, hour+day or neither should be used as additional features. Then seven different algorithms are compared by analysing their metrics.

The temporal aspects of the tweets are further discussed in relation to the different datasets including both the input and predicted data. Then the qualitative results are discussed, first summary data from an NLTK analysis is given, followed by analysis of word clouds and word trees of the data. Then a brief discursive analysis is given followed by a grounded theory analysis of the data.

The chapter concludes with a discussion of accounts and the results of machine learning processes aimed at predicting which accounts are racist, using metrics from the accounts, from the grounded theory analysis and a combination of these.

The final chapter discusses the results in light of the theoretical framework and research questions.

Chapter 2

Literature Review

The chapter contains a review of existing work on the automatic detection of hate speech on Twitter.

2.1 Machine Learning and Hate Speech Detection

Greevy and Smeaton (2004) used Support Vector Machines (SVM) to automatically classify webpages as racist or not. They represented the text of each page as both Bag Of Words (BOW) and bigrams. They justify the use of SVMs by noting their popularity and efficacy, in particular their avoidance of overfitting and their applicability to textual analysis. However they do not compare SVMs with any other type of algorithm. They used a corpus of 3 million words which was split into an equal number of racist and non-racist documents although they give no details on what determines a document as being racist or nonracist. They found that the BOW representation was more accurate and had higher recall than the bigram one, but the bigrams had higher precision[1].

Razavi et al. (2010) examined the automatic detection of messages regarded to be

[1] Precision is the ratio of true positives to all those classified as positives. See Section 5.12.2.2 for more details.

flames which are messages that include one or more of hate speech, slurs, extremism, crude language disguised offensive terms, ‘four letter words’, provocative language, taboos and unrefined or squalid language. Their data was log files and USENET newsgroup messages pre-annotated by other researchers. In total they analysed 1,525 messages, 487 (32%) of which were flames. They created an ‘Insulting and Abusing Language Dictionary (IALD)’ containing 2,700 words, phrases and expressions. They performed three-level classification, initially using Naive Bayes (NB) to select the most discriminating features. They then used another NB classifier using new labelled sentences, and then a third NB classifier using the IALD. However, they did not compare their Naive Bayes classifier with any other models. They performed tenfold classification and achieved high levels of accuracy and precision, but they did not compare the results using other algorithms.

Sood et al. (2012b) focused on automated insult detection in comments on the social news site, Yahoo! Buzz. They collected 1,655,131 comments between March 2010 and May 2010 and they annotated 6,500 comments from these data using Amazon’s Mechanical Turk workers. They had some concern about the quality of the Mechanical Turk data and so required a consensus of three raters, however they do not give any details of checking of the consensus ratings. They used SVM with tenfold cross validation and achieved accuracy[2] of 0.8345 and F-score[3] of 0.5432. They did not compare SVM against any other algorithms.

Warner and Hirschberg (2012) used data from websites that contained hate speech along with thousands of comments flagged as hate speech by users of Yahoo! groups. They then annotated around 9,000 paragraphs from the websites into seven categories of hate: anti-semitic, anti-black, anti-asian, anti-woman, anti-muslim, anti-immigrant and other-hate. They rated interrater reliability for the anti-Semitic versus other paragraphs and this gave a Fleiss’ Kappa[4] $\kappa = 0.63$. They used a template-based strategy to determine their features which they compared with a baseline unigram feature set which were used as input to a SVM algorithm with tenfold cross validation. They found that the unigram feature set performed best. They also did not compare SVM against any other algorithms.

[2] Accuracy is the percentage of predictions that are correct. See Section 5.12.2.1 for more details.

[3] F-score is the harmonic mean of precision and recall. See Section 5.12.2.6 for more details.

[4] Fleiss’ Kappa is a statistic that quantifies agreement between multiple raters of categorical or binary data (Fleiss, 1971).

Chen et al. (2012) used lexical syntactic features (LSF) of sentences to try to detect offensive language. They compared BOW, bigrams, trigrams and 5grams, dependency sets and LSF as features sets. They used unsupervised classification using tenfold cross validation with NB and SVM algorithms but did not consider other algorithms. They found SVM worked best as did their LSF features. Their algorithms detected the very broad category of 'offensiveness' and provide little discussion of the very subjective nature of offensiveness.

Dadvar et al. (2012) explored the use of gender as a feature in machine learning classification of cyber bullying, since research shows males and females use language to bully in different ways. The data were posts from MySpace, numbering 381,000 posts, of which the gender of the author was known. From these 2,200 posts were annotated by three annotators as either harassing or non-harassing. They used four types of features: profane words, second person pronouns, other personal pronouns and Term Frequency-Inverse Document Frequency (TF-IDF) values. They selected these based on the work of Yin et al. (2009). They used tenfold cross validation and an SVM algorithm implemented in WEKA. They did not compare SVM with any other classifiers.They experimented with all the annotated posts and compare this with just the male posts and just the female posts. They found the gender specific classifiers performed better although all had poor scores compared with the literature in general.

Dadvar et al. (2013) analysed the effect of user context in the automated detection of cyberbullying. They collected 4,626 comments from YouTube which were annotated as bullying and non-bullying, 9.7% of which were bullying. They created three different feature sets as input into their machine learning algorithm: features based on content, features based on cyber bullying and features based on user information. The content-based features contained: number of profane words, number of first and second person pronouns, profanity windows (whether profanity follows a second person pronoun within the size of the window), the number of emoticons and the ratio of capital letters (this was used as an indicator of shouting within the comment). The cyber bullying-based features included a number of cyber bullying words and the length of the comment. The user-based features included the history of users activities and the age of the users. They used SVM and tenfold cross validation and compared content-based features versus cyber bullying features versus user-based features. They found user-based features to give the best precision, recall and F-score. They did not compare the use of other algorithms.

Bartlett et al. (2014) investigated racial and ethnic slurs on Twitter. They collected 126,975 publicly available English-language tweets that contained a slur from the ethnic slurs list in Wikipedia, during the period from 19th to 27th November 2012. They found the most prevalent terms were: ‘white boy’, ‘paki’, ‘whitey’, ‘pikey’, ‘nigga’, ‘spic’, ‘crow’, ‘squinty’ and ‘wigga’. They manually annotated around 200 to 500 tweets and used them as input into machine learning classifiers, although they give no details of what these classifiers were.

Hosseinmardi et al. (2015) investigated cyber bullying on the Instagram network. They snowball sampled[5] Instagram users annotating 998 of what they termed *media sessions* as cyberbullying or not, using the not necessarily reliable crowdsourcing platform Crowdflower. As features they used text of comments and meta data of users and images and these were input into NB, SVM and SVM with Principal Component Analysis (PCA) of features keeping the first 20 components. They found the SVM with PCA classifier performed optimally.[6]

Gitari et al. (2015) analysed a sample of paragraphs from blog postings from each of ten hate sites totalling 180 paragraphs and paragraphs of text related to the Israel-Palestinian conflict totalling 320 paragraphs. This data was rated by three researchers as not-hateful, weakly hateful and strongly hateful. Their percentage agreement was 68% with Kappa $\kappa = 0.45$. They settled disagreement using a majority vote. They used a rule-based approach to classify sentences using a combination of sentiment lexicons and ‘hate’ verbs generated from their data. They created an algorithm to determine the level of hate of text using rules and their lexicon. They did not perform any ML. Their method performed favourable in terms of F-score compared with Kwok and Wang (2013), and

[5] Snowball sampling is a sampling method that gains participants by referrals from existing participants (Biernacki and Waldorf, 1981).

[6] They found that a naive Bayes classifier improved accuracy to 0.72, from a baseline 0.52. Using an SVM classifier with PCA reduction to 20 components increased accuracy to 0.87. Precision and recall figures both showed similar gains.

Warner and Hirschberg (2012)'s NB models.

Dinakar et al. (2012) investigated automated cyber bullying detection using a common sense knowledge base, *BullySpace*. They obtained data from YouTube comments on videos discussing sensitive issues and also nonsensitive issues, and 1,500 of these were annotated by three annotators, for: negative remarks based on gender or sexual orientation, negative remarks based on race or ethnicity and negative remarks based on intelligence. They used four classifiers: NB, rule-based learner, tree-based learner and SVM. They found the rule-based learner had the best accuracy, but the SVM had the highest F-score.

Xiang et al. (2012) collected a total of over 696 million tweets which were used, along with a small dictionary of offensive terms, to bootstrap a set of training tweets, which were then used as input to a semi-supervised machine learning approach. Four algorithms were used: DT, SVM, LR and RF and these were run with tenfold cross validation. ANN was not compared and nor were other feature sets. A random sample of 4,029 tweets were coded by the researchers for evaluation. They found the LR algorithm to be the most successful and their lexicon-based technique outperforming a classifier without a lexicon, as shown by its AUROC being higher.

Xu et al. (2012) used data from Yotube comments from over 2 million users. They auto corrected the spelling and grammar of these comments. The corrected sentences were then rated for offensiveness using a lexicon and syntactic features. They used SVM and NB machine learning models with tenfold cross validation and found their model performed well with precision of 77.9% and recall of 77.8%. No other algorithms and features were compared.

Kwok and Wang (2013) looked at racist tweets directed against Blacks. They used a NB classifier to automatically classify tweets as racist or not. To build a training dataset they use samples of tweets from known racist Twitter accounts, which after data pre-processing left them with 24,582 tweets. Their pre-processing consisted of the removal of URLs, mentions, stopwords and punctuation; conversion to lowercase and spelling correction of misspelled slurs. They used a BOW model. Their method produced an accuracy of 76% (Kwok and Wang, 2013). No other algorithms and features were compared.

Burnap and Williams (2014) investigated how hate speech manifests and diffuses in social media in relation to a triggering hate crime event. They collected what they term as 'Big Data': tweets from a two-week period following the murder of Lee Rigby in Woolwich, UK in May 2013. They collected tweets containing either the word 'woolwich' or the hashtag '#woolwich', collecting 450,000 tweets. They randomly sampled 2,000 tweets which were annotated for hate speech using CrowdFlower, requiring at least four annotators per tweet, and only keeping those tweets where there was agreement among at least three annotators. This left 1,901 tweets which contained 222 examples of hate speech. They compared a limited set of features: ngrams of words with lengths 1 to 5, hateful terms from Wikipedia and ngram typed dependencies (which represent grammar relationships and can be used to capture the use of othering in language) and also combinations of these. They use tenfold cross validation and NB, LR, SVM and Random Forest Decision Tree (RFDT) algorithms. There was very little difference between these classifiers and they performed best with the typed dependencies although the results for those were almost the same as for using hateful terms alone.

Djuric et al. (2015) used a two-step method for hate speech detection, using paragraph2vec[7] to transform paragraphs into a feature set, which they compared with BOW for both TF and TF-IDF encodings. They used comments from the Yahoo Finance website with 195,456 non-hate comments and 56,280 comments that contain hate speech. They performed fivefold cross validation using a LR algorithm. They found that paragraph2vec worked best and, interestingly, TF worked better than TF-IDF.

Burnap et al. (2015) performed an analysis of Twitter data in their investigation into potential racist tweets relating to a Premier League football match. Tweets were collected one month prior to and succeeding the occasion in which a Manchester United player, Patrice Evra alleged a Liverpool player, Lewis Suarez, had racially abused him. Tweets were collected by searching for the keyword 'Suarez'. Police officers coded 1,022

[7] Paragraph2vec is an alternative to BOW. Both algorithms reduce text to fixed-length feature vectors, with BOW words are treated as equally distant to one another, whereas Paragraph2vec predicts surrounding words by sampling contexts from the paragraph (Le and Mikolov, 2014).

tweets for levels of tension on a numerical scale. The level of agreement of the officers' annotation was rated with Krippendorf's $\alpha = 0.67$. To analyse their data they created a tension detection algorithm which consisted of a series of decisions to determine the tension level of a text. They compared this algorithm with a machine learning approach, using a multinomial NB classifier and a SVM classifier and also with a sentiment analysis classifier, SentiStrength, which rates short texts on a scale from extremely negative to extremely positive, but no other algorithms were compared. Their results were mixed, for example for the tension class the precision of the NB classifier was better than the tension engine and so was the SVM classifier for the high tension class, but with a much lower recall.

Tulkens et al. (2016) performed experiments aimed at automatically detecting racist discourse in Dutch language social media. They analysed posts from two Belgian social media pages that attracted large number of racist comments: a noted anti-Islamic organisation's community hub and one that 'was used to post articles by a well-known right-wing organization' (ibid., p.7). From both these pages they scraped the first 100 posts which produced 5,759 comments, which were annotated by two annotators with tie-breaks decided by a third annotator. They repeated this procedure for an independent test set, for which they retrieved 616 comments. There were three possible annotations, racist, nonracist and invalid. Interrater reliability was measured using Kappa $\kappa = 0.60$ for the first two annotators on the training corpus, and $\kappa = 0.54$ for all three annotators for the first 25% of comments. The researchers suggested that this low value of interrater reliability is probably due to the misinterpretation of certain Flemish words and the different cultural backgrounds of the annotators. For features they used ngrams, stylistic features including word length, average sentence length and vocabulary richness, and dictionaries containing racist terms, finding trigrams to perform the best. They performed tenfold cross validation using an SVM algorithm implemented using scikit-learn. All of the models performed fairly poorly compared with the literature in general. They then repeated their experiment with a more neutral (i.e. less likely to attract racist comments) set of data from a Belgian newspaper website which they annotated as before. The interrater reliability for this second set of annotations was better $\kappa = 0.7$ and 0.57 which the researchers surmised could be due to a training effect. The classifier performed worse with the neutral data than the racist data.

Mehdad and Tetreault (2016) use the same dataset as Djuric et al. (2015) to explore whether character ngrams are more beneficial than word ngrams as features. They used three classifiers: a distributional representation of comments, a recurrent neural network language model and a SVM with NB features model. For most measures the NB/SVM model work best and the results were mixed with respect to words versus characters.

Waseem and Hovy (2016) annotated 16,914 tweets collected over two months, 3,383 of which had sexist content and 1,972 which had racist content. They performed tenfold cross validation using LR and bigrams to 4grams as features as a baseline. They compared this baseline against a feature set with the addition of gender of the tweeter and another with gender and location. There was very little difference in performance with the addition of gender or gender and location, but no other algorithms were compared.

Burnap and Williams (2016) studied 'cyber hate' that is hateful and antagonistic content on the Internet. They selected data from Twitter for a short time period following certain 'trigger' events. The justification for this was that they wanted to collect a large number of tweets containing cyber hate speech. Although this is a sensible strategy from the viewpoint of collecting lots of these kinds of tweets, it is unlikely to lead to representative Twitter streams, and so any classifier that is trained on this data may not perform well on 'normal' Twitter data. In addition, even these targeted analyses have large amounts of 'noise' related to similar but distinct incidents (Roberts et al., 2017). The events they used were Barack Obama's re-election in November 2012, Jason Collins' announcement that he was gay in April 2013, and the opening of the 2012 Paralympic games. They collected data for two weeks after these events. From the collected tweets they randomly sampled 2,000, which they uploaded to the crowdsourcing website CrowdFlower. Then at least four human annotators rated tweets as to whether they were hateful or not, and from these the researchers rejected any with less than 75% agreement. These produced 1,803 tweets with 183 of them hateful for sexual orientation, for race there were 1,876 tweets with 70 of them hateful and for disability 1,914 tweets containing 51 hateful ones. As in their previous research (Burnap and Williams, 2014), they compared the following features: ngrams of words with lengths 1 to 5, hateful terms from the Wikipedia list of ethnic slurs and ngram typed dependencies (which represent grammar relationships and can be used to capture the use of othering in language) and also combinations of these. They used tenfold cross validation and SVM and Random Forest Decision Tree (RFDT) algorithms. The SVM outperformed the RFDT, and the use of a linear kernel outperformed RBF and polynomial kernels. They found that the combination of all three feature sets performed best with respect to race for recall and F-score, but precision was best for the hateful terms alone. Typed dependencies did not provide significantly better results than the hateful terms alone. They did not investigate other algorithms such as ANN. Hasanuzzaman et al. (2017) utilised demographic information of tweeters. They used a word embeddings methodology with age, gender and location embeddings. Their data consisted of 17.2 million tweets selected by those containing racist keywords along with 41.3 million tweets that did not contain any racist keywords. They predicted gender using Sap et al.'s (2014) text-based models[8] and location using Chen and Neill's (2014) rules.[9] Their word embeddings follow the method outlined by Bamman et al. (2014). They performed tenfold cross validation with a SVM model using the WEKA platform, although they did not compare this against any other algorithms. They found that the demographic variables improved model metrics for bigram, trigram and word2vec feature sets but marginally worsened the outcomes for ngrams (1 to 4grams). They used the crowdsourcing platform CrowdFlower to obtain annotated tweets. These were judged as racist, not racist, or unsure by a minimum of 3 annotators. Of 8,000 tweets 7,358 had a majority vote that was not unsure, and of these 3,267 (44%) were considered to be racist, a high percentage.

Mondal et al.'s (2017) aim was to understand rather than predict online hate speech. Their data were from two online social networks: Whisper (a social network that allows only anonymous posts) and Twitter. The data were all collected between June 2014 and June 2015. They downloaded 48.97 million Whisper posts[10]. They reduced these posts to

[8] Sap et al. (2014) used Facebook and Twitter data to create lexica of words and their weights using regression and classification models. The lexica include demographic information such as gender.

[9] To determine location they used the first location source found from three sources in the following order: searching for a location mention in the text message, then verifying if the user enabled the geo-coding function of their device, and, finally looking for location information from a user's profile (Hasanuzzaman et al., 2017, p. 929).

[10] A post on Whisper contains amongst other things text, location, timestamp, the number of favourites,

27.55 million which were English-only and contained an exact location. They downloaded all publicly available tweets during this time period, giving 1.6 billion tweets of which only 1.67% had location information. When reduced to English-only this left 512 million tweets. Eschewing the use of hateful keywords they attempted to identify hate posts by using a sentence template thus: I *<intensity><user intent><hate target>*, where *<intensity>* is a modifier that may amplify emotions such as 'really', *< user intent>* would be one of the synonyms for hate, and *<hate target>* includes group identifiers such as 'white people'. This strategy produced 20,305 tweets and 7,604 whispers that contained hate speech. They noted that this method is likely to have low recall, since tweets and whispers will be missed which do not match their sentence structure.

Davidson et al. (2017) used a hate speech lexicon derived from data from Hatebase.org. Using this they searched Twitter for tweets with terms from the lexicon, producing tweets from 33,458 users. They then retrieved tweets from each of these users' timelines, producing 85.4 million tweets. They randomly sampled 25,000 of these, and annotated them using three or more people sourced from CrowdFlower into three categories: hate speech, offensive language without hate speech, and no offensive content. This produced a sample of 24,802 tweets where there was a majority decision regarding their content. Only 5% of these were deemed to be hate speech according to the majority, and only 1.3% were annotated by all annotators as hate speech. They used five different classifiers: LR, NB, DT, RF and SVM but no ANN, and fivefold cross validation. They found that the LR and SVM classifiers worked best.

Malmasi and Zampieri (2017b) investigated the classification of hate speech into three categories: hate speech, offensive language without hate speech and no offensive content. They used a SVM classifier with tenfold, cross validation and surface ngrams and words skipgrams as feature sets. They used the data that was that produced by Davidson et al. (2017) containing 14,509 English tweets. They found that character 4grams performed best with accuracy in line with the literature in general, although they only reported accuracy. In further work Malmasi and Zampieri (2017a) investigated the use of ensemble classifiers on the same data and features. They found an RBF kernel SVM meta-classifier

number of replies and username.

performed the best, with accuracy of 79.8%. They did not investigate the use of ANN.

Fortuna (2017) investigated the automatic detection of Portuguese-language hate speech. To do this they annotated 5,668 tweets, of which they determine hate speech was contained in approximately 22%. In this small sample she found that hate speech tweets peaked between midnight and 6 AM, at midday and between 6 PM and 7 PM, and that racist tweets were equally distributed throughout the week. They did not perform any ML analysis.

Badjatiya et al. (2017) compared LR, RF, SVM's, Gradient Boosted Decision Trees (GBDTs) and Deep Neural Networks (DNNs) with different feature sets, baselines of ngrams TF-IDF, and BOW versus embeddings created by three deep learning algorithms: FastText, CNNs and Long Short-Term Memory Networks (LSTMs). They used Waseem and Hovy (2016)'s data. They found the deep learning algorithms to outperform the baselines, a LSTM and GBDT combination performing the best. They did not look at any non-textual features.

Park and Fung (2017) performed a two-step classification process using a convolutional neural network (CNN). They used three CNN-based models: having input features characters, words, or both. They used Waseem and Hovy (2016)'s data. They tested to see if a one-step process, that is detecting whether language was sexist, racist or neither, performed differently to a two-step process, that is detecting if language was abusive followed by classifying it into sexist or racist. They compared their CNN-based models with LR, SVM and FastText classifiers as baselines, with tenfold cross validation. They found that two-step classification did perform well and the CNN with both input features gave the best precision, however the results also showed the baseline SVM to perform best in three measures.

Pitsilis et al. (2018) used Recurrent Neural Network classifiers, but did not compare them with others. Their data was that of Waseem and Hovy (2016) from which they used users' history of posting abusive messages. They found, similar to Badjatiya et al. (2017) high F-score values in the region of 0.929 to 0.931. They did not analyse the results of using other algorithms and different feature sets.

2.2 Chapter Summary

The review of the existing work on the automatic detection of hate speech on Twitter[11] concluded that the most commonly used algorithms were: NB, SVM and LR. Many researchers just used one algorithm, others compared NB and SVM. Two groups of researchers compared five different algorithms: Badjatiya et al. (2017) who compared LR, RF, SVM's, Gradient Boosted Decision Trees (GBDTs) and Deep Neural Networks (DNNs) although they did not use NB and they did not use any non-textual features, and Davidson et al. (2017) who used LR, NB, DT, RF and SVM but no ANN. Algorithm results were mixed, some researchers found SVM to be the optimum algorithm, whereas others found NB worked better. Those that used ANN, generally found that these were well, although they were not necessarily compared with SVM.

The performing of the data annotation was divided: some researchers perform their own annotation, whilst others paid workers such as those from Amazon's Mechanical Turk.

In terms of features, again results were mixed. Tokens of the text were treated as BOW or Ngrams (bigrams, trigrams or 5grams). Of the researchers that compared these, the majority found that 5grams worked best, although this was not unanimous. Many combinations of textual and metadata features were compared. Gender, geographical and time zone features showing some benefit to the machine learning process, although again this was not always the case.

The following chapter discusses the methodological issues related to the investigation of racism on Twitter.

[11] Nearly all of the researchers used machine learning to perform the automated classification.

Chapter 3

Racism and Twitter

This chapter is concerned with an understanding of racism on Twitter. In order to answer all of the research questions and to automatically classify racist tweets and tweeters an introduction to Twitter is given followed by various conceptualisations of 'race' and 'racism' and how they can be used to identify and understand the content of the tweets. In the literature the concepts 'race' and 'racism' are often used, sometimes interchangeably, with the concept 'ethnicity', and so an understanding of 'ethnicity' and how it differs from race will be considered. In addition an analysis of what is meant by 'racist speech' is given. To do this the more encompassing term of 'hate speech' is discussed, first from a legal viewpoint, then from a sociological perspective. Then an understanding of racist speech is explored along with how it sits within the broader field of hate speech[1]. Finally the implications these ideas have on the methodology of this research are discussed, included in this is the difficulty of determining the meaning and intent of a racist tweet, and how the word 'nigga' is particularly problematic.

[1] Although a lot of the literature combines hate speech into one category, classifiers often perform better when separated (Burnap and Williams, 2016), and for this reason, the research for this PhD focuses solely on racism.

3.1 Social Networks and Social Media and Twitter

Twitter is an online microblogging service. Murthy (2012, p.3) defines microblogging as

> an internet-based service in which (1) users have a public profile in which they broadcast short public messages or updates whether they are directed to specific user(s) or not, (2) messages become publicly aggregated together across users, and (3) users can decide whose messages they wish to receive, but not necessarily who can receive their messages; this is in distinction to most social networks where following each other is bi-directional (i.e. mutual).

Murthy (ibid.) notes a distinction between 'social network' and 'social media' technologies. Social networks sites are

> [...] web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system (Ellison and Boyd, 2007, p.210).

Although the distinction between social networks and social media is often ignored, they are distinct in that, social media are a medium that allows 'ordinary' people to disseminate 'news', in contrast to traditional news media disseminating traditional news. Social networks on the other hand are the structures that allow this dissemination (Murthy, 2012). So Twitter is both a social medium and a social network. In a social medium capacity it allows people to express what they consider to be of value, and in its social network capacity allows the dissemination of an individual's content.

The distinction between social media and social network can also be seen when Twitter and other social networks are compared. Twitter provides a publicly accessible forum, unlike other social networks such as Facebook, which, although they contain some public data, much of their content exists in small groups accessible only by group members. The public nature of tweets means Twitter's social medium aspect is more to the forefront within its social network, when compared with the status updates and other posts on Facebook, which may be public although normally are only visible to a small number of friends. Twitter users can follow other users they find of interest, including strangers and public figures, and they may engage in conversations with such people (ibid.).

Users of Twitter create status updates known as *tweets* which may be forwarded by other users, these forwarded messages are known as *retweets* (Romero et al., 2011). A user may also have *followers* who subscribe to their updates. Tweets are short messages of 140 characters maximum length.[2]

At the time of writing twitter.com was rated by Alexa[3] as the thirteenth most popular site in the world, and the eighth most popular in the United States, popularity based on a combination of visits and page views (Alexa, 2018). Quantcast (2018) rated Twitter as the fifth most popular site in the United States with over 190 million visits in January 2018 .

On Twitter users are restrained by the 140 character limit. This limitation has somewhat paradoxically appeared to be beneficial in Twitter's uptake and utility, as is often the case in creative endeavours (Alexander and Levine, 2008). Text messages sent by mobile phones similarly illustrate the popularity of short messages.

Originally seen as a site of 'pointless babble' in a short time Twitter,

> increasingly has come to be studied as an emergency communication channel in times of disasters and other major events, as well as an event-following and aid machine for revolution and uprising in the Middle East and beyond (Rogers, 2014, p21).

[2] Although the limit was expanded to 280 characters in 2017 for some users, and this is planned to be extended to all users (Rosen, 2017).

[3] Alexa and Quantcast are companies that provides web traffic data and analytics.

Twitter 'effectively supports a digital agora that promotes real-time interactive exchange of thoughts, opinions and beliefs' (Burnap and Williams, 2016, p.3).

A medium that does this is likely to harbour the views of extremists, and these are often seen on Twitter where racism is not uncommon. How this should be handled will be discussed in Section 3.5. Prior to this, however, it is necessary to gain a theoretical understanding of the concepts of race and racism, which are given in the next section.

3.2 Race and Racism

Morning (2011) conceptualises *race* as 'as system for classifying human beings that is grounded in the belief that they embody inherited and fixed biological characteristics that identify them as members of racial groups' (ibid., p. 21). The notion of humans as belonging to different racial groups is relatively recent, emerging from the 16th Century as a result of European encounters with other cultures during colonialism (Mills, 2014), although the word 'race' was first recorded in English in 1508 (Banton, 1998). European 'scientific' ideas of the eighteenth century promoted the classification of humans into groups delineated by physical traits, and certain colonial practices were justified using these ideas of hierarchical racial differences (Golash-Boza, 2016). These pseudoscientific ideas behind racial categorization have been discredited, although their popular use is still widespread today. This rise of 'scientific racism' during the Victorian era culminated in the eugenics movements of the UK, USA and Germany, during the early part of the 20th century. More latterly 'scientific' ideas of race were no longer seen as acceptable or realistic, since it was no longer plausible to classify people into separate races, as the 'degree of variation within postulated races came to be recognised as greater than the variation between them' (Fenton, 1999). Despite this 'modern' racism still contains notions of racial group hierarchy, and is given credence by institutional acknowledgement of difference, for example by the definition of race in legal statute (Malik, 1996).

Golash-Boza (2016) contends that sociological theorists tend to separate the concepts of race and racism and focus on one or the other. Instead she argues that any analytical framework needs to include both, since they are inextricably linked. She argues that racism includes both:

> (1) the ideology that races are populations of people whose physical differences are linked to significant cultural and social differences and that these innate hierarchical differences can be measured and judged and (2) the micro- and macrolevel practices that subordinate those races believed to be inferior (ibid., p.131).

In contrast Garner (2017) defines racism as having three key elements:

> 1. **A historical power relationship** in which, over time, groups are *racialised* (that is, treated as if specific characteristics were natural and innate to each member of the group).
> 2. **A set of ideas** (*ideology*) in which the human race is divisible into distinct 'races', each with specific natural characteristics.
> 3. **Forms of discrimination** flowing from this (*practices*) ranging from denial of access to resources through to mass murder
>
> (ibid., p. 21).

Both of these definitions state that one aspect of racism is an ideology. For Golash-Boza the notion of ideology is that there are significant tangible differences between races, whereas for Garner ideology is more straightforward in that it only requires each race to have its own set of characteristics. Both definitions include the dimension of ‘practices’ of discrimination. The main difference between the two definitions is that Garner (ibid.) sites racism within a historical process, as he includes the notion of ‘a historical power relationship’ from which the categorization of people into races arises.

These definitions have a number of similarities, all arguing that an ideology that classifies people into different groups, along with practices that perpetuate these divisions are key aspects of racism. Up to now the assumption has been that racism is a binary situation, with acts being classified as either racist or not racist. However Rattansi (2007) questions this dichotomous identification of racists versus non-racists. He argues that there indeed needs to be a spectrum of racism and that this might mean that currently there is too much focus on acts of extreme racism. This is discussed further in Section 8.7.

While this provides an understanding of race and racism, they are often conflated with ethnicity. Racism will be contrasted with ethnicity in the next section.

3.3 Ethnicity

Ethnicity is a fluid process of self-identification, it is the social construction of an identity based on ideas of shared attributes such as heritage, culture, religion, parentage, and language (Quraishi and Philburn, 2015). The boundaries of inclusion and exclusion in an ethnicity are fluid, and can change over time, influenced by political and social factors, especially the struggle for power and resources (Chattoo and Atkin, 2012). The model of ethnicity posited by Jenkins (1997, p.18) argues that ethnicity is about cultural differentiation, and this differentiation includes both similarity and difference. Ethnicity is the outcome of social interaction but at its heart is is concerned with culture. Ethnicity is not an unchanging label, instead it can shift and morph with the culture and situation in which it arises. Ethnicity is both a group and individual form of identity. For an individual it is part of their sense of self, their internal identity, and also part of their external identity, in the form of social interaction. Historically the two concepts race and ethnicity were regarded as mutually exclusive, race was the classification of groups by their genealogy and ethnicity was used to classify groups by a shared culture and history (Chattoo and Atkin, 2012). Although ethnicity appears to be, at face value, a more 'acceptable' concept than race, Webster (2007) argues that race, culture and ethnicity are often used to mean the same thing. He argues that the acceptable, or at least benign, form of ethnicity 'refers to a group possessing some degree of coherence and solidarity based on an awareness of common origins and interests' (ibid., p.2). However ethnic groups, according to Webster, may be seen, either by themselves or by external observers, as 'homogenous, self-perpetuating, defensive and unchanging' and that such groups are often seen as a 'race' (Cashmore, 1996). Indeed in the UK the words 'ethnic' and 'ethnicity' are often used to identify minority cultures, and so, some argue, are tools of 'cultural racism' (Chattoo and Atkin, 2012, p.23).

3.4 Hate Speech

Now that the concepts of race, racism and ethnicity have been considered, concepts related to racist speech need to be explored.[4] In relation to the automated detection of offensive speech of one form or another, there are a number of terms used in the literature to refer to such speech. These are summarised in Table 3.1, which is an updated version of Schmidt and Wiegand (2017)'s work.

It can be seen from the table that there is a plethora of terms used to refer to offensive speech of one form or another. This definitional multitude is much less pronounced in the wider literature where commonly, although not exclusively, the term 'hate speech' is used.

This research analyses *racist speech* on Twitter, as opposed to *hate speech* the term often used in the literature, but it is informative to first look at the broader category of hate speech, since much of the legal theoretical work has been included in this category, and it is directly related to racist speech. From a legal perspective hate speech, and racist speech in particular, is not a single monolithic concept, and is bounded by a series of complex legislative definitions and laws, that are further complicated by the transnational nature of the internet. This makes a transnational study of hate speech legislation important since, to some extent, legal definitions of which language is proscribed will, for example, determine the lexicons used in research in this area. The relationship between hate speech and legislation will be considered in the following sections.

[4] Technically *speech* refers to spoken language, so *hate speech* would not apply to the written form such as that encountered in tweets. However, much of the literature refers to both spoken and written forms of hate as 'hate speech' and this terminology will be used in this thesis as well.

Table 3.1: Different terms related to hate speech, adapted from the work of Schmidt and Wiegand (2017).

Terminology	**Authors**
Abusive messages, hostile messages or flames.	Spertus (1997).
Cyberbullying.	Xu et al. (2012), Hosseinmardi et al. (2015), Zhong et al. (2016), Van Hee et al. (2015), Dadvar et al. (2013) and Dinakar et al. (2012).
Hate speech.	Warner and Hirschberg (2012), Burnap and Williams (2015), Silva et al. (2016), Djuric et al. (2015), Gitari et al. (2015), Williams and Burnap (2015), and Kwok and Wang (2013).
(Personal) insults, swearing and user posts that are characterised by malicious intent.	Sood et al. (2012a).
Offensive language.	Razavi et al. (2010).
Vulgar language and profanity-related offensive content.	Xiang et al. (2012).
Jokingly formulated teasing.	Xu et al. (2012).
Othering language.	Burnap and Williams (2014).
Slurs.	Bartlett et al. (2014).

Legislation regulating ‘hate speech’, that is the use of written or oral means to victimise minorities or other marginal groups, is a relatively recent phenomenon, having originated in Europe in the second half of the 20th century. Partly as a result of its youthfulness, legislation between nations is not always consistent, even within the EU. For example victims of hate crimes are defined differently in England and Wales and Sweden, anyone can become a victim of a crime from a legal perspective in England and Wales, whereas in Sweden only members of ethnic minorities can be victims of hate crimes. If a minority attacks a majority group, this is not regarded as a crime in Sweden (Hardy and Chakraborti, 2017).

Discussions of hate speech have mostly taken as their starting point legal ideas and legislation, but before how hate speech seen from a legal perspective is discussed, a review of the legislative situation with respect to racism in England and Wales[5] will be given, along with a brief summary of the law in some other countries which use the adversarial system of justice, Canada, Australia and the United States and a summary of the differences between how Europe and the United States differ in their legal handling of hate speech.

3.4.1 Hate Speech and the Law of England and Wales

In England and Wales hate speech is not protected as free speech as it is in the United States. Similar to Canada and Germany England and Wales has laws prohibiting racist discourse (Pitsilis et al., 2018). England and Wales’ law is largely divided into these two types of racist offences: those directed against a person or group, and those that generally incite racial hatred.[6]

In order to further discuss the legal situation with respect to hate speech, a brief

[5] Within the UK Scotland and Northern Ireland have their own legislation, much of which is similar to that of England and Wales.

[6] Racist language in tweets can be directed at an individual or a group, or not directed at anyone in particular but instead being general declarations. For example a tweet that contains the phrase ‘Jack is a nigger’ is obviously targeted towards Jack and the phrase ‘Jews from Margate are scum’ also obviously targets Margate based Jews, whereas the phrase ‘I hate wops’ is a generic attack against people of Italian ethnicity.

discussion of the elements that make up a crime, and the criminal law of England and Wales is needed. The criminal law of England and Wales adopts the principle of 'actus non facit reus, nisi sit mens rea' which can be roughly translated as 'a person is not criminally liable just by an act but must be accompanied by a guilty mind'. This means the criminal law breaks down a criminal act into two parts: the *actus reus* and *mens rea.* The actus reus is the act that the defendant performed, and for a prosecution to be successful in England and Wales this must be proved beyond reasonable doubt. As well as the actus reus having to be proved beyond reasonable doubt, the same test is applied to the mens rea. The mens rea is the intention to commit a crime, and there are different levels of intention with respect to different offences. For example speeding is a 'strict liability' offence meaning the actus reus is enough, and no mens rea is necessary for a successful prosecution. Murder, on the other hand, is an offence which can only be successfully prosecuted if a deliberate intention is shown, beyond a reasonable doubt. This discussion of mens rea is relevant to a discussion of hate speech, since there is inconsistency in legislation with respect to whether intention is required to commit racist crimes. For example the EU's *Convention on Cybercrime* was designed to combat racist discourse mediated by the internet. In Articles 2, 5 and 6 of the convention, there are references to 'hate'. However this is an unusual requirement in hate speech legislation, most hate crime legislation does not require this aspect of the mens rea to be present, and racist acts can occur and be prosecuted without any hatred on the part of the perpetrator (Garland and Chakraborti, 2012).

With this understanding of what constitutes a criminal acts under the law of England and Wales, there now follows an exploration of the relevant pieces of legislation, that tackle racism and hate speech in England and Wales. There are a large number of such pieces of legislation and focus will be given to the main ones that have been used against (or have the potential to be used against) racist speech. In England and Wales legislation related to racial hatred, initially framed such offences as public order offences. The first such attempt at prohibiting racial hatred was the *Race Relations Act 1965* which was an amendment to the *Public Order Act 1936* (Malik, 1999). Under the section headed 'Public Order' was included the new offence of 'incitement to racial hatred', which stated that,

> a person shall be guilty of an offence under this section if, with intent to stir up hatred against any section of the public in Great Britain distinguished by colour, race, or ethnic or national origins
>
> (a) he publishes or distributes written matter which is threatening, abusive or insulting; or
>
> (b) he uses in any place or at any public meeting words which are threatening, abusive or insulting,
>
> being matter or words likely to stir up hatred against that section on grounds of colour, race, or ethnic or national origins. (*Race Relations Act 1965 s. 1 (1)*).

This was the first piece of legislation in England and Wales to specifically refer to hatred as a result of race or ethnicity. It was amended by the *Race Relations Act 1968*, which broadened the protections, making it an offence to discriminate 'on the ground of colour, race or ethnic or national origins' (*Race Relations Act 1968 s. 1 (1)*) with respect to housing, accommodation and public services (legislation.gov, 2018a). Both these acts were regarded as weak legislation and were replaced by the *Race Relations Act 1976* (Sooben, 1990). Both of the 1960s acts only prohibited discrimination directly aimed at an individual (Quraishi and Philburn, 2015). The 1976 act no longer required hatred as motivation, instead being treated 'less favourably' as a result of belonging to a 'racial group' was now prohibited (*Race Relations Act 1976, s. 1(1)*). This act was later amended by the *Race Relations Amendment Act 2000*, and repealed by the *Equality Act 2010*, which was designed for England and Wales to meet EU requirements in racial discrimination law (ibid.).

The public order nature of racist offences in England and Wales are also shown by Section 21 (1) of the *Public Order Act 1986* which prohibits individuals from distributing, showing or playing recordings (both visual and audio) that are 'threatening, abusive or insulting' and by which '(a) he intends thereby to stir up racial hatred, or (b) having regard to all the circumstances racial hatred is likely to be stirred up thereby.' Strickland and Douse (2012) argue that use of the word 'insulting' might be seen as an unnecessary infringement of free speech, and there is an uneasy tension in England and Wales between the necessary curtailment of racist speech, and the impact this has on denying people's rights to express themselves freely. This is not the case in the United States, where the First Amendment to the US constitution protects a citizen's right to express themselves freely, whether that expression is racist or not. A case that illustrates this difference between the two countries' attitudes to the right to espouse racist views, and also use of the *Public Order Act 1986* to prosecute racist speech, is that of *R v Sheppard and Whittle* (2010) EWCA Crim 65. In this case the defendants published holocaust denying material on their website, and were successfully prosecuted under the *Public Order Act 1986*.[7] Their defence had argued that despite the fact that the material was uploaded in England, it was hosted on a website in California in the US, and thus the US protection of freedom of speech meant that they had done nothing wrong. The judge ruled that despite this US-based hosting of the website, most of the activities needed to publish the material, its creation, editing and uploading, were performed in England and Wales. On appeal this ruling was upheld by the Court of Appeal and the defendants convictions were allowed to stand (Gillespie, 2010).

The *Public Order Act 1986* was amended by the *Racial and Religious Hatred Act 2006* to expand the protections already afforded to racial groups to religious groups. Although this amendment only applied to England and Wales. Prior to this amendment there was no protection for hatred against, for example Muslims.[8] The *Crime and Disorder Act 1998* added to the public order offences of the *Public Order Act 1986*, a number of new racially aggravated criminal offences including harassment and threatening behaviour. For example section 18 (1) states:

[7] Digital crime is not constrained by physical or legal boundaries and often involves technology situated in many countries. Therefore issues of both material jurisdiction (under which jurisdiction was the offence committed?) and procedural jurisdiction (which jurisdictions' procedural rules govern the investigation and evidence) are important (Walden, 2007, p. 298 and p. 310).

[8] There was some concern prior to the act being enacted, that it would lead to prosecutions of people criticising religions. To counter this Section 29J provided a 'Protection of freedom of expression' (Barendt, 2011) which stated:

> Nothing in this Part shall be read or given effect in a way which prohibits or restricts discussion, criticism or expressions of antipathy, dislike, ridicule, insult or abuse of particular religions or the beliefs or practices of their adherents, or of any other belief system or the beliefs or practices of its adherents, or proselytising or urging adherents of a different religion or belief system to cease practising their religion or belief system. (legislation.gov, 2018b).

> A person who uses threatening, abusive or insulting words or behaviour, or displays any written material which is threatening, abusive or insulting, is guilty of an offence if
>
> (a) he intends thereby to stir up racial hatred, or
>
> (b) having regard to all the circumstances racial hatred is likely to be stirred up thereby (*Public Order Act 1986*).

Section 19 has a similar offence, but is concerned with 'publishing or distributing written material'. The *Anti-terrorism Crime and Security Act 2001*, amended the *Public Order Act 1986* to include religious versions of these offences, which were added as new sections 29B and 29C, in addition to the existing racial ones. Prosecutions under section 19 and 29C 'must be referred to the Special Crime and Counter Terrorism Division (SCCTD) and proceedings require the consent of the Attorney General' (CPS, 2018). There has to be a public element to these offences, and the Crown Prosecution Service (CPS) has determined the internet to be a 'public space' (Williams and Pearson, 2016).

As well as public order legislation, England and Wales has laws aimed at preventing 'malicious communications' and there are instances when racist or other hate speech can be classed as these types of communications. The *Malicious Communications Act 1988* prohibited the sending of 'letters etc. with intent to cause distress or anxiety' (*Malicious Communications Act 1988 s. 1*). Its interpretation with respect to whether this included electronic communications, was in some doubt until it was amended by the *Criminal Justice and Police Act 2001* and its replacing of 'letter or other article' by 'letter, electronic communication or article of any description' (s. 43 (1) a) as a prohibited form of malicious communication. Even messages that are sent, for example, by social media such as Twitter, can be subject to prosecution under these acts, even if they do not reach, or are read by, their intended recipients.

The CA 2003 contains similar provisions although is more aimed at prohibiting the use of a public network than protecting an individual (CA 2003 s 127 (2); Walden, 2007: 151).[9] Section 127 of the CA 2003 was also used to successfully prosecute a man who

[9] Paul Chambers succeeded in a High Court Appeal against his prosecution under section 127(1) of

sent racist tweets to Stan Collymore, a former footballer (Edwards, 2012). The *Protection From Harassment Act 1997* (PFHA 1997) provides protection similar to the MCA 1988 but for a series (2 or more) of malicious communications, and contains provision for issuance of restraining orders (PFHA 1997 s 2).

This section has given some insight into the legislation available to prosecute hate speech in England and Wales. Ultimately decisions about whether to prosecute are made by the CPS, who take into account available legislation and other factors as discussed in the next section.

3.4.2 CPS and Racism

The police in England and Wales pass cases they investigate onto the Crown Prosecution Service (CPS) who handle prosecution. When a case arrives at the CPS it may already have been flagged by the police as being racially and/or religiously aggravated. The CPS may also similarly flag the case at any stage of the prosecution. The definitions of what is and what is not racially or religiously aggravated were jointly created with Association of Chief Police Officers (ACPO, now the National Police Chiefs' Council) and they are:

> Any incident/crime which is perceived by the victim or any other person to be motivated by hostility or prejudice based on a person's race or perceived race
> or
> Any incident/crime which is perceived by the victim or any other person to be motivated by a hostility or prejudice based on a person's religion or perceived religion.

The flag can mean: prosecution for a particular racial or religiously motivated crime, or an increased sentence for a different crime (under s145 of the *Criminal Justice Act 2003*)

the CA 2003. In 2010 Chambers tweeted: 'Crap! Robin Hood airport is closed. You've got a week and a bit to get your shit together otherwise I'm blowing the airport sky high!!' which was originally ruled to be menacing (Bowcott, 2012).

but it can also act as a signifier showing that the CPS take such matters seriously even if it has no effect on sentencing. These definitions and the legislation they are based on require consideration of important questions: what is meant by a racial group, how is membership of such a group determined and what constitutes hostility (CPS, 2017)? Racial group is defined under the *Crime and Disorder Act 1998* s 28 94) as: 'a group of persons defined by reference to race, colour, nationality (including citizenship) or ethnic or national origins.' Case law has supported the inclusion of Romany gypsies and Irish Travellers as racial groups (Commission for Racial Equality v Dutton [1989] QB 783 and O'Leary v Punch Retail (HHJ Goldstein, Westminster County Court, 29 August 2000)) and the European Court of Human Rights have determined the Roma are also such a group (ibid.). There is some support in case law for the notion that religious belief can determine membership of a racial group, for example Sikhs (Mandla v Dowell-Lee [1983] 2 AC 548) and Jews (King-Ansell v Police [1979] 2 NZLR 531, R v JFS [2009] UKSC 15 and Seide v Gillette Industries Ltd [1980] IRLR 427, ibid.). In the appeal of R v Rogers (2007) 2 W.L.R. 280 the House of Lords (HOL) advised that membership of ethnic groups also extended to nationality.

The definition of hostility is not addressed by legislation. But case law is more useful and judgements have suggested that there must be evidence of either spoken or written hostile words or hostile action. These might include swearing (for example "black bastard" (R v Woods [2002] EWHC 85) or "African bitch" (R v White [2001] EWCA Crim 216) or the use of an offensive banner (RG & LT v DPP [2004] EWHC 183, ibid.). The CPS (ibid.) state that the hostility must be contemporaneous with the act either immediately prior, during or after the event. When considering racist motivation, as opposed to an actual racist act, the racist element might be the membership of a racist group or use of racist speech in the past. In the appeal of R v Rogers (2007) 2 W.L.R. 280 the House of Lords (HOL) advised that membership of ethnic groups also extended to nationality. The HOL also noted that hostility directed mistakenly at a person who is not a member of the racial group being denigrated, still counts as hostility under the law, and that it did not matter if other factors as well as racial ones led to the hostility. The CPS (ibid.) also note that the racial abuse does not require any dislike or ill feeling towards the group in question and the victim's reaction to it is also irrelevant. The CPS specifically mention 'social media postings;' as a form of evidence that may be part of a case concerning a racially motivated offence or offences.

Due to the global nature of Internet communications, any discussion of relevant legislation should take into account that of other countries. Space does not allow for a global review but the relevant legislation from some example countries will now be discussed.

3.4.3 Hate speech and Other Countries

3.4.3.1 Canada

Canada has hate speech laws at both the federal and provincial levels, and these include both criminal law and human rights legislation (Martin, 2018). Section 319 of the Criminal Code of Canada prohibits the incitement of hatred 'where such incitement is likely to lead to a breach of the peace' and also the wilful public promotion of hatred. Both offences prohibit behaviour targeted at 'any identifiable group' and must occur in a public place, (*Criminal Code of Canada, RSC, 1985 c C-46*). This section has been challenged, unsuccessfully, in court a number of times since, some argue, it infringes upon the right of freedom of expression, also enshrined in Canadian law (ibid.). The law only applies if the hate speech occurs in a public venue, and if what was said is deemed to be true, or the basis of a religious belief, or is in the public good are all legitimate defences. The law covers the use of *statements*, which include electronic messages.

3.4.3.2 Australia

Australia is federally constituted with six states and two self-governing Territories, and Australian law tackles hate speech at both the national and subnational level (Gelber and McNamara, 2015). The primary federal legislation is the *Racial Discrimination Act, 1975* in which s. 18C states:

> It is unlawful for a person to do an act, otherwise than in private, if:
>
> a the act is reasonably likely, in all the circumstances, to offend, insult, humiliate or intimidate another person or a group of people; and
>
> b the act is done because of the race, colour or national or ethnic origin of the other person or of some or all of the people in the group.

Similar to Canada, this requires the act to be performed in public, although the state of Victoria does allow for prosecutions of private hate speech (Sellars, 2016). At the subnational level, those that have both civil and criminal laws, i.e. New South Wales, Queensland, Victoria, the Australian Capital Territory, and South Australia, have never successfully prosecuted a criminal hate speech case. In contrast, Western Australia, which only has criminal hate speech laws, has had successful prosecutions (Gelber and McNamara, 2015).

3.4.3.3 European Union vs the United States

The US and Europe have very different views on the regulation of racist hate speech. It is very difficult in the United States to prosecute for racist speech or for the use of racist symbolism, unless they are related to imminent harm or danger. In Europe, however, there is a very different attitude towards prosecution of racist speech and other expressions, and there is a history of racist speech legislation dating back to the 1960s.[10] Bleich (2014) argues that from a legal perspective these differences are largely a result of the differences between rulings from the US Supreme Court and the European Court of Human Rights (ECtHR). He argues that a series of rulings by the Supreme Court, during the rise of the civil rights movement in the United States, underlined the protection of free speech afforded by the first amendment of the US Constitution. In contrast the ECtHR's rulings on interpretation of the European Convention on Human Rights (ECHR) have given member states protection of any antiracist speech laws. The ECHR has conflicting

[10] The aim of the European Union's Framework Decision on Combating Certain Forms and Expressions of Racism and Xenophobia By Means Of Criminal Law (EU framework decision 2008/913/JHA) was to harmonise the legislation aimed at tackling racism and xenophobia across the EU. Most European states had not implemented the full measures of the directive by its deadline of 28 November 2010, (Lobba, 2014) but is now largely in force (EJN Secretariat, 2018).

provisions, its article 10 states:

> Everyone has the right to freedom of expression. This right shall include freedom to hold opinions and to receive and impart information and ideas without interference by public authority and regardless of frontiers.

whilst article 14 states:

> The enjoyment of the rights and freedoms set forth in this Convention shall be secured without discrimination on any ground such as sex, race, colour, language, religion, political or other opinion, national or social origin, association with a national minority, property, birth or other status.

and the possibility exists that any rights exerted under article 10 to use racist speech might curtail an individual's rights under article 14. However the ECHR has a 'get out' article 17, which reads:

> Nothing in this Convention may be interpreted as implying for any State, group or person any right to engage in any activity or perform any act aimed at the destruction of any of the rights and freedoms set forth herein or at their limitation to a greater extent than is provided for in the Convention.

This article has meant that the ECtHR has been able to rule in favour of laws prohibiting racist speech, since this article prohibits behaviours that infringe any of the other rights afforded by the convention (Bleich, 2014).

These brief examinations of different legal systems handling of hate speech, illustrate the heterogeneity of how hate speech is handled in different countries. Despite this many definitions of hate speech are based on legal ideas and these definitions will now be reviewed.

3.4.4 Definitions of Hate Speech

Delgado (1982) looked at racist speech, and argued that racist speech should be sanctioned by law, contending that, for a successful prosecution of a racial insult, the following would need to be proved:

> [that the] language was addressed to him or her by the defendant that was intended to demean through reference to race; that the plaintiff understood as intended to demean through reference to race; and that a reasonable person would recognise as a racial insult (ibid., p.179).

This definition focuses solely on the intended impact of racist speech on a victim, while saying very little about the content. It attempts to be objective by using the legal idea of the 'reasonable person' (Sellars, 2016), which, itself, is a difficult concept open to different interpretations (Moran, 2003). Matsuda (1989) also take a legal approach and builds on the work of Delgado (1982), while trying to address criticisms that she feels proponents of the US First Amendment of the Constitution would aim at any definition of racist speech, as curtailing freedom of speech, by the following requirements for racist speech to be prosecutable:

> The message is of racial inferiority; 2. The message is directed against a historically oppressed group; and 3. The message is persecutorial, hateful, and degrading (Matsuda, 1989, p.2, 357).

This definition of racist speech more overtly states requirements for the content to be seen as racist speech, noting that it is about 'racial inferiority' and that the target is 'a historically oppressed group' (Sellars, 2016). Massey (1992) criticises these approaches for using subjective terms, or terms that are too specific, thus predetermining the outcome. Instead he defines hate speech as:

> any form of speech that produces the harms which advocates for suppression

> ascribe to hate speech: loss of self-esteem, economic and social subordination, physical and mental stress, silencing of the victim, and effective exclusion from the political arena (Massey, 1992, p.105).

As Massey (ibid.) himself notes, this is something of a catchall definition, treating all levels of hate the same. Moran (1994) was also critical of these attempts, noting that any definitions of a concept like hate speech, are subject to a number of subjective choices that arise out of particular worldviews, biases, political standpoints and so on. Nevertheless she argues that there are certain common elements seen in attempts to define hate speech, and provides her own definition of hate speech as: 'speech that is intended to promote hatred against traditionally disadvantaged groups' (ibid., p.1, 430). This again is a very wide definition, and, according to Moran (ibid., p.1, 430), written pornography could fall 'within the class of speech plausibly encompassed by the term hate speech'. There is no content element in this definition, it also includes more subtle speech within its boundaries, since it targets speech that *promotes* hatred, as opposed to, for example, Matsuda's (1989) requirement that the speech is 'hateful' itself (Sellars, 2016).

Ward (1997, p.765) defined hate speech as 'any form of expression through which speakers primarily intend to vilify, humiliate, or incite hatred against their targets.' This definition combines hate speech as both that which incites hatred or is directly hateful itself.

Benesch (2013) examines, what she calls *dangerous speech*, a subset of hate speech that is speech with an element of incitement to violence. According to her, the dangerousness of speech depends on five variables which are: the speaker, the audience, the speech act itself, the social and historical context and the means of dissemination. The most dangerous speech is when these five variables are maximised, such as when there are:

- A powerful speaker with a high degree of influence over the audience.
- The audience has grievances and fear that the speaker can cultivate.
- A speech act that is clearly understood as a call to violence.
- A social or historical context that is propitious for violence, for any of

> a variety of reasons, including longstanding competition between groups for resources, lack of efforts to solve grievances, or previous episodes of violence.
>
> - A means of dissemination that is influential in itself, for example because it is the sole or primary source of news for the relevant audience.
>
> (ibid.).

Benesch's focus is unusual in that it gives a lot of weight to the context of the speech act, including societal and situational factors that may influence it (Sellars, 2016).

Parekh (2012) aimed to draft a definition of hate speech by looking at examples of hate speech, noting the wide variety of forms that hate speech can take. In his definition, hate speech comprises three dimensions:

> It is directed against a specified or easily identifiable individual or, more commonly, a group of individuals based on an arbitrary or normatively irrelevant feature (ibid., p.40).
>
> It stigmatises the target group by implicitly or explicitly ascribing to its quality is widely regarded as undesirable (ibid., p.41).
>
> Because of its negative qualities, the target group is viewed as an undesirable presence and a legitimate object of hostility (ibid., p.41).

In this definition, there is no requirement for harm, and so it is likely to be problematic for free-speech advocates, since the lack of a requirement for harmful speech means it is likely to include what might be seen as 'less serious' speech, and so more likely to lead to curtailment of 'borderline' views (Sellars, 2016).

Marwick and Miller (2014) examined existing definitions of hate speech, and found that there are three elements that are used to define hate speech, those of content, intent and harms. The content element includes offensive words and phrases, and symbols and iconography. The intent element requires that the speech intends:

> only to promote hatred, violence or resentment against a particular minority, member of a minority, or person associated with a minority, without communicating any legitimate message (Marwick and Miller, 2014, p.17).

The harm element is speech can cause either mental or physical harm to the victim or affects their self-esteem, social standing or affects them economically (ibid., p.17). While there is some disagreement about what constitutes hate speech in the literature, the elements noted by Marwick and Miller (ibid.) provide a good basis to identify hate speech on Twitter.

Although this research focuses on *racist* rather than *hate* speech these three elements still apply, the only difference being that instead of them applying to *minorities*, that they apply to *racial or ethnic groups.*

However even if there is a basis for identifying the dimensions of racist speech, the subtleties of written language create further difficulties and these will be discussed in the next section.

3.4.5 Difficulties in Interpreting Tweets

While the legal and definitional aspects of what constitutes racist speech are of course important, any system attempting to analyse racist tweets needs to go beyond this and and try to determine what was the intent[11] of the tweeter. Determining intent can be tricky as can be identifying racist or potentially racist tweets. If a tweet contains a single racial epithet, it is very difficult to determine whether it is racist without any other

[11] Intention is particularly important in terms of determining whether a racist tweet (or any behaviour) meets requirements for action by law enforcement. Even if the tweet does not meet the standard for prosecution as determined by the CPS, the tweet might still be problematic and perhaps trigger a warning, or its removal by Twitter. See Section 3.4.1 for more on this.

information. The tweeter may be drawing attention to use of the word, they may be using it in a racist sense, they may have misspelt the word, they may be replying to another tweet and so on.

There is also some debate as to whether slurs can have both expressive and descriptive uses. Some argue that slurs are only expressive, that is they have no descriptive value, and are only used to express negative emotions towards others, whereas others argue that they do have a descriptive role (O'Dea and Saucier, 2017). The latter view is more convincing since, as Anderson and Lepore (2013, p.25) note:

> these are expressions that target groups on the basis of race ('nigger'), nationality ('kraut'), religion ('kike'), gender ('bitch'), sexual orientation ('fag'), immigrant status ('wetback') and sundry other demographics.

Also there is an issue of context. Some scholars argue that slurs and racial stereotypes are derogatory and negative no matter what context they are used in, this view being a *context insensitive* position. Others argue that context is important, and that slurs are sometimes not derogatory, particularly when used within the group that the slur is aimed against, this conflicting position is a *context sensitive* position (Croom, 2015).

Whether context is important or not, and whether tweets are descriptive or not, it is still impossible to determine what a tweeter actually intends by the use of a particular word, and all research on discourse has this underlying problem. Meanings can only be guessed at, and heuristics and subjective guesswork is often used in this type of research. Some words, however, are more difficult to handle than others. A major difficulty in determining the meaning of a tweet from this data is understanding the use of the words 'nigger' and 'nigga'. The tweet

> @KDTrey5 you're a nigger,[12]

[12] The data discussed in this section are from the popular predicted racist tweets as discussed in Section 7.3.1.

seems to be a fairly obvious racist attack on @KDTrey5, although even here there is some ambiguity as, for example, this could be a response to @KDTrey5 asking what racist comment had been used. Unless context suggests otherwise, it seems reasonable to assume that uses of the word 'nigger' can be taken to be racist in nature. However the word 'nigga' is more problematic. There is generally a consensus that often 'nigga' is used as a positive identifier of the self, although whether this is acceptable is another matter, since it can be argued that using it in this way helps reinforce power differentials (Andrews, 2014). It is also impossible to determine how each of the participants: creator, retweeter, target of the tweet, and the audience of the tweet react to its content. The word 'nigga' might have been used by the original sender of the tweet, to show affection or solidarity with its target, or it might be used in the same form as 'nigger', that is its use is intended to be racist. For example the tweet,

> The nigga was raping multiple women; he mad folks don't fw him . Niggas are literally figg*g stupid. `https://t.co/kqyHdhHOos`,

Relates to Ian Connor the fashion designer who is alleged to have raped 21 women (Law, 2018). Here 'nigga' is used as a wholly pejorative term.

The most popular retweet in the dataset was:

> Why black people don't swim , nigga almost drowned and tried to play it off ??? https://t.co/kihaADVhBb,

was retweeted 20,465 times, although this number comprises a number of tweets that all contain the same text, but have different URLs. The URLs all pointed to the same video of a black man struggling in a swimming pool, then acting nonchalantly at the end of the video, as if he had not had any problems and was enjoying himself. This video with the concomitant text, plays to the stereotype of black people being poor swimmers. This stereotype might seem harmless at first sight, but it has been suggested that it has led to black children being afraid to swim, believing they are unable to do so well, and has led to a disproportionate number of child drowning deaths in the black population of the United States (Rosemond, 2017). This is a convincing argument for the harm potentially done by this, casual stereotyping, but despite this the aim of this research is to determine whether the tweet is racist or not. Again the use of the word ‘nigga’ is problematic, and it is very difficult to determine the motives behind the sending of this tweet, especially since there are so many retweets of it. Determing motive of the sender is complicated by issues related to ethnicity. It might be argued that it is acceptable (or at least not racist), for a person of black ethnicity to retweet this tweet, or, alternatively there is no such assumption that people cannot be classed as racist if they ridicule their own ethnicity.

Despite these issues an attempt was made when coding tweets that were collected, to determine whether they were racist or not. Coding was performed by five people and to minimise subjectivity a series of rules were used designed to minimise variability between coders, such as classifying a tweet only containing the use of a racial epithet as racist (see Section 5.10 for further details).

3.4.6 Personal Racism

The racist discourse seen in the texts in this research is what might be called *personal racism*, that is it is used to elicit a negative response in other people, presumably in the hope that the people who read it are shocked. There are of course other more subtle ways of expressing racist ideas, but this research does not analyse this kind of data, instead it focuses solely on tweets which have an ethnic slur in them, i.e. overt racist discourse. The sending of racist tweets would include Garner (2017)’s three elements of racism since like any racist text they are underpinned by power relationships between groups such as UK ethnic whites and blacks, these groups are seen as distinct by those sending racist tweets targeting a particular group, and the recipient or the recipients of the tweet are discriminated against since the abuse is aimed against them.

Now that racism and racist language has been explored, it is necessary to discuss the roles and responsibilities of the different actors involved in policing racist language on the internet and Twitter in particular, and these are discussed in the following section.

3.5 Policing the Internet

There is a growing trend for crime to be perpetrated via the internet. When cybercrime was added to the Crime Survey of England and Wales as a pilot in 2015, the overall amount of crime doubled (Office for National Statistics, 2015). These new offences require significant resources to combat, and whether this is possible, has been questioned at the highest level of government (Horsman et al., 2017). As a result any procedures that aid in the automation of policing the internet are extremely useful. Automatically identifying racist tweets can be beneficial to identify potentially illegal behaviour but also - if one accepts the ‘broken windows’ theory of policing[13] - identifying racists by their tweets can aid law enforcement in early intervention and thus hopefully reduce furthermore serious offences that may occur at a later date.

The regulation and governance of cyberspace, involves a number of ‘stakeholders and actors including nation states, public and private organisations and movements such as the activist group "Anonymous"’ (Bryant and Day, 2014, p.84). Indeed there is debate as to whether cyberspace should even be controlled at all. As well as this there are other issues that come into play in policing the internet, such as technical standards of the internet, censorship by organisations and states, and the concept of *net neutrality* (ibid., p.84). Some argue that the traditional power of the nation-state has been eroded and in its place there has been a concomitant rise in the influence of organisations that are transnational and distributed in nature. Some examples of these new organisations are the Internet Engineering Task Force (IETF)

> ‘which standardises the internet’s protocols via memos known as Request for Comments (RFCs) and the ‘Internet Corporation for Assigned Names and Numbers’ (ICANN) which was created in 1998 to administer IP addresses and

[13] ’Broken windows’ policing is based on the notion that that minor offences can lead to serious crimes (Ratcliffe, 2016, p. 49).

the DNS system (Mueller, 2010).

The complicated nature of the internet's regulation and the way it has been shaped by varying interests especially its decentralisation of control [14] has created tensions between nations attempting to control the internet. Countries ranging from China to the UK have attempted strong internet regulation which is in conflict with the multi-stakeholder nature of other parts of internet regulation (Kleinwächter, 2017).

As well as the control the nation-states try to exert, there are many other non-governmental voices attempting to influence the debate over regulation of the internet. For example, perhaps the best-known 'anti-hate' organisation in the United States, the Anti-Defamation League promote their Best Practices for Responding to Cyberhate, a series of rules for both providers and the wider internet community aimed at combating cyberhate (ADL, 2014). Many providers, including Twitter, are listed on the ADL webpage as supporters of the initiative. These Best Practices stress the importance of cooperation within the whole 'Internet Community' and stress that providers should 'offer users a clear explanation of their approach to evaluating and resolving reports of hateful content', something arguably that Twitter is lacking in, or at least lacking in publicising their efforts, which hampers their role as a capable guardian.

The United States often takes the informal lead in such initiatives, and similarly much of the research on racism comes from the US, and many tweets analysed in this research are created by US accounts. It can, therefore, be seductive to see the US as 'the home of racism'. This however is problematic. Smith (2017) argues that in the UK[15], discourse about what constitutes a 'racist society' is in large part determined by deflection to other countries, particularly the apartheid system in South Africa and the Jim Crow

[14] The internet began life in 1969, with four computers connected via the world's first packet switching network (packet switching allows for data to be passed along shared network channels, which improves upon circuit switching which requires dedicated channels). In 1983 this network was the first to use the 'Internet Protocol' and the 'Domain Name System' to standardise communication. At the same time the network was separated into a military-only MILNET and the remaining ARPANET. The ARPANET was replaced by an academic network, NSFNET which was replaced by a commercially open internet in the 1990s. During the late 1980s and early 1990s the network was connected to European and Asian TCP/IP networks (Bryant and Day, 2014, p.85).

[15] And its former colonies that are predominantly 'white, or have a history of white rule': Australia, New Zealand, Canada, the United States and South Africa.

laws of the southern United States. He argues that this ‘othering’ of racist practices in other countries, is widespread, and masks the local racist practices, which while anti-racist discourse within the country abhors what is occurring in another country, is oblivious to practices within its own country that may be seen as racist by others.

Indeed racism, hate crime and hateful language are significant problems within the UK, and initiatives to tackle the problem have been discussed at a governmental level. In 2016 the UK government published ‘Action against hate: the UK government’s plan for tackling hate crime’ which outlines the UK’s strategy for combating hate crime both in the ‘real world’ and online, the government states its aim is to work ‘in partnerships with communities and joining up work across the hate crime strands to ensure that best practice in tackling hate crime is understood and drawn upon in all our work’ (Home Office, 2016, p.10). The government aims, inter alia, to tackle online hate crime by convening a: ‘ministerial seminar on hate on the internet that brings together victims’ groups, stakeholders and industry representatives’ (ibid., p.27).[16] However, the UK government’s online hate crime initiatives have been criticised as having very little detail on how they will work (Walters and Brown, 2016).

Other government initiatives include Mayor of London Sadiq Khan announcing a new online hate crime unit run by the Met police in April 2017 (Mayor of London, 2017). The government also funded a new online hate crime hub although it has been criticised for not providing sufficient funding for this (Doward, 2017), and at the time of writing (February 2018) it is still not operational.

3.6 Chapter Summary

Twitter is both a social medium: it allows people to express what they consider to be of value, and a social network: it allows the dissemination of an individual’s content. The

[16] They also note existing initiatives including True Vision: an online hate crime reporting portal, the flagging of any online element of a crime by UK police, the Violence against Women and Girls Group online working-group, the CyberHate working-group and work with the Internet Watch Foundation (IWF) to counter online hate crime.

public nature of tweets means Twitter's social medium aspect is more to the forefront within its social network, when compared with the status updates and other posts on Facebook, which may be public although normally are only visible to a small number of friends.

Ideology that classifies people into different groups, along with practices that perpetuate these divisions are key aspects of racism. Rattansi (2007) questions the dichotomous identification of racists versus non-racists, arguing that there needs to be a spectrum of racism and that this might mean that currently there is too much focus on acts of extreme racism. Race and racism, are often conflated with ethnicity. Ethnicity is a fluid process of self-identification, it is the social construction of an identity based on ideas of shared attributes such as heritage, culture, religion, parentage, and language. However ethnic groups may be seen, either by themselves or by external observers, as homogenous and such groups are often seen as a race. Indeed in the UK the words 'ethnic' and 'ethnicity' are often used to identify minority cultures, and so, some argue, are tools of 'cultural racism'.

The criminal law of England and Wales breaks down a criminal act into two parts: the *actus reus* and *mens rea*. The actus reus is the act that the defendant performed, and the mens rea is the intention to commit a crime. For a prosecution to be successful in England and Wales these must be proved beyond reasonable doubt. In England and Wales legislation related to racial hatred, initially framed such offences as public order offences. Relevant legislation includes the *Public Order Act 1986*, which proscribed the stirring up of racial hatred. The *Public Order Act 1986* was amended by the *Racial and Religious Hatred Act 2006* to expand the protections already afforded to racial groups to religious groups. The *Crime and Disorder Act 1998* added to the public order offences of the *Public Order Act 1986*, a number of new racially aggravated criminal offences including harassment and threatening behaviour. There has to be a public element to these offences, and the Crown Prosecution Service (CPS) has determined the internet to be a 'public space'.

As well as public order legislation, England and Wales has laws aimed at preventing ‘malicious communications’. The *Criminal Justice and Police Act 2001* amended the *Malicious Communications Act 1988* to make illegal the sending of any electronic message intended to cause distress. The police in England and Wales pass cases they investigate onto the Crown Prosecution Service (CPS) who handle prosecution. Racial group is defined under the *Crime and Disorder Act 1998* s 28 94) as: ‘a group of persons defined by reference to race, colour, nationality (including citizenship) or ethnic or national origins.’ Case law has supported the inclusion of Romany gypsies and Irish Travellers as racial groups and the European Court of Human Rights have determined the Roma are also such a group. There is also some support in case law for the notion that religious belief can determine membership of an racial group and that membership of ethnic groups also extended to nationality. The definition of hostility is not addressed by legislation. But case law is more useful and judgements have suggested that there must be evidence of either spoken or written hostile words or hostile action. Hostility directed mistakenly at a person who is not a member of the racial group being denigrated, still counts as hostility under the law, and that it did not matter if other factors as well as racial ones lead to the hostility. Racial abuse does not require any dislike or ill feeling towards the group in question and the victim’s reaction to it is also irrelevant.

The CPS specifically mention ‘social media postings;’ as a form of evidence that may be part of a case concerning a racially motivated offence or offences.

The criminal law of England and Wales breaks down a criminal act into two parts: the *actus reus* and *mens rea.* The actus reus is the act that the defendant performed, and the mens rea is the intention to commit a crime. For a prosecution to be successful in England and Wales these must be proved beyond reasonable doubt.

In England and Wales legislation related to racial hatred, initially framed such offences as public order offences. Relevant legislation includes the *Public Order Act 1986*, which proscribed the stirring up of racial hatred. The *Public Order Act 1986* was amended by the *Racial and Religious Hatred Act 2006* to expand the protections already afforded to racial groups to religious groups. The *Crime and Disorder Act 1998* added to the public order offences of the *Public Order Act 1986*, a number of new racially aggravated criminal offences including harassment and threatening behaviour. There has to be a public element to these offences, and the Crown Prosecution Service (CPS) has determined the internet to be a 'public space'.

As well as public order legislation, England and Wales has laws aimed at preventing 'malicious communications'. The *Criminal Justice and Police Act 2001* amended the *Malicious Communications Act 1988* to make illegal the sending of any electronic message intended to cause distress. The police in England and Wales pass cases they investigate onto the Crown Prosecution Service (CPS) who handle prosecution. Racial group is defined under the *Crime and Disorder Act 1998* s 28 94) as: 'a group of persons defined by reference to race, colour, nationality (including citizenship) or ethnic or national origins.' Case law has supported the inclusion of Romany gypsies and Irish Travellers as racial groups and the European Court of Human Rights have determined the Roma are also such a group. There is also some support in case law for the notion that religious belief can determine membership of an racial group and that membership of ethnic groups also extended to nationality.

The definition of hostility is not addressed by legislation. But case law is more useful and judgements have suggested that there must be evidence of either spoken or written hostile words or hostile action. Hostility directed mistakenly at a person who is not a member of the racial group being denigrated, still counts as hostility under the law, and that it did not matter if other factors as well as racial ones lead to the hostility. Racial abuse does not require any dislike or ill feeling towards the group in question and the victim's reaction to it is also irrelevant. . Legislative handling of hate speech differs from country to country, as shown by the examples of Canada and Australia. The US and Europe have very different views on the regulation of racist hate speech. It is very difficult in the United States to prosecute for racist speech or for the use of racist symbolism, unless they are related to imminent harm or danger. In Europe, however, there is a very different attitude towards prosecution of racist speech and other expressions, and there is a history of racist speech legislation dating back to the 1960s.

There are three elements that are used to define hate speech: content, intent and harms. The content element includes offensive words and phrases, and symbols and iconography. The intent involves the promotion of hatred or other negative ideas towards a minority without there being any legitimate reason.The harm element is speech can cause either mental or physical harm to the victim or affects their self-esteem, social standing or affects them economically. Although this research focuses on *racist* rather than *hate* speech these three elements still apply, the only difference being that instead of them applying to *minorities*, that they apply to *racial or ethnic groups*. There is debate as to whether context is important or not, and whether tweets are descriptive or not. A key question regarding these messages, is whether what the tweeter writes is what they actually mean? If there is no link then this research is of greater value to industry than Law Enforcement, and it would have a social benefit aiding the regulatory/enforcement aspect of industry. If there is a link between language and belief, then, as mentioned, the research could help the police, by identifying racist tweeters, such people are likely to be of interest under the broken windows hypothesis of policing.

It is not possible to determine what a tweeter's intent was, although reasonable guesses can be made. However for certain words and phrases within tweets it is more problematic to determine their meaning, and for this research the word 'nigga' is particularly difficult to interpret.

The racist discourse seen in the texts in this research is what might be called *personal racism*, that is it is used to elicit a negative response in other people, presumably in the hope that the people who read it are shocked. It focuses solely on tweets which have an ethnic slur in them, i.e. overt racist discourse.

The regulation and governance of cyberspace, involves a number of stakeholders including nation states, public and private organisations and movements. Some argue that the traditional power of the nation-state has been eroded and in its place there has been a concomitant rise in the influence of organisations that are transnational and distributed in nature.

The UK government's online hate crime initiatives have been criticised as having very little detail on how they will work (Walters and Brown, 2016).

The next chapter introduces the key theoretical perspectives that underpin the research.

Chapter 4

Theoretical Background

In order to investigate racist tweets it will be assumed that they are instances of cybercrime (see Section 4.1) and so a criminological theoretical framework will be used. Such a framework must first consider what is meant by cybercrime and this is discussed in the next section. Then a brief summary of criminological theory is given followed by a discussion of the differences between situational and dispositional criminology. Within criminology, a particularly useful perspective is that of Routine Activity Theory (RAT), which is discussed in detail along with the similar Lifestyle Exposure Theory (LET). Then RAT's applicability to cybercrime is discussed. Psychological perspectives are also discussed as they can shed light on why people may sent racist tweets.

4.1 Cybercrime

The terminology relating to computer crime has had a complex history, with a number of competing terms covering different threads of computer-related crime. For example e-enabled crime, cybercrime and netcrime might all be used to describe crime mediated by the internet, however e-enabled crime might also include other crimes mediated by computers without a networking element (Bryant, 2008). Bryant (ibid.) uses the term *digital crime* and notes that

> we have deliberately chosen to extend the discussion beyond the realms of what may be termed ‘conventional cybercrime’ (despite the apparent inherent contradiction of the phrase) to other forms of criminality that exploit digital technologies to a lesser or greater extent. Hence we examine telecommunications fraud, video game piracy and ‘chip and PIN’ credit and debit cards, in addition to considering well-recognised problems such as cybercrime and internet grooming.

but he notes that this can be criticised as something of a catchall definition. Bryant’s (2008) aim was to analyse the policing of all forms of ‘digital crime’, and so a catchall definition was useful. For this research the focus is on one potentially criminal behaviour, that of sending racist tweets. So whilst this behaviour would fall under the umbrella of digital crime, it also falls under the narrower definition of *cybercrime*. Wall (2004) argues that the term cybercrime is meaningless since it has no legal or scientific basis, and instead is used ‘metaphorically and emotionally’ (ibid., p.2). Others agree that there is a certain amount of confusion regarding the term cybercrime, for example Hunton (2012, p.202) notes that

> [...] the concept of cyber criminality is often clouded by the interchangeable, inaccurate and even contradictory terms commonly used to describe the vast array of illicit activities and behaviours associated with cybercrime and cyber security.

Bryant (2014, p.23-24) suggests the reasons for this confusion are threefold. The first reason is that it is still the early stages of the transformation of society into a ‘Network Society’[1] and there will be concomitant changes and transformations in the forms of cybercrime. The second reason is that the striving for a universal definition of cybercrime is hindered by the fact that there is no universal consensus between countries as to what a legal definition of cybercrime should be. The third reason is that there are vested interests

[1] The idea of the ‘Network Society’ arose in the 1990s and is associated in particular with Manuel Castells (Castells, 2011) who argued that the rise of digital technology in the latter part of the 20th century, meant a shift in society towards horizontal networked organisation would emerge.

involved in reacting to the 'threat of cybercrime'. For example, it may be beneficial for law enforcement agencies to maximise the dangers of online activity in order to obtain resources to combat them, whilst it may be more prudent for others to downplay the risks if they wish to entice people to their websites. These voices can influence the discourse around what is and what is not cybercrime.[2]

Whilst a 'one size fits all' definition of cybercrime is difficult, if not impossible, the notion of cybercrime as crime mediated by the internet works well from a pragmatic perspective. In order to obtain a theoretical perspective that underpins research into racist tweets, the exact parameters of what is and what is not cybercrime, are not necessarily that important. Hence the term cybercrime will be used as the umbrella term for which this research falls into.

In the following sections, theoretical explanations of criminal activity will be explored, in particular in relation to cybercrimes.

4.2 Crime and deviance

Before explanations of crime and deviance are discussed, it is necessary to at least briefly consider what is meant by 'crime and deviance'. There is insufficient space here to focus on the debate about the different definitions of crime and deviance, and these topics have been discussed in great detail elsewhere. However it is informative to examine Gottfredson and Hirschi's (1990) definition of crime and deviance in relation to the sending of racist tweets. Their definition of crime was limited to 'acts of force or fraud undertaken in pursuit of self-interest' (ibid., p.15), which they delineated from deviant behaviour by the reaction to the act: there would be a state response to acts of crime, but deviance

[2] It can also be useful to consider the distinctions between *cyber-assisted*, *cyber-enabled* and *cyber-dependent* crimes (Wall, 2007; Wall, 2017). Cyber-assisted crimes are ones that use the internet in their commission, but could still occur without the use of the internet. For example a person may be stalked by information gleaned from their Facebook account. Cyber-enabled crimes are something of a halfway house between cyber-assisted and cyber-dependent crimes. They are crimes that could still occur without the use of the internet, but the internet affords them a force-multiplier effect. An example is the use of a phishing email allowing a fraud to occur at a much larger scale. Cyber-dependent crimes are ones that only occur on the internet, for example a DDOS attack.

would not generate a state response, although there may be an informal response from other actors. Their theory also attempts to include deviance along with crime; they talk of ‘crime equivalents’ or acts ‘analogous to crime’ which include smoking, drinking, drug use, gambling, having children whilst unmarried and engaging in ‘illicit sex’ (Gottfredson and Hirschi, 1990, p.90). Of course, their idea that ‘having children whilst unmarried’ is a ‘crime equivalent’ is seen as somewhat quaint in 2018, and this illustrates how crime and deviance are subject to shifts in societal norms and values. It also illustrates how the definition of crime and deviance varies between countries. With respect to racist tweets, Gottfredson and Hirschi’s definition will make them criminal in the UK but not in the USA. Even in the UK there is the possibility that they would not be seen as criminal acts under this definition, since only a small percentage of online ‘crime’ actually incurs a state response, due to law enforcement having insufficient capability to handle the vast number of acts performed on the internet.

Whether racist tweeting is a crime, deviant behaviour, or neither depends on many factors, but from a UK legal perspective (as discussed in chapter 3) it can be considered (potentially at least) a criminal act. So if it is accepted that racist tweeting is a crime, then theoretical explorations of such behaviour need to be examined, and such explorations will be discussed in the next section.

4.3 Criminological Perspectives on Crime and Deviance

Criminological theory is a vast and complex area and any review will struggle to synthesise such a large body of work and only a very brief discussion is given here.[3]

Classical criminology contrasted with *positivist criminology* in the philosophical notion of *free will.* Classical criminology’s theories assumed that people freely chose their actions based on rational decisions whereas positivist theories argued that actions were (somewhat at least) determined by features of an individual.

[3] There are many more in-depth criminological analyses available for further reading, for example Tierney and O’Neill (2013).

Classical criminology has been criticised since, for example, it ignores the possibility of non-rational humans, such as those suffering from mental illness committing crime (although *neo-classical criminology* allows for such possibilities). The idea of rational human actors underpinning classical criminology, are also the basis for a contemporary criminology that includes Lifestyle Exposure Theory (LET) and Routine Activity Theory (RAT) which are further discussed in Sections 4.3.3 and 4.3.4, respectively.

During the nineteenth century a contrasting view of crime arose: that of positivism, which was couched in scientific language and method, arguing that crime was the result of human biology. Positivism has been criticised as being deterministic, that is it contends that individual factors compel people to behave in certain ways, negating the role of free will. Biological positivism has largely been ridiculed but psychological positivism is still highly influential (Newburn, 2017).

Psychological positivism focuses on personality, psychology and learning and how these relate to criminal behaviour. Learning theories see behaviour as arising from interaction with others, in particular it is seen as the result of influence from close social contact with family and others. Sutherland's theory of *differential association* argued that it was the balance of association of an individual with groups whose norms were either deviant or non-deviant. Learning theories in general have been criticised as not explaining why some deviate yet others do not.

Biosocial theories such as Burgess and Akers's (1966) work adapted Sutherland's thesis and argued that learning is the result of *operant learning*, that is behaviour resulting in favourable outcomes will outweigh behaviour resulting in negative outcomes. This theoretical stance towards human behaviour also underpins Rational Choice Theory (RCT) and Routine Activity Theory (RAT) which are discussed further in Section 4.3.4.

Durkheim theorised that society contains a moral code known as the *conscience collective* and that crime is a normal part of society which can have positive functions, for example reinforcing the collective notion of 'right and wrong' (Durkheim, 1933). In his study of suicide, he argued that one of the motivations for suicide was that of *anomie*, a form of normlessness of those unsuccessfully integrated into society.

Merton (1938) adapted Durkheim's ideas with what became known as *strain theory*. In essence this theory stated that individuals that lacked means to a goal that they desired, suffered from strain resulting in anomie. He theorised five types of behaviour that result from this strain: *conformism* - those that are interested in the goals and have the means to reach them; *innovation* - those that want the goals but do not have the means; *ritualism* - those that are not interested in the goals but 'play along' anyway; *retreatism* - those that do not have the goals nor are they interested in obtaining them, instead they 'drop out' from society and *rebellion* - those that are interested in rebelling against the means and/or the ends (ibid.).

Agnew adapted Merton's work, arguing that in addition to being unable to reach goals, strain also included the removal of things of positive value as well as the presence of negative factors.

Strain theory in general has been criticised for ignoring the crimes of the powerful, instead focusing on those without the means to achieve certain goals. It has also been criticised as treating goals as a homogenous entity that everyone strives towards (Box, 2002, p.35).

Sykes and Matza's (1957) posited what they called *techniques of neutralisation* as justifications for deviant behaviour, including denial of responsibility, denial of injury, denial of the victim, condemnation of the condemners, and appeal to higher loyalties. All of these justifications have been seen in, for example, online music piracy (Cohn and Vaccaro, 2006).

Mead (1967) introduced the idea of *symbolic interactionism*, which saw the 'self' as created in a process of social interaction involving symbol interchange, in particular the use of language to signify things. It requires individuals to perform 'role-taking', that is putting themselves in the role of another person and interpreting the other's behaviour towards them and it is from this that the self emerges. The self is thus made up of a number of interpretations of others' views of the self's social roles e.g. father, student, and these are identities (Stryker, 1980). In pre-modern societies an over-arching identity dominated for an individual, but this is less the case for modernity where identities have become more fragmented (Bradley, 2015). The post-modern view of identity sees identities as fluid, and matters of choice (e.g. Bauman, 1996; Hall and Du Gay, 2006) but societal structures still have a constraining role on which identities are available to the individual (Bradley, 2015).

Labelling theory arose out of these symbolic interactionist ideas. Lemert (1951) introduced the idea of *primary and secondary deviance.* Primary deviance is the breaking of social rules on an occasional basis. Secondary deviance results from the social labelling that occurs when an individual repeatedly performs deviant behaviour. One of the key criticisms of labelling theory, is its lack of explanation of the deviancy that arises without labelling (Davis, 1972).

Control theories attempt to explain why people *do not* commit crime or behave deviantly, with the assumption that humans are predisposed to deviance, and will commit deviant acts unless they are subject to some form of control. One of the key criticisms of control theory is this assumption, the inevitability of crime is a philosophical position unsupported by empirical evidence (Newburn, 2017).

Other criminology schools of thought include: radical and critical criminology, realist criminology, feminist criminology, and late modernity, governmentality and risk and these have little bearing on this thesis and will not be discussed here.[4]

It is now useful to discuss dispositional and situational theories of crime, concentrating on two theories that are particularly relevant to this research: control theory which is a dispositional theory, and routine activity theory which is a situational theory.

[4] This is not to say that they cannot potentially be applied to cybercrime research, but it is necessary to focus on a small number of particular theories, otherwise the task of applying theory to research can become too unwieldy, and the work too convoluted.

4.3.1 Dispositional and Situational Theories of Crime and Deviance

It has long been argued that explanations of crime and deviance, (and the large body of criminological theory that has been discussed in the previous section), can be classified into two categories: those that take dispositional (or 'historical') stances and those that take situational stances (Sutherland (1947), (as cited in Birkbeck and LaFree (1993))).

Dispositional theories, try to answer the question why do *people* commit or desist from crime, from the point of view of a causal reason for lawbreaking.

Those that take a situational stance, see criminal disposition as a 'fact of life'; they are not interested in why a person commits a crime, other than how the situation the person finds them-self in is criminogenic.[5]

Hirschi and Gottfredson (1986, p.58) saw crime and criminality as two distinct concepts. They saw crime as an event comprising a number of necessary conditions, one of which is criminality. Criminality is the propensity of an individual to commit crime, and thus is a necessary condition for crime to occur but it is not sufficient since other conditions must be met, for example the finding of a suitable target (Birkbeck and LaFree, 1993). So for Hirschi and Gottfredson (1986, p.58) Sutherland's notion of dispositional theory would be theories about criminality, whereas his idea of situational theory would be part of the theories of crime (Birkbeck and LaFree, 1993).

4.3.2 Control Theory

Gottfredson and Hirschi's (1990) *A General Theory of Crime* introduced their self-control theory which deemed low self-control to be the primary factor for criminality. They saw individuals with low self-control as people who are 'impulsive, insensitive, physical (as opposed to being mental), risk-taking, shortsighted, and nonverbal' (ibid., p.90). Such

[5] Of course ideally all the circumstances around the behaviour would be considered; it is easy to over-interpret the circumstances of the act of sending racist tweets. The tweeter may not actually be racist, just having a bad day, and their anger causes them to momentarily use an ethnic slur.

individuals, in the authors' view, were less likely to exhibit self-control as a result of one or more of: impulsivity causing them to ignore the negative consequences of their acts, insensitivity meaning they have fewer possible negative outcomes to consider and a lack of intelligence meaning they have less to lose. They argue that the traits that make up an individual with low self-control, are conflated with one another and are all seen together in an individual.

In relation to the efficacy of control theory explanations of cybercrime, much of the early research investigates the link to digital piracy (Donner et al., 2014). To broaden the investigation of control theory's relevance to cybercrime Donner et al. (ibid.) intended to widen its study to other forms of cybercrime, surveying 488 criminology and social science undergraduate students, as to whether they had committed one or more of 10 deviant online acts. Of these 'threatening/insulting others through email or instant messaging' (ibid., p167) was the most closely related to this research. They performed a correlational analysis of counts of these behaviours against a variable measuring low self-control, which was formulated using a 24 item attitudinal measure. Of the 10 behaviours only the email/instant messaging harassment and the posting of hateful information were highly significantly correlated with low self-control.

Vazsonyi et al. (2012) also analysed the effect of low self-control in relation to cybercrime. They used data from the EU Kids Online II study, which randomly sampled a minimum of 1,000 youths, aged 9 to 16, in 25 European countries (N=25,142) to investigate the links between low self-control and cyberbullying. They found low self-control correlated with both online and off-line bullying, with the effect on online bullying the greater. If the example of racist tweeting is examined in the light of this theoretical viewpoint, it can be seen that control theory explains well why tweets might be sent 'in heat of the moment'. Tweeters with low self-control are more likely to use extreme expressive language, perhaps when enraged, and the ease and routine nature of tweeting means that such individuals can easily express such angry language. However control theory falls down when considering racist tweets from, for example, far right organisations or individuals. Racist tweets from such sources are likely to be part of their everyday worldview, rather than results of low self-control.

Control theory also has a monolithic view of culture and gender, attributing differences in offending between males and females, and between cultures as largely due to differences in self-control. Cheung (2013) in his study of online music piracy by adolescents in Hong Kong, found no difference in levels of piracy between male and female adolescents, contrary to what would be expected under control theory.

These studies show some, albeit limited, support for the idea that low self-control is correlated with cybercrime. Indeed, it can be theorised that control theory may give at least a partial explanation for why some individuals might be predisposed to criminality and thus more likely to send racist tweets. However, control theory does not consider the causal conditions under which such an event occurs. Since the overall philosophy of this research is pragmatism, with a desire to automatically spot racist tweets and tweeters, a theoretical framework that can be used to analyse the event itself, as opposed to just the offender's motives is needed. The dominant theory that is used to analyse the components that make up a criminal event, is that of RAT. The focus of Routine Activity Theory (RAT) is very much situational; theorists (certainly prior to the advent of the internet and explanations of cybercrime) look at how the spatial and temporal situations and actors find themselves in, affect the criminal event (Yar, 2005).

4.3.3 Lifestyle Exposure Theory

Prior to the development of RAT, Hindelang et al. (1978) examined data from victimisation surveys in eight major American cities collected in 1972 by the US Bureau of the Census. Their aim was to develop a theory of personal victimization that could be applied to rape, robbery, assault, and 'larceny'. Their *Lifestyle Exposure Theory* (LET) argues that *lifestyle* is the most important factor in the likelihood of victimisation for these crimes. Lifestyle comprises 'routine daily activities, both vocational activities (work, school, keeping house etc.) and leisure activities' Hindelang et al. (ibid., p.241). The theory contends that the observed differences in victimisation rates between different demographic categories are the result of personal lifestyle differences. Daily activities can bring people into situations that are more likely to lead to criminal events, lifestyles that have higher exposure to dangerous locations, and at more dangerous times of day are likely to lead to higher chance of victimisation. For Hindelang et al. (ibid.) structural constraints and role expectations influence people's choices of lifestyle. Status attributes such as gender and income have concomitant expectations of behaviour and structural obstacles to overcome. Individuals that conform to these expectations, come into contact with others of similar lifestyle choices, and routine activities and patterns of behaviour result in similar exposure to crime (Meier and Miethe, 1993).

There is considerable overlap between RAT and LET. Both theories contend that crime victimisation is highly dependent on routine patterns of activity, and neither theory considers motivational aspects of offenders' behaviour. Their main difference is that they were developed to explain different things: RAT 'was originally developed to account for changes in crime rates over time whereas lifestyle exposure theory was proposed to account for differences in victimisation risks across social groups' (ibid., p.470). Although much of the theorising on cybercrime uses RAT without reference to LET, Choi's (2008) formulation of Cyber Routine Activity Theory (CRAT) focuses on LET. RAT is further discussed in the next section.

4.3.4 Routine Activity Theory and Environmental Criminology

Up to the time of RAT most criminological theories were dispositional in nature, attempting to explain why people offend, but RAT, shifted the focus onto situations that are conducive to criminality, with the assumption that crime and criminality will happen under the right circumstances, ignoring the question of offender motivation, except as an element of a cost benefit analysis on their part (Cohen and Felson, 1979). RAT and Situational Crime Prevention (SCP) arose as critiques of dispositional theories and their attention to a small number of individuals disposed to commit crime, despite the fact that research of the time suggested that most crimes were actually committed by people who were not normally thought of as criminal, and dispositional theories' lack of focus on the situational aspects of crime (Clarke, 1980). Cohen and Felson (1979) were interested in explaining how despite structural improvements in the United States, (including increased income levels, and education), crime levels had also risen from 1960 to the late 1970s (Cohen and Felson, 1979). They theorised that the routine activities of an individual within a particular environment influenced their criminal opportunities within that environment, thus affecting the prevalence and incidence of *direct contact predatory violations*, which they defined as 'illegal acts in which 'someone definitely and intentionally takes or damages the personal property of another' (Glaser (1971, p.4), as cited in Cohen and Felson (ibid.))). They argued that it was not structural changes in society that caused the increase in crime rates, i.e. there were not more offenders, but instead changes in everyday routines as a result of systemic changes in the rapidly growing and industrialising post-World War II US economy, resulted in more opportunity for crime.

Cohen and Felson argued that 'criminal acts required the convergence in space and time of likely offenders, suitable targets, and the absence of capable guardians,' (ibid., p.588), that is the coming together in a particular place and time of these three factors are the necessary conditions for a criminal act. A crime is more likely to occur when these three factors converge, and the absence of any of these means a crime is unlikely to occur. Cohen and Felson (ibid.) argued that crime rate trends are affected by changes in 'routine activities' in the day-to-day lives of criminals. RAT suggests that the interaction and organization of events, actors and activities explains crime 'rate trends and cycles' (ibid., p.588). They explicitly stated that they were not interested in why or whether certain people were or were not criminally inclined but instead assumed a universal criminal inclination. Their intention was to:

> examine the manner in which the spatio-temporal organization of social activities helps people to translate their criminal inclinations into action (ibid., p.589).

For Cohen and Felson (ibid., p. 593) these social activities or *routine activities* were 'recurrent and prevalent activities which provide for basic population and individual needs, whatever their biological or cultural origins', which included any day-to-day activities which could include work, leisure, the provision of food etc. Although the use of the word, 'basic' is perhaps misleading since they allowed for activities that go beyond this as long as they were everyday activities. They stressed that a criminal event occurs at a specific *time* and *place*. One of the underlying ideas of RAT and its theorists is that crime is not a random event, instead it occurs in a disproportionate amount in areas that are criminogenic, because they either have a high number of motivated offenders, have a high number of suitable targets or lack capable guardians, or some combination of these. Like most criminological theories, the original formulation of RAT has a undergone a number of transformations resulting from the work of different theorists. Eck and Madensen (2015) provide a summary of the evolution of RAT as can be seen in figure 4.1. They argue that RAT has gone through four stages of development: RAT 1 - The original theory, RAT 2 - Handlers as controllers, RAT 3 - Places and place management and RAT 4 - Controlling controllers. The first diagram in figure 4.1 shows the original formulation of RAT, with offender, target and guardian. The guardian acting as a controller in relation to the target, to stop the offender committing a crime when they encounter a suitable target. In the second diagram (RAT 2) the *handler* has been added. The concept of the handler was added to RAT by Felson (1986), the handler being another controller, but this time in relation to the offender rather than the target. Emotional and social bonding with the offender and the handler means the offender is less likely to commit crime. In the third diagram, RAT 3 shows the addition of place and place management, with Sherman et al.'s (1989) RAT analysis of calls to police in Minneapolis suggesting that most calls, and therefore crime, arose from a small number of hotspots, stressing the importance of place in RAT. Eck (1994) (as cited in Eck and Madensen, 2015) suggested the idea of *place controllers*, that is the owners (or people delegated by the owners) of the place. Madensen (2007) in his study of bar management theoretically analysed this idea of place management and how it can help to prevent crime. This new formulation with handler, guardian and manager as controllers of offender, target and place respectively, is what Eck and Madensen (2015) denote as RAT 3.

RAT 4 has an additional layer beyond the controllers of *super controllers*. Sampson et al. (2010) introduced this concept in order to explain why crime still occurs with handlers, guardians and managers in place. They argued that all these types of controllers belong to networks known as 'super controllers', these networks provide incentives for controllers to reduce criminal activity. Super controllers arise from social, political and economic

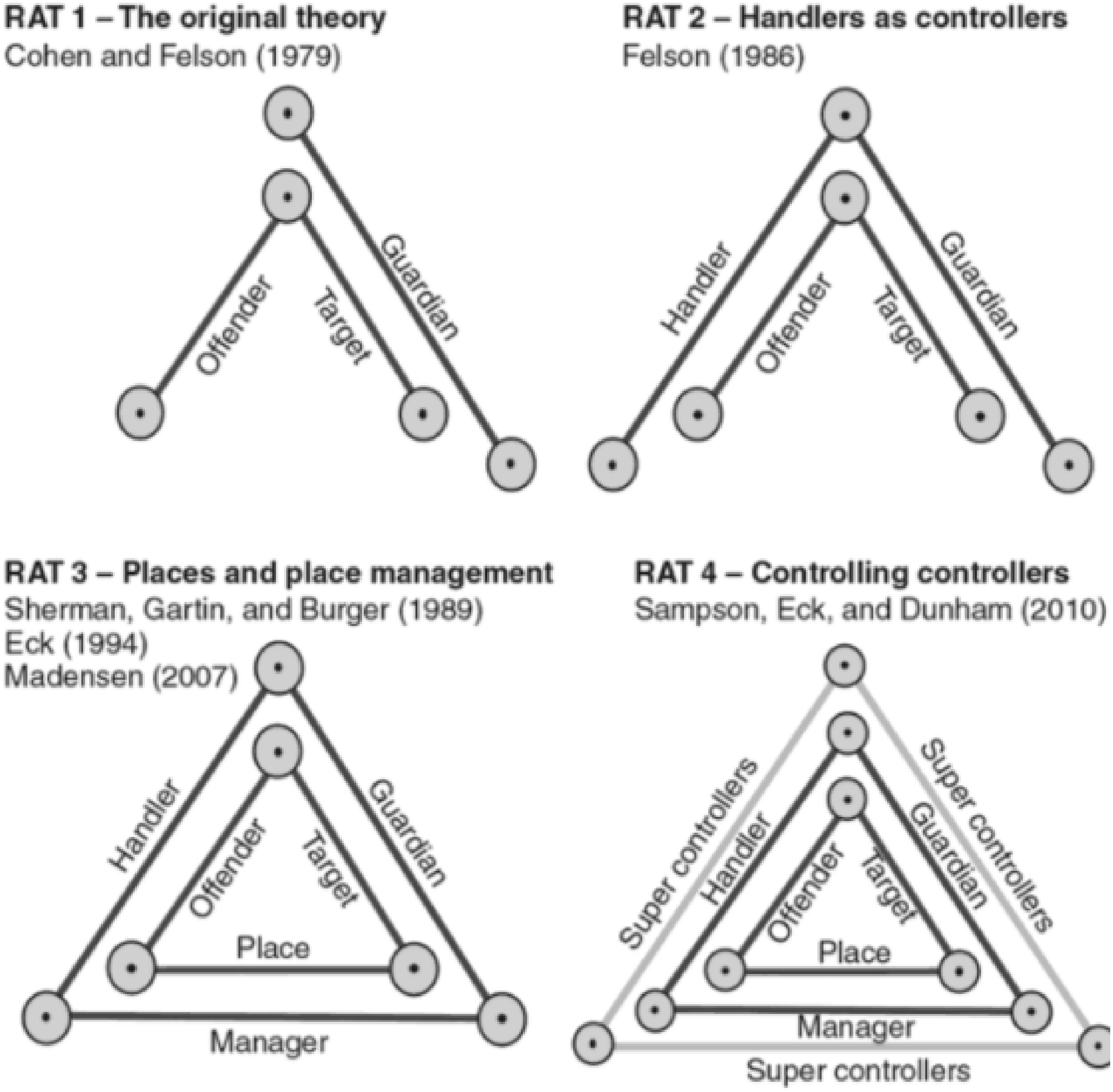


Figure 4.1: The development of routine activity theory, from Eck and Madensen (2015).

structures and would include entities such as families and friends and the relationships within them, the criminal justice system including courts and forms of deterrence and financial institutions such as banks and insurance companies.

RAT can be criticised as it ignores the disparate levels of victimisation experienced by, for example, different genders. Being female is not a routine activity, yet females are disproportionately victimised and RAT does not explain how they make 'better' targets than males (Finkelhor and Asdigian, 1996). Jeffery (1993, p.492) argues that the theory is lacking in explanatory power, completely ignoring the 'why' question of crime, and is, he argues, a 'description of crime not an explanation'.

Another criticism levied at RAT, is its underlying theoretical view of human agency. RAT is often tied to Rational Choice Theory (RCT, Hechter and Kanazawa, 1997), especially the RAT work of Felson (for example, Clarke and Felson, 1993). RCT posits that humans are rational actors that make cost-benefit choices, and that criminal acts result when opportunities present themselves; the actor considers the reward gives them greater benefit than the risk and effort involved costs them (Cornish and Clarke, 1986). Yar (2005) notes that the reliance on RCT, potentially makes RAT a poor choice for expressive and other non-instrumental crimes. He argues, however they even these kinds of crimes can be said to have a rational element since, emotional responses may be rational reactions to situations. Perhaps the main criticism of RAT is that is difficult, if not impossible, to measure its main components: the routine activities of people and whether targets are suitable and guardians are capable, and Eck (1995) notes that normally macrolevel aggregate data is used to test what is, in fact, a microlevel event.

Despite these criticisms, RAT, in its various guises, has been successfully applied to a number of real-world settings, and its simplicity lends itself to application by law enforcement (Hollis et al., 2013). These applications are usually in the form of what is known as Situational Crime Prevention (SCP). RAT, RCT and Crime Pattern Theory (CPT) provide the theoretical basis of SCP, which focuses on ‘suitable targets’ and their attractiveness to offenders. Crime prevention strategies that have evolved from SCP theory, focus on ‘target hardening’, that is making targets less attractive to offenders (Clarke, 1980; Clarke, 1983; Felson and Clarke, 1998). Felson and Clarke (1998) proposed a model of target attractiveness corresponding to the acronym *VIVA (value, inertia, visibility, and accessibility).* They theorised these four dimensions of a target’s attractiveness as attributes which are rationally analysed and observed by offenders, and are key components in the offender’s rational consideration, as to whether a target is suitable for them.

The first dimension *value* is a measure of how much value an offender gives the target or goal, Felson and Clarke (ibid.) giving the example of the theft of CDs (which at the time were likely to have had significant monetary value). It does not necessarily have to be a monetary value; it can be a real value or something symbolic (Miró, 2014).

Inertia is a factor that signifies the difficulty of removing an item, according to Felson and Clarke (1998), it is merely the object's weight, and they give the example that small electronic goods are more likely to be stolen than weighty items, due to their lower inertia. However this statement of inertia of an item merely being its weight is somewhat limiting, since, for example, the difficulty of taking an item can also relate to its volume since relatively lightweight objects that are large, are also difficult to steal. Indeed Cohen et al. (1981, p. 508) uses a more expansive notion of inertia, for them it is 'the weight, size, and attached or locked features of property inhibiting its illegal removal and the physical capacity of persons to resist attack'. *Visibility* obviously relates to how visible an item is to an offender, Felson and Clarke (1998), give the examples of 'when someone flashes money in public or puts valuable goods by the window.'

Accessibility refers to the geography of a location, and layouts of buildings and rooms, where items are placed and any other factors that interact with the ability of offenders to access items. The applicabilty of VIVA to cybercrime is discussed in Section 4.3.5.

Brantingham and Brantingham's work on environmental criminology expands on these ideas of accessibility. They look at how targets are distributed spatially, and how offenders interact with their environment, and other factors of that environment, and how these interactions can determine locations of criminal activity (Brantingham and Brantingham, 1981).

They postulate the notions of geometry of crime and crime pattern theory. Geometry of crime examines how the structure of cities influences the distribution of crime. They note that all human behaviour is related to a backcloth which is 'composed of social, economic, political, and physical dimensions' (Brantingham et al., 2016). Choices are impacted by this backcloth, and activities constrained and determined within a number of influences, be they social pressures to conform or rebel, economic restrictions and political and physical landscapes of society. Within this backcloth they note that activities are performed within what they term as an *activity space*, this is a series of *nodes* that individuals are familiar with, including home, work and leisure destinations and a series of *paths* which are the usual routes they take between the nodes. The area that is visible from the activity space is termed the *awareness space*. They argue that criminals are like anybody else, in that they perform most of their activities within their activity and awareness spaces, so are more likely to commit crimes around familiar nodes and along familiar paths, particularly when these intersect with target activity and awareness spaces. Popular areas such as shopping centres attract motivated individuals and provide them with opportunities by attracting people who act as targets (Brantingham et al., 2016; Brantingham and Brantingham, 1993; Brantingham and Brantingham, 1981). They note the need to map these ideas to what they call *hyperspace*, indeed they argue that the

> ideas in environmental criminology in general, and Routine Activity Theory and its associated enhancements in particular, need to be transformed into a not-limited-to-physical, n-dimensional hyperspace where *adjacency* and *convergence* do not refer simply to *physical* adjacency and *physical* convergence. (Brantingham and Brantingham, 2015, p.135)

They argue that the various concepts behind RAT and environmental criminology, such as routine activities, activity space and so on still can be applied when considering hyperspace instead of the physical world. There are of course changes, networks of activity are far more connected in hyperspace than physical space, and there is more overlap between activity spaces such as the overlap between Twitter, Facebook and other social media sites.

Theoretical analyses of cybercrime are now explored.

4.3.5 Routine Activities and Cybercrime

According to RAT, proximity to offenders leads to higher risk of victimisation in the real world, however there is difficulty with applying this to cyberspace, where the notion of proximity is not as straightforward as it is in the physical world. Anyone is, in some sense, at the same place as anyone else on the internet, where everything is potentially 'just one click away' (Yar, 2005). Yar (ibid.) however has identified some spatial similarities between cyberspace and the real-world in that at the time of his writing much of the internet use was from the United States, and that cyberspace can be mapped to the geographical world with similar disparities caused by economic and social hierarchies in the world. He also notes that despite there potentially being no distance between anybody on the internet, there are also can be barriers to access certain sites and structures on the internet, some sites are password restricted, others are hidden behind pay walls, others, such as those in the darknet are very difficult to reach for the everyday user.

Despite these geographical mappings there is still a discord between real-world RAT and its application to cyberspace. Some such as Geer (2007) argue that all locations on the internet are equidistant. However this does not take into account as already mentioned, barriers to access, such as the requirement to use the Tor browser to access the darknet. Although there is certainly a difference between RAT as formulated for the terrestrial world, and the environment in cyberspace, it does not immediately exclude the use of RAT to investigate cybercrime. Whilst RAT requires motivated offenders and suitable target to meet in a location, the collapsing of physical space, as seen on the internet, means that, a different version of RAT can be utilised, where the geographical component is no longer required, since there is potential on the internet from motivated offender to always be in the same proximity as one, (or more likely many), suitable targets.

Yar (2005) discusses how well RAT applies to cybercrime. He noted the novelty of social interactions afforded by the new computer-mediated communication (CMC) technologies, chiefly the internet. Interactions on the internet are different from ‘real-world’ interactions due to 3 main factors:

> the collapse of spatial-temporal barriers, many-to-many connectivity, and the anonymity and plasticity of online identity) that make possible new forms and patterns of illicit activity (ibid., p.411).

Leukfeldt and Yar (2016) reviewed the empirical tests of RAT’s applicability to cybercrime[6] by examining 11 studies and performing one of their own. These are summarised

[6] There is an issue of the spectrum of cybercrimes (Gordon and Ford, 2006) being treated as a ho-

in Table 4.1 which is adapted from their work, with the inclusion of their study's data. The table consists of five columns: the name of the study, the population of participants, which cybercrimes the study investigated, the parts of RAT that were measured, and how well RAT explains victimisation. Of the 12 studies half were conducted on college students, one on Australian residents, one on Floridian adults, two on Dutch households, and two on representative samples of the Dutch population aged 15 and above. The sample sizes ranged from 104 to 9,161. Each study examined one or more aspects: value, visibility and accessibility from the acronym VIVA (none of the studies measured inertia). The +/- column in Table 4.1 denotes which element of VIVA (and therefore RAT) explained victimisation. The results of this column were largely negative, that is elements did not explain RAT, but there were some exceptions. Of the studies finding positive results all found visibility to have good explanatory value for RAT. In addition Marcum et al. (2010) in their study of undergraduate internet usage, found that accessibility explained victimisation as did the survey of Dutch residents on victimisation by malware, hacking, identity theft, consumer fraud, stalking, and threats (Leukfeldt and Yar, 2016). Also Choi (2008)'s study of the experience of college students' victimisation by computer viruses also suggested that capable guardianship explained victimisation.

The application of VIVA to cybercrime will now be discussed.

4.3.5.1 Value

Even in the terrestrial sphere the value of an an object or goal is a complex function of what an offender has in mind. They may value an object since they may wish to keep an object for themselves, they may wish to sell it they may wish to give it away, they may wish to use it for something else and so on. Values also rise and fall according to external forces such as fashion, scarcity and public opinion. Such forces are also applicable to the online world as evidenced by the rise and fall in the value of crypto currencies. In the digital world value is often in the form of information, which can have a monetary

mogenous type of crime. It needs to be considered whether there is any similarity between the sending of racist tweets and, for example, obtaining bank details by a phishing attack. If there is no similarity then tests of RAT that show good explanation for certain types of cybercrime, may not have any validity for other types.

Table 4.1: Empirical studies applying RAT to cybercrime, adapted from Leukfeldt and Yar (2016).

Study	Population	Cybercrimes	RAT[*]	+/-[**]
Choi, 2008	N = 204, college students	Computer viruses	Va: No	+
			In: No	+
			Vi: Yes (3 items)	
			Ac: No	
			Cg: Yes (2 items)	
Bossler and Holt, 2009	N = 570, college students	Malware infection	Va: No	-
			In: No	-
			Vi: Yes (11 items)	
			Ac: No	
			Cg: Yes (13 items)	
Hutchings and Hayes, 2009	N = 104, residents of Brisbane Metropolitan area	Phishing	Va: No	-
			In: No	-
			Vi: Yes (3 items)	
			Ac: No	
			Cg: Yes (6 items)	
Holt and Bossler, 2009	N = 578, college students	Online harassment	Va: No	-
			In: No -	-
			Vi: Yes (9 items)	
			Ac: No	
			Cg: Yes (8 items)	
Marcum et al., 2010	N = 744, college students	Sexually explicit material	Va: No	+
		Non-sexual harassment	In: No	+
		Sexual solicitation	Vi: Yes (6 items)	-
			Ac: Yes (4 items)	
			Cg: Yes (5 items)	
Pratt et al., 2010	N = 992, adults in Florida	Consumer fraud	Va: No	+
			In: No	
			Vi: Yes (2 items)	
			Ac: No	
			Cg: No	
Ngo and Paternoster, 2011	N = 295, college students	Computer virus	Va: No	-
		Harassment (non) stranger	In: No	-
		Unwanted pornography	Vi: Yes (4 items)	-
		Sexual solicitation	Ac: Yes (3 items)	
		Phishing	Cg: Yes (3 items)	
		Defamation		
Reyns et al., 2011	N = 974, college students	Cyberstalking	Va: No	-
			In: No	-
			Vi: Yes (5 items)	-
			Ac: Yes (3 items)	
			Cg: Yes (3 items)	
Van Wilsem, 2011a	N = 4,353, on-line panel, repr. for Dutch households	Threat	Va: No	+
			In: No	
			Vi: Yes (7 items)	
			Ac: No	
			Cg: No	
Van Wilsem, 2011b	N = 6,201, on-line panel, repr. for Dutch households	Consumer fraud	Va: No	+
			In: No	
			Vi: Yes (6 items)	
			Ac: No	
			Cg: No	
Leukfeldt, 2014	Repr. for Dutch population (15+) (N = 8,379).	Phishing	In: No	-
			Vi: Yes (12 items)	-
			Ac: Yes (6 items)	-
			Cg: Yes (3 items)	-
Leukfeldt & Yar, 2016	Repr. for Dutch population (15+) (N = 9,161).	Malware infection	Va: Yes (4 items)	-
		Hacking	In: No	
		Identity theft	Vi: Yes (12 items)	+
		Consumer fraud	Ac: Yes (6 items)	+
		Stalking	Cg: Yes (3 items)	-
		Threats		

[*] Parts of RAT that have been measured. Va = Value, In = Inertia, Vi = Visibility, Ac = Accessibility, Cg = Capable guardian. The number of items correspond with the variables shown in the final analysis. If, for example, authors present a construct of three variables, this is seen as one variable.

[**] + means this part of RAT (largely) explains victimization, - means this part of RAT (largely) does not explain victimization.

value, for example pirated movies which have value since the downloader will want to watch them, and they may reduce revenue for the producers of the films. Value might not be explicitly monetary related however, computer systems can also be targeted by hacking and/or malware, to gain unauthorised access to take or damage information. They may be used to obtain or trade illicit material such as Child Sexual Abuse Images (CSAI). They may be used for stalking, or to abuse people based on some characteristic, such as their ethnicity. They may be used to broadcast messages of hate (Yar, 2005). All these activities have value of some kind for offenders, and this wide array of values with its concomitant wide array of targets, shows the difficulty of treating cybercrime as one homogenous entity. A racist tweet might have value to the tweeter for a number of reasons. It might provide a release for their anger and frustrations. It might give them a sense of security, that attacking the outgroup can provide. The tweeter might be trying to impress or repel other users. It might signify a 'call to arms' or have content that is intended more directly to mobilise others and so on.

4.3.5.2 Inertia

In terrestrial form inertia is the inherent resistance of an object or person in becoming a target. Large and heavy objects are harder to steal, and larger people tend to be harder to assault, or at least are more likely to give this impression to a potential assailant. Data in the cyber sphere has no weight or physical size, but file sizes can be limiting factors in, for example downloading. Somebody illegally downloading a film, may give up if the file size is too large since it may take a considerable time to download (ibid.). There are also limits to storage, although the ubiquity of cheap large capacity hard drives, means this is not limiting factor it was at the time of Yar (ibid.)'s writing.[7]

[7] Films downloaded from torrent sites vary in size, but often are around the 700 MB size, meaning nearly 1500 films would fit onto one 1 TB drive.

4.3.5.3 Visibility

According to RAT people and property that are visible to offenders are more likely to become targets and Yar (2005) argues that the visibility on the internet is virtually unlimited since there are no physical barriers. He also argues that the public nature of the internet is important in this visibility. While this is undoubtedly true for, for example, Twitter, where messages are by default public, it is questionable whether this applies to, for example, Facebook, where messages may be public, but also much more commonly restricted to a small number of 'friends'. It can be argued that, while indeed there are no physical barriers by default on the internet, there certainly are barriers created by websites, pay walls and other limits to visibility. In a further test of the visibility dimension of RAT, Hawdon et al. (2017) studied youths and young adults in Finland, Germany, the UK, and the US in relation to their exposure to online hate materials. They surveyed young people in four countries about whether they had been exposed to online hate within the past three months. They found that the US had the highest exposure to hate with 53%, and Germany the lowest with 30%. They performed a regression analysis and found that the effect of using multiple SNS and dangerous sites were significant predictors but online anonymity and online trust were not. Age was significant in the US and UK and gender was only significant in the US. They concluded that their results gave support to RAT, since the visiting of dangerous sites and SNS meant more exposure to hate materials.

4.3.5.4 Accessibility

In the terrestrial world accessibility is the 'ability of an offender to get to the target and then get away from the scene of the crime' (Felson and Clarke, 1998, p.58). So it is related to the human geography of an area; a house on a busy street is more likely to be seen by a burglar and therefore targeted (although possibly business might act as a deterrent). In cyberspace there are no analogous routes to and from the 'scene of a crime' (Yar, 2005). However accessibility in the digital realm varies with factors such as search engine ranking, signposting by social media and other communications, barriers to access and so on.

As well as VIVA, RAT's notion of 'capable guardianship' needs to be considered in relation to cybercrime. A guardian in the physical world is a person or object, whose presence is a deterrent to criminal behaviour, if they are absent then a location becomes much more criminogenic. Presence includes an actual physical presence, or the presence of some direct effect on an area performed by the guardian such as monitoring by CCTV cameras. Guardians may be formal law enforcement agents such as the police, although RAT usually focuses on informal guardians, 'ordinary citizens' whose routine activities mean that they are present in places and so have a deterrent effect (ibid.). This emphasis is due to the fact that in reality most policing is by informal guardians, and the police is often called as a last resort (Felson and Clarke, 1998, p.58).[8] For human guardians RAT suggests they merely have to be present in a location.[9] Guardianship on the internet is also largely the work of informal guardians including

> in-house network administrators and systems security staff who watch over their electronic charges, through trade organizations oriented to self-regulation, to 'ordinary online citizens' who exercise a range of informal social controls over each other's behaviour (Yar, 2005, p.423).

Guardianship in RAT is not limited to human guardianship but also includes physical and technological measures, such as locks, CCTV, alarms, street lights and so on. In the virtual realm there are analogues to these physical guardians, including virus and intrusion checking software, firewalls, government monitoring systems such as those employed by GCHQ and the CIA (ibid.) to automated detection systems that can detect illicit visual and textual material, such as Child Sexual Abuse Images (CSAI) (Sae-Bae et al., 2014) and Twitter's efforts at identifying and blocking racist tweets and tweeters.

[8] More formally policing can be thought of in terms of *primary, secondary and tertiary social control.* Primary social control consists of the policing activities carried out by organisations and individuals for which social control is their main focus, for example the work of the police and private security companies. Secondary social control is exerted by people whose main role is not that of social control, but social control does play a part in their activities. Examples of occupations that include a degree of secondary social control are teachers, caretakers and so on. Tertiary social control is the informal control which arises from groups such as churches and clubs, or from any social or familial group (Jones and Newburn, 2002).

[9] Indeed it has been suggested that mere likenesses of people have a similar deterrent effect, and so cardboard cutouts of policeman are placed at the entrance to shops.

Choi (2008) argues that RAT expands upon LET, in that the lifestyle variables of LET are the target suitability aspect of RAT. He notes that routine activities in cyberspace, related to both work and leisure, can increase the chances of an individual's victimisation. He argues that differences in online lifestyles and digital guardianship are correlated with online victimisation. He developed a theory that integrates the concepts of LET and RAT to computer crime victimisation, which he later calls called Cyber-Routine Activities Theory (CRAT, Choi and Lee, 2017). Choi and Lee (ibid.) used CRAT as the theoretical basis of an investigation into online victim offender overlap. They looked at cyber harassment and cyber impersonation by surveying 272 American college students in 2014. With regards to cyber interpersonal violence they were asked questions which operationalised their experiences of online victimisation, offending and risky online activity. Choi and Lee (ibid.) performed a logistic regression on their answers and found that risky online behaviours and lack of adequate cyber security did indeed mean, as theorised under CRAT, that the occurrence of cyber-interpersonal violence victimisation was more likely. In regards to offending the only behaviour that had significant relationship with offending was that of risky social network behaviour which was significantly positively correlated with engagement in cyber-interpersonal violence.

This thesis takes the lead of Yar (2005) and adopts RAT in its original form i.e. RAT suggests

> that structural changes in routine activity patterns can influence crime rates by affecting the convergence in space and time of the three minimal elements of direct-contact predatory violations: (1) motivated offenders, (2) suitable targets, and (3) the absence of capable guardians against a violation. (Cohen and Felson, 1979, p.589).

So, in simple terms RAT suggests that one of the reasons crimes are likely to occur is that there is no capable guardian present. Although Twitter is attempting to police racist tweets, whether there is currently a capable guardian present on Twitter is debatable. In any case the racist tweeters would need to be aware that there is a guardian for RAT to apply. There is some evidence of this, as shown in Figure 1.1 the tweet 'jack and twitter isn't going to like this' which was a reply to a tweet containing a number of racial epithets and swearing, suggests an awareness of surveillance by Twitter. In 2017 there was publicity regarding Twitter's attempts at curbing racist and other extreme language and behaviour on its platform, which this is presumably a reference to.

In using the RAT theoretical framework, it is hoped that the result of this research, that is an efficient and reliable system for automatically identifying both racist tweets and tweeters, would be a first step in capable guardianship on Twitter, or similar online communities, and its publicity within the platform, such as warning messages sent to known racist tweeters prior to their sending of tweets, might act as a deterrent, and thus limit the amount of offensive material on Twitter. Twitter's current attempts at blocking offensive material seem to consist of a temporary block on certain accounts, although it is not clear which accounts. The methods used to identify them are also not divulged, and some accounts seem to be evading this blocking. While RAT is the main theoretical foundation of this thesis, it is also important to gain some insight into the psychological dimensions of racist tweeting.

4.4 Psychological Perspectives on Online Offending

Up to this point, the theoretical perspectives that have been examined, are largely sociological in nature. Now psychological evaluations of online offending will be discussed.[10]

4.4.1 Disinhibition

It has been theorised that the internet has a disinhibiting effect on users, with Danet (1998, p.131) noting that the

> anonymity and dynamic, playful quality of the medium have a powerful, dis-

[10] Since the main focus is criminology and due to space restrictions, only the psychological theoretical work that is most relevant to this thesis will be discussed.

> inhibiting effect on behavior. People allow themselves to behave in ways very different from ordinary everyday life, to express previously unexplored aspects of their personalities, much as they do when wearing masks and costumes at the carnival or a masked ball.

Suler (2004) expanded on this noting an 'online disinhibition effect' which has two manifestations: *benign disinhibition* which includes behaviours like revealing intimate information and *toxic disinhibition* which includes expressions of hatred. It is this toxic disinhibition that is relevant to this research, however the evidence for whether this disinhibition is a factor in online behaviour is, at best, mixed (Bryant, 2014). Suler (2004) gave six factors he thought were involved in the cause of online disinhibition and Bryant (2014) added one more. These are shown in Table 4.2. The theoretical notion that dissociation and disinhibition has an effect on criminal behaviour, is relevant to this research. It might explain, for example any correlation between the amount of racist tweets and the time of day, that differs from everyday tweeting. If, say, there are proportionally more racist tweets around the hours of 11 PM to 1 AM, and it may be that alcohol is having a further disinhibiting effect on the tweeters. Further discussion of this is given in Chapter 8.

4.4.2 Anonymity

One of the reasons for the Internet having a disinhibiting effect on users is the anonymity[11] (both perceived and real) it affords.

The relationship between anonymity and whether a person directs abuse towards another on the Internet is complex. Wulczyn et al. (2017) investigated anonymity and online personal attacks using Wikipedia comments data and a machine learning classifier. They found that although each individual anonymous comment was six times as likely

[11] Both Beyer (2012) and Qian and Scott (2007) see anonymity on the internet in the form of an 'anonymity continuum' with identities ranging from fully anonymous through to clearly identified. Nagel and Frith (2015) regard this continuum as limiting and instead discussed anonymity practices that are more fluid and include the use of pseudonyms, mononyms, stage names, anglicised names, user names that avoid using a real name or that partly incorporate a real name and interactions between these types of identity labelling.

Table 4.2: Factors involved in the cause of online disinhibition, based on Suler (2004) and adapted from Bryant (2014).

Factor	**Explanation (exp.)**	**Everyday exp.**	**Notes**
Dissociative anonymity	The anonymity provided by online environments lends itself to feelings of 'dissociation' from usual self.	'You don't know me'	Suler considers this to be a 'primary' factor.
Invisibility	Users are not physically visible to others (unless they choose so).	'You can't see me'	Invisibility 'overlaps' with anonymity.
Asynchronicity	Delayed reaction; interaction with others does not have to happen in 'real time'.	'See you later'	Sometimes manifests itself as an emotional 'hit and run'.
Solipsistic introjection	The impression that an individual's mind has 'merged' with that of an online correspondent: a form of talking to oneself.	'It's all in my head'	Particularly the case with text communication e.g. through email.
Dissociative imagination	The online world is seen as part of a 'game' with different rules to everyday life.	'It's just a game'	A combination of solipsistic introjection with the 'escapability' of the online world.
Minimizing status and authority	The absence of traditional cues to status and authority.	'We're equals'	Status in cyberspace is determined by communication skills etc.
Solitude	People tend to access online environments whilst alone.	'I'm alone'	The absence of surveillance offline.

to be an attack as one left by registered users, less than half of all attacks were left by anonymous users. They also found that users with both high and low levels of activity contributed significant amounts of attacks.

Mondal et al. (2017) investigated the relationship between anonymity and hate speech online. They looked at the percentage of tweets that were posted anonymously. To do this they looked at the names on the Twitter accounts of their hate speech data and compared these against a lexicon of real names harvested from Facebook, that contained 4.3 million unique first names and 5.3 million unique last names. If the name of the Twitter account was not found then the post was deemed to be anonymous. They compared the percentage of tweets posted anonymously for random tweets, and hate speech tweets where the target was: race, sexual orientation, physical, behaviour and other. They found that in each case of the hate speech categories the percentage of anonymous posters was greater (between 46% and 55%) than for the random tweets (40%).

4.4.3 Motivation

Of course attributing motivation to tweeters is problematic. It is often the most ‘obvious’ explanation that people choose to label others. This may be successful, but it may not be, due to human interpretive biases. Judgements are based on intuitive interpretation that is informed by personal experience, and these interpretations can become reinforced, leading to confidence in possibly incorrect judgements. Psychological constructs such as beliefs, preferences and dispositions are especially difficult to interpret due to subjective worldviews, even more so since, what is revealed in public of a person’s persona, is often very different to their private one. Research also suggests the prevalence of hard-to-identify subconscious motivations, that can mean that people act on habit, without much thought as to why they behave in particular ways. This unavoidable reflexive focus on the culture, beliefs and knowledge of the researcher, can mean the complex attributes of the person under study are ignored. Humans also often suffer from *confirmation bias*, that is observations and evidence that support conclusions are sought out and focused upon (Hoffman, 2015). Misinterpreting why people behave as they do can also occur since there is a many-to-many relationship between motives and behaviours. An observed behaviour can be the result of different motives in different people, and people can behave in a number of ways even though they start with the same motivation. As well as motives other stimuli and influencing factors can also cause people to behave differently. Motives may be misinterpreted as a result of conflating them with personality. It is problematic to label people with particular motivation types, and assume that since they are a particular type their motivations conform to that type. Such assumptions may be true in some instances, but situational factors can override typical behaviour (ibid.).

Analysis of motivation is also complicated by the interaction of emotions with motivation and behaviour. Emotional strain can alter behavioural patterns, making analysis of motivation far more complex (ibid.).

Despite this it is informative to consider whether particular *types* of people may be more predisposed to racist tweeting, or other behaviours that may be indicative of a likelihood to send racist tweets. Psychopathic or Machiavellian traits are strongly predictive of violent and criminal acts and, although the research is sparse, they have also been linked with violence online (Peterson and Densley, 2017). Studies of online trolls have shown correlations between trolling behaviour and psychopathic characteristics including sadism, Machiavellian-ism and lack of empathy (ibid.).

Crandall and Eshleman (2003) examined the psychological processes that lead to expressions of prejudice, and proposed a justification-suppression model (JSM) to explain these processes. They defined prejudice as: 'a negative evaluation of a social group or a negative evaluation of an individual that is significantly based on the individuals group membership' (ibid., p.414). They argued that previous definitions of prejudice struggle with the notion of rationality and whether the prejudice is rational or not. They argued that for a definition of prejudice, rationality should be avoided because it is impossible to determine whether an actor is acting rationally or not, and that the psychological processes involved in prejudice, are the same whether the prejudice is performed rationally or irrationally. They argued that much theoretical work on prejudice can be summarised as people at an early age gaining bias and prejudice towards other racial groups and as they grow older, they become more influenced by societal norms, which are negative towards prejudice, and so they suppress their innate prejudices. They summarise this with the equation: $prejudice + suppression = expression$ (Crandall and Eshleman, 2003, p.416). They argued that older models of prejudice that see beliefs, values and ideology as underlying causal factors of prejudice, are misguided, and instead, suggest that these are not causes, but are instead justifiers of an underlying prejudice within them. The JSM contends that people have a 'genuine prejudice' that is a powerful motivational force. This prejudice is shaped by social, cultural, cognitive and developmental factors, and is acted on by other forces, both societal and personal, which act to suppress expressions of prejudice. At the same time there are justification forces also acting on an individual, that pull in the opposite direction to the suppression forces, and can facilitate the expression of prejudice (ibid.). The JSM in many ways is a good fit with RAT. For RAT a crime is likely to occur when there is a motivated offender, there is a suitable target and there is a lack of a capable guardian or manager, with the assumption of an underlying criminality of offenders. For JSM prejudice may occur when justification forces overcome suppression forces. Like RAT there is an assumption of an underlying propensity to 'do bad'. For RAT offenders *will* commit crime given a suitable scenario, and for JSM people *will* express prejudice. For both theories there is a balance between factors suppressing an event and factors encouraging an event. In terms of suppression, JSM contends that social and personal pressures stop people from expressing prejudice, and for RAT a capable guardian or lack of suitable target suppress the likelihood of crime. In terms of encouragement, JSM states that other forces can lead to the expression of prejudice, and for RAT a suitable target and the absence guardian are criminogenic factors.

4.5 Chapter Summary

In order to conceptualise racist tweets as *cybercrimes* this chapter presents a criminological theoretical framework. Initially what is meant by cybercrime is discussed. Then a brief summary of criminology is given followed by a discussion of the differences between situational and dispositional criminology. Routine Activity Theory (RAT) is discussed in detail along with the similar Lifestyle Exposure Theory (LET). Then RAT's applicability to cybercrime is discussed. Psychological perspectives are also discussed. The chapter concludes with a review of other work in the field of the automated detection of racist language.

It is noted that the terminology relating to cybercrime is complex and not clear-cut. Various terms, such as *digital crime*, have been used in the literature to refer to crimes somehow related to computing, although such terms are often criticised for being catchall terms. Use of the more narrow term *cybercrime* is also problematic since there is no consensus definition. However this is the term that will be used in this research. Whether racist tweeting is a crime, deviant behaviour, or neither depends on many factors, but from a UK legal perspective it can be considered (potentially at least) a criminal act. Hence criminological perspectives on crime and deviance are discussed.

A brief summary of criminology theory was given for the following theoretical groupings: Classicism and Positivism, Biological Positivism, Psychological Positivism, Durkheim, Anomie and Strain, The Chicago School, Subcultures and Cultural Criminology, Interactionism and Labelling Theory, Control Theories, Other Criminology including: radical and critical criminology, realist criminology, feminist criminology, and late modernity, governmentality and risk. The difference between dispositional and situational theories was discussed. Dispositional theories, try to answer the question why do people commit or desist from crime, from the point of view of a causal reason for lawbreaking. whereas those that take a situational stance, see criminal disposition as a ‘fact of life’; they are not interested in why a person commits a crime, other than how the situation the person finds them-self in is criminogenic. Control theory is an example of a dispositional theory, which argues that low self-control is the primary factor for criminality. Studies show some, albeit limited, support for the idea that low self-control is correlated with cybercrime. However, control theory does not consider the causal conditions under which such an event occurs. The pragmatic nature of the research, means a theoretical framework that can be used to analyse the event itself, as opposed to just the offender’s motives is needed. The dominant theory that is used to analyse the components that make up a criminal event, is that of RAT. Prior to RAT *Lifestyle Exposure Theory* (LET) argued that *lifestyle*, that is everyday routine activities related to work, school, and home, is the most important factor in the likelihood of victimisation for certain crimes.. There is considerable overlap between RAT and LET. They both look at how the spatial and temporal situations and actors find themselves in, affect the criminal event. Their main difference is that they were developed to explain different things: RAT theorists focused on crime rates whereas LET was developed to explain victimisation rates. Most of the literature ignores LET in favour of RAT but Choi's (2008) formulation of Cyber Routine Activity Theory (CRAT) is a notable exception.

RAT contends that crime is not a random event, instead it occurs in a disproportionate amount in areas that are criminogenic, because they either have a high number of motivated offenders, have a high number of suitable targets or lack capable guardians, or some combination of these. RAT evolved from this 'crime triangle' with the addition of three types of controllers: handlers that interact with the offender, place controllers who manage the location, and super controllers that are networks that the other controllers are part of, which provide incentives for controllers to reduce crime. RAT does not explain, for example, the disproportionate victimisation of females, nor does it consider the motivations of offenders. Its model of human behaviour as entirely rational is also debatable and its main theoretical constructs are very difficult if not impossible to measure. Despite these criticisms, RAT has been successfully applied to a number of real-world settings usually in the form of SCP which focuses on 'suitable targets' and their attractiveness to offenders. Crime prevention strategies that have evolved from SCP theory, focus on 'target hardening', that is making targets less attractive to offenders based on the model of target attractiveness corresponding to the acronym VIVA (value, inertia, visibility, and accessibility), which are rationally assessed by offenders as to whether a target is suitable for them. Value is a measure of how much value an offender gives the target or goal, inertia signifies the difficulty of removing an item, visibility relates to how visible an item is to an offender, and accessibility refers to the geography of a location, and layouts of buildings and rooms, where items are placed and any other factors that interact with the ability of offenders to access items.

Geometry of crime examines how the structure of cities influences the distribution of crime. All human behaviour is related to a backcloth: activities are constrained and determined within a number of influences. Within this backcloth activities are performed in an activity space, that individuals are familiar with, and the area that is visible from the activity space is the awareness space. Criminals are like anybody else, in that they perform most of their activities within their activity and awareness spaces. Brantingham and Brantingham (2015) argue that the various concepts behind RAT, such as routine activities, activity space and so on still can be applied when considering cyberspace instead of the physical world. There are of course changes, networks of activity are far more connected in hyperspace than physical space, and there is more overlap between activity spaces such as the overlap between Twitter, Facebook and other social media sites. A review of the literature on the application of RAT to cybercrime shows that VIVA applies well to the cybersphere. RAT's capable guardianship also applies well, in both the off-line and online world most human guardianship is of the informal nature and technological guardianship can be seen in both 'worlds'. Choi (2008) argues that RAT expands upon LET, in that the lifestyle variables of LET are the target suitability aspect of RAT. He notes that routine activities in cyberspace, related to both work and leisure, can increase the chances of an individual's victimisation. He argues that differences in online lifestyles and digital guardianship are correlated with online victimisation. Cyber-Routine Activities Theory (Choi and Lee, 2017) integrates the concepts of LET and RAT to computer crime victimisation, arguing that differences in online lifestyles and digital guardianship are correlated with online victimisation.

As well as the sociological discussion already given. psychological perspectives can give insight on online offending. It has been theorised that the internet has a disinhibiting effect on users, although the evidence for this is mixed. The disinhibition can be both benign and toxic. One of the reasons for this effect is theorised to be anonymity, although again research is mixed on the correlation. between anonymity and disinhibition. Psychopathic or Machiavellian traits are strongly predictive of violent and criminal acts and, although the research is sparse, they have also been linked with violence online. Crandall and Eshleman (2003) examined the psychological processes that lead to expressions of prejudice, and proposed a justification-suppression model (JSM) to explain these processes. The JSM contends that people have a 'genuine prejudice' that is a powerful motivational force. This prejudice is shaped by social, cultural, cognitive and developmental factors, and is acted on by other forces, both societal and personal, which act to suppress expressions of prejudice. At the same time there are justification forces also acting on an individual, that pull in the opposite direction to the suppression forces, and can facilitate the expression of prejudice. Like RAT there is an assumption of an underlying propensity to 'do bad'. For both theories there is a balance between factors suppressing an event and factors encouraging an event. In terms of suppression, JSM contends that social and personal pressures stop people from expressing prejudice, and for RAT a capable guardian or lack of suitable target suppress the likelihood of crime. In terms of encouragement, JSM states that other forces can lead to the expression of prejudice, and for RAT a suitable target and the absence guardian are criminogenic factors.

The following chapter discusses the methodological issues related to the investigation of racism on Twitter.

Chapter 5

Methodology

This chapter explores the use of mixed methods, that is applying a combination of quantitative (ML) and qualitative (grounded theory) methods to answer the research questions posed. First a theoretical discussion of mixed methods is given. Then the use of machine learning is explored, including an explanation of the difference between supervised and unsupervised learning. Then the application of machine learning to textual data is discussed, including a section on how the data is preprocessed prior to input into an ML algorithm. Finally the qualitative analysis of the tweets is discussed, including various techniques, the main one being that of grounded theory.

5.1 Mixed Methods

There is, at first glance at least, a clear dichotomy in the nature of the research questions listed in Section 1.2: some are clearly quantitative, for example, ‘is it possible to have a reliable, efficient and accurate automated racist tweet identifier?’, whereas others are qualitative, for example, ‘can grounded theory and criminology be used to further the understanding of racist tweets?’ In the past there has been something of a tension between qualitative and quantitative methods, and it is necessary to discuss the theoretical background behind the methods used, and how they can be integrated within a particular *epistemological*[1], *ontological*[2] and *axiological*[3] stance (Tuli, 2011).

Traditionally in social research there was (and still is) a conflict between two different paradigms, on the one side those of *positivism* (and the later, *postpositivism*) and on the other side the paradigms of *constructivism/interpretivism* (Feilzer, 2010). Positivism is a philosophical stance closely tied to the hypothetico-deductive method (Ponterotto, 2005). It is a *realist* philosophy, that is, from an ontological view, it assumes an objective or 'true' reality that exists beyond subjective interpretations and constructions, and this truth can be measured and discovered (Bhaskar, 2009). There are two forms of realism: *naive realism* which is in contrast to *critical realism*, both contend that there is an objective truth, but critical realism stresses that measuring and uncovering this truth can only be done imperfectly (Ponterotto, 2005). Epistemologically positivists see the relationship between researcher and the subject of the research as an objective relationship. Under this paradigm's axiological stance, research aims to eradicate or, at least, reduce the effect of the researcher on the subject, the idea being that the result is an observation of 'truth', which is not the case if researchers biases are allowed to affect the result (ibid.).

In contrast to positivism, constructivists deny the existence of a single observable truth, instead believing there are multiple constructed realities, which are always subjective. This is a *relativist* ontological position; there are no objective truths, only those constructed by the interaction of researchers with subjects under particular situations. For constructivists, epistemologically and axiologically, not only does there not have to be a disjuncture between researcher and researched, their interaction is often regarded as a necessary requirement to gain a rich understanding of a social situation (ibid.).

This research uses both quantitative and qualitative methods, the qualitative methods chiefly being grounded theory, which is very much situated within the constructivist paradigm, and the quantitative method of machine learning which is within the positivist paradigm (ibid.). The solution to the use of methods with potentially conflicting theoretical bases, is to take a *mixed methods* approach, with such an approach following the

[1] Epistemology is the study of the nature of knowledge.

[2] Ontology is the study of the nature of reality.

[3] Axiology is the study of the nature of value. For research it includes debates about what role the researcher should play in the research process (Ponterotto, 2005).

paradigm of *pragmatism* (Feilzer, 2010). For Feilzer (ibid., p.8), pragmatism,

> when regarded as an alternative paradigm, sidesteps the contentious issues of truth and reality, accepts, philosophically, that there are singular and multiple realities that are open to empirical enquiry and orients itself towards solving practical problems in the 'real world.'

In other words its emphasis is on 'getting things done' using research, it allows that both sides of the 'paradigms wars' are equally valid. With respect to this research, the initial driver was the pragmatic aim to develop computer software aimed at automatically identifying racist tweets and tweeters. To do this machine learning routines were written with the viewpoint that it is possible, albeit imperfectly, to identify what 'racist tweets' and who 'racist tweeters' are. This postpositivist work was combined in a mixed methods approach, with an analysis, using grounded theory, of the identified tweets, and the concepts which arise from this analysis, were used to identify racist tweeters. The grounded theory portion of the work, is very much situated in constructivism, yet it also fits within the pragmatic paradigm.

The idea driving the research was to create a computer system that would perform a task, and the aim was that creating this system, would be done as objectively as possible. Of course it is naive to assume that such a process is, in reality, completely objective, since many decisions and choices were made by the researcher meaning the creation of the system would always be somewhat subjective. However this is always a criticism that can be directed towards any positivist or postpositivist work, and while it might be valid, the pragmatic paradigms within which the work was performed justifies this approach. Even if it was assumed that the computer system creation task was wholly objective, this research could potentially still be criticised because of the postpositivist stance of computer system research being in conflict with the constructivist stance of the grounded theory analysis.

Nearly all research is a combination of both inductive and deductive processes. While aspects of induction and deduction can be seen within most research methods, grounded theory tends to be much more inductive and atheoretical. It makes no prior theoretical assumptions, and so provides a good balance to the more theory driven deductive computing side of this research. It, in some ways, makes the research more 'well-balanced' since it provides qualitative human analysis of the data. It is well-suited to the analysis of the tweet data, since it provides theoretical insight and avoids preconceptions about the data. It is also well-suited to the short format of the tweets data, whereas other qualitative methods, such as narrative analysis, are unlikely to work well with small pieces of textual data.

The grounded theory analysis also arose from a pragmatic desire, since grounded theory is a very popular qualitative method, that has had much success in research aimed at understanding the meaning behind texts. As can be seen in the discussion of the difficulties in interpreting tweets in Section 3.4.5, it is very difficult to determine meaning and intention behind tweets. Ideally any such understandings gleaned from the data would be objective and free from researcher bias, however this is impossible since any researcher has to draw on their life experiences, cultural awareness and other factors that make up their worldview. The example of the word 'nigga' having an array of meanings and interpretations that differ internationally, and between different cultures and also within cultures, shows the difficulty of interpretation and its interaction with the interpreter. Despite this the resulting concepts were used as input to the machine learning algorithms which were aiming to identify racist tweeters.

So the racist tweets created by the postpositivist computer system research were analysed using constructivist grounded theory, and the results of this were input into more postpositivist computer programs. While it would be very difficult to site both forms of research in either positivism or constructivism, they both fit well in pragmatism where the end justifies the means.

A mixed methods approach like this is also justified since it has a proven track record in social sciences, in particular policing (see, for example: Williams and Stahl (2008), Brown et al. (2010), Elliott et al. (2011), Beletsky et al. (2016), Kuehl et al. (2016), Dwyer et al. (2017), and Kiedrowski et al. (2017). One of the main drivers behind the choice of method in this research is the type of data being analysed. Hence the data used in this research will be now discussed: its nature, collection, storage and annotation. This research stores and performs intensive processing on big data scale data. The Twitter data collected in this research is an example of big data, and the methods used to analyse this data are techniques associated with the handling of such data. The definition of big data used for this research is:

> big data is data that requires parallel processing either because of its size or analysis requirements.

This definiotn and big data in general are discussed further in Chapter 6. How computing resources can be utilised to handle big data is discussed in the next section.

5.2 Scalability

Big data systems need to be highly scalable so that further increases in volume, variety or velocity of data can be handled efficiently. A computer system needs to be 'scaled' when the input it is working on starts becoming too big for its hardware and/or software to handle. A system that is scalable is one that allows for the addition of increasing processing power relatively easily, with very little disruption to the underlying software and hardware structure. Scaling can be thought of in terms of both software and hardware. Software is needed that can efficiently allocate data and processing tasks, and hardware is needed to run the tasks on. Big data systems normally allow for scaling by being designed to allow distributed parallelisation of their tasks. In such distributed parallel systems the architecture is normally a cluster of many commodity hardware machines that have a shared file system. The machines work in parallel, splitting a large task into many independent smaller sub tasks that can be worked on at the same time (Taylor, 2010). If scaling is required one or more commodity machines can be added to a cluster. Such systems require software that both allows new computing resources to be added to a cluster seamlessly, and that can organise and delegate tasks in parallel to the machines that comprise the cluster.

The most common paradigms for distributed software parallelization are the implementations of Message Passing Interface (MPI) and the Hadoop framework (Taylor, 2010). Hadoop has two advantages over MPI: it is regarded as easier to use and it aims to process data locally. In Hadoop the machine with the data usually processes the data, whereas MPI moves data across the network reducing its efficiency (rjurney, 2012). Because of these advantages, this research followed the lead of Murthy and Bowman (2014) who implemented a Hadoop infrastructure, in what they called a ‘small-scale’ big data system. Further details of Hadoop are given in the next section.

5.3 Hadoop

Hadoop is the most prevalent distributed parallel computing paradigm. It originally was an open source variation of Google’s Google File System (GFS, Hadoop’s equivalent is known as Hadoop Distributed File System (HDFS)) and MapReduce architecture (Ghemawat et al., 2003; Dean and Ghemawat, 2008). In 2006 Doug Cutting, an employee of Yahoo, developed Hadoop based on GFS/MapReduce (Harris, 2013). Hadoop has a number of advantages for this research: it is free, it works best on big and static data, it is schema-less i.e can handle different data sources, such as text files, html files, NOSQL databases etc. and it efficiently distributes processing to nodes[4] (White, 2012; Lam, 2010). Hadoop is designed to meet the requirement that nodes can easily be added to scale up the system. HDFS is a component of Hadoop that allows files to be stored in distributed storage and MapReduce is the original computational framework in Hadoop. With Hadoop version 2.0 a new computational framework, YARN (Yet Another Resource Negotiator), was introduced to replace MapReduce. YARN allows users of Hadoop to utilise different processing frameworks against their HDFS data, rather than just MapReduce, although MapReduce can still be used as one of the frameworks that runs on YARN (Vavilapalli et al., 2013). Hadoop is a distributed master-slave architecture which is discussed in the following sections for HDFS, MapReduce and YARN.

[4] Nodes in Hadoop are logical constructs that perform specific tasks. They can usually be thought of as individual machines, although it is possible to have, for example, a machine that is both a NameNode and a DataNode.

5.3.1 HDFS

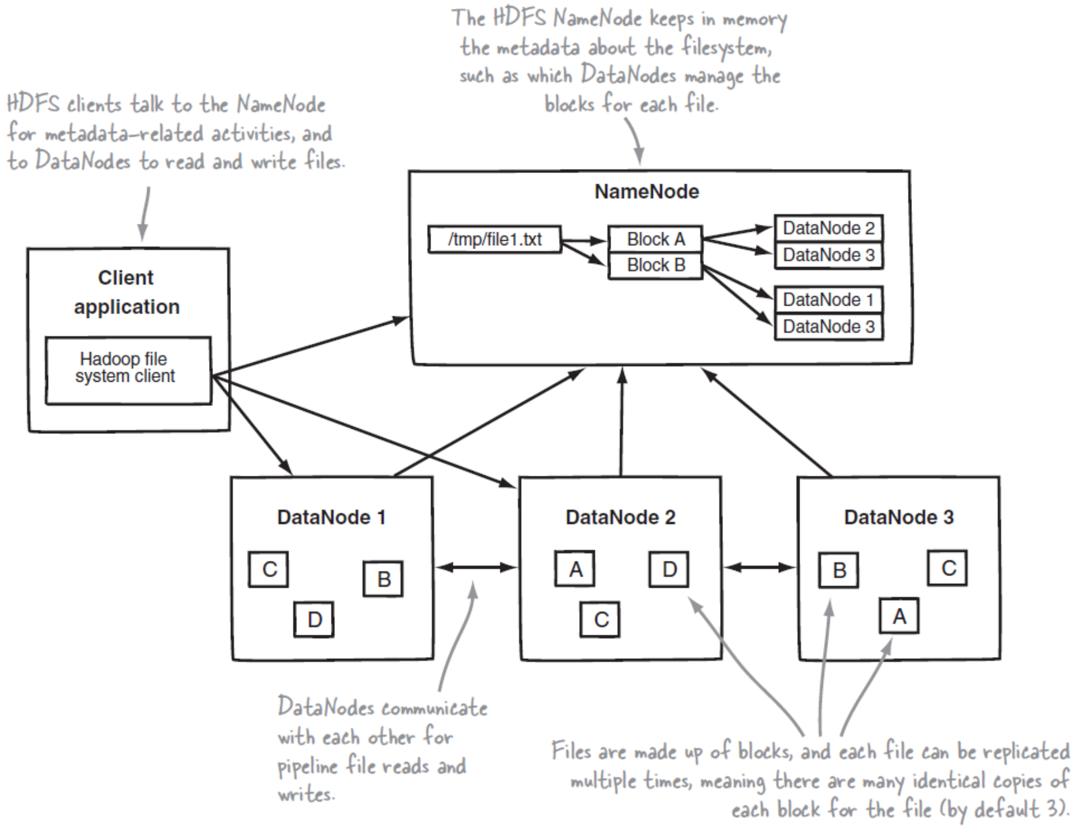

The HDFS NameNode keeps in memory the metadata about the filesystem, such as which DataNodes manage the blocks for each file.
HDFS clients talk to the NameNode for metadata-related activities, and to DataNodes to read and write files.
NameNode
/tmp/file1.txt
Block A
Block B
DataNode 2
DataNode 3
DataNode 1
DataNode 3
Client application
Hadoop file system client
DataNode 1
C
B
D
DataNode 2
A
D
C
DataNode 3
B
C
A
DataNodes communicate with each other for pipeline file reads and writes.
Files are made up of blocks, and each file can be replicated multiple times, meaning there are many identical copies of each block for the file (by default 3).

Figure 5.1: HDFS architecture shows an HDFS client communicating with the master name node and slave data nodes, from Holmes (2012).

HDFS's architecture is outlined in Figure 5.1. Hadoop creates a process known as *NameNode* which has overall control over the HDFS system. The NameNode liaises with a number of *DataNodes* on which the actual data resides. The NameNode is responsible for the following tasks:

- Handling the file systems metadata such as where the files are stored and the folder structure.
- Managing permissions and access to the data.
- Mapping of the data to the data nodes.
- Opening and closing files and other file system operations.
- Other tasks such as registration of DataNodes.

DataNodes perform the actual storage of the data, and any replications required. The data never travels to the NameNode since there is only one such node (although it is possible to include a standby node in case the NameNode goes down). Instead the data always reside on the potentially very many DataNodes, and as the requirements for data processing grow, more DataNodes can be added to scale the system (Alapati, 2016, p.39-40).

By default each HDFS block of data (of size either 64MB or 128MB) is replicated twice (in total there will be 3 copies of it in HDFS), to allow for machine failure and efficient allocation of tasks. Hadoop also tries to process data blocks on a machine on which they reside, in order to reduce network traffic, since minimising network traffic increases efficiency and processing speed (Shvachko et al., 2010).

If, say, a file is uploaded to Hadoop for processing, HDFS will store that file across the network, on one machine if possible. A MapReduce task can be run against this data. Hadoop will run multiple copies of this task, perhaps one on each node, and each of these will handle a different part of the file, most likely the block(s) of data stored by HDFS on the same node. Hadoop automatically aggregates the results from all the tasks running on the various nodes. More details of this and a conceptual view of MapReduce are given in the following section.

5.3.2 MapReduce

MapReduce is a computing framework that runs in batch mode, that is MapReduce jobs are run with input data, they perform the processing with no additional human input, and produce results, usually with some delay. This means that MapReduce's speed of processing is not suitable for computational tasks that require fast response times, such as online user interaction. MapReduce makes distributed parallel computing much easier since it handles the parallelisation of any processing necessary and it also handles the management of any resources required in the processing. It also reduces the problems created by unreliable hardware and software, since Hadoop makes it very easy to add and remove resources for MapReduce jobs (Holmes, 2012).

MapReduce's architecture can be seen in Figure 5.2. MapReduce splits jobs up into a number of parallel map and reduce tasks. MapReduce is a *shared-nothing* system[5], that is each map and reduce task is independent, there is nothing shared between them. Hadoop takes care of the splitting and combining of the data between the MapReduce tasks. Maps take as input key value pairs, and output lists of key value pairs. The details of the map are up to the programmer. As can be seen in Figure 5.3 key value pairs that are output from the mappers, are shuffled and sorted by the MapReduce framework, which involves them being *partitioned*, that is divided up between the different reducers, and also the key value pairs list input to each reducer is sorted.

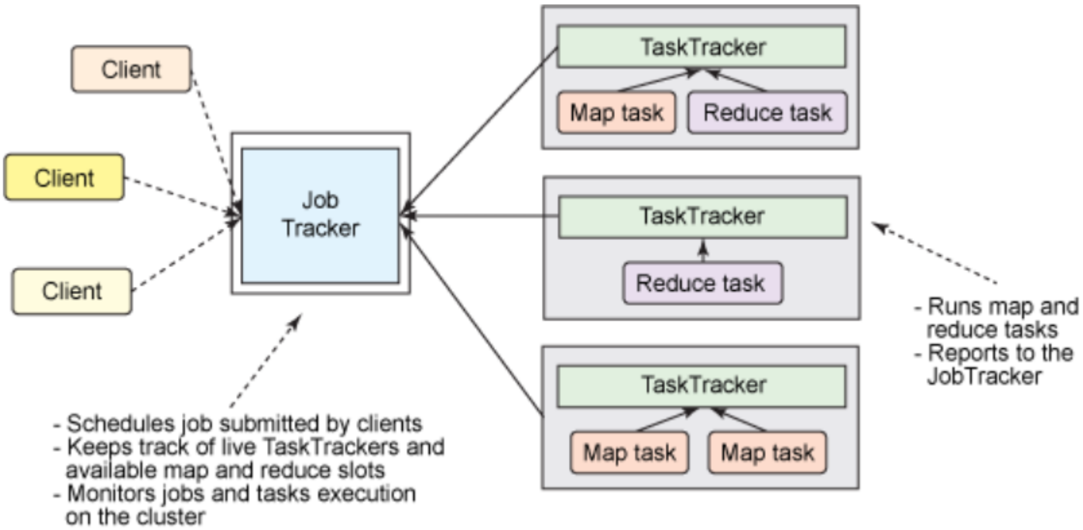
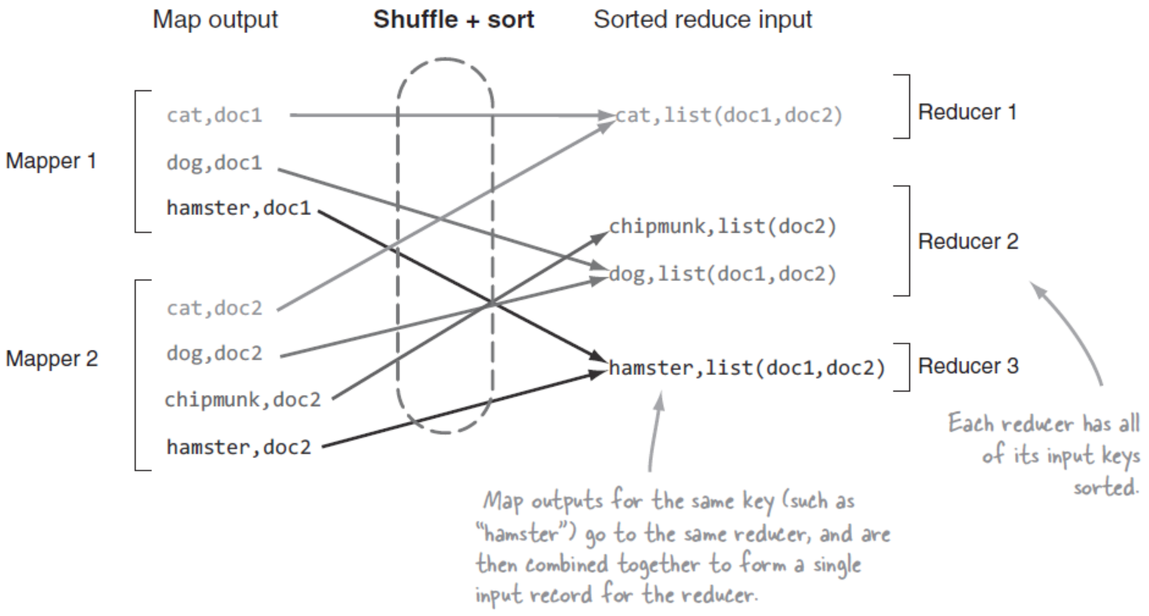

Client
Client
Client
Job Tracker
TaskTracker
Map task
Reduce task
TaskTracker
Reduce task
TaskTracker
Map task
Map task
- Runs map and reduce tasks
- Reports to the JobTracker
- Schedules job submitted by clients
- Keeps track of live TaskTrackers and available map and reduce slots
- Monitors jobs and tasks execution on the cluster

Figure 5.2: MapReduce architecture from Kawa (2014).

MapReduce jobs are executed using a structure similar to the HDFS NameNode/-DataNodes structure. A single *JobTracker* coordinates the process and handles the coordination of map and reduce tasks, assigning each to a particular *TaskTracker*. The problems with this architecture are that using a single JobTracker can lead to bottlenecks, and that it only allows for one kind of computation. These drawbacks led to the development of YARN (Kawa, 2014), which addresses both issues. A discussion of YARN is given in the following section.

[5] Both HDFS and YARN are also shared-nothing systems.



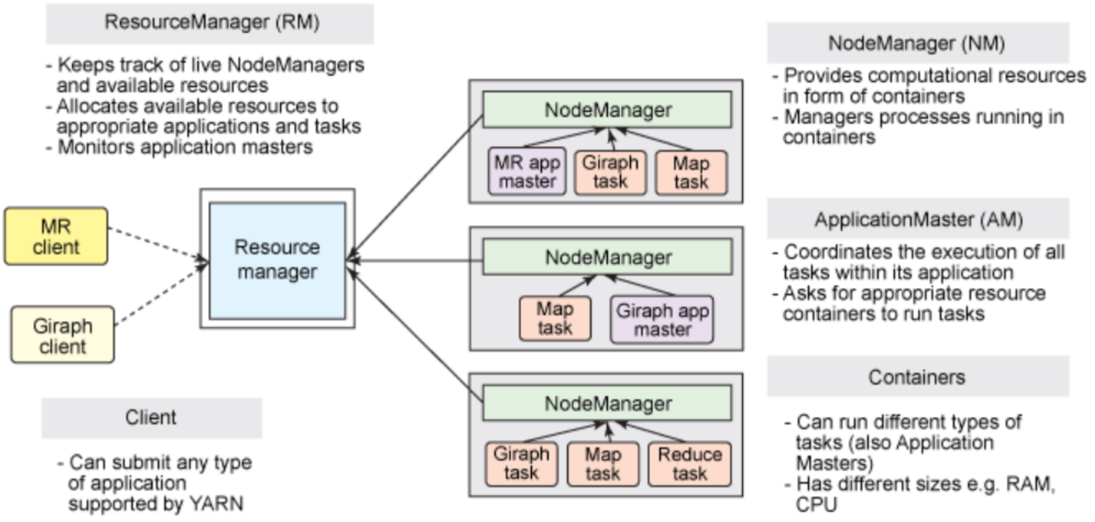
Each reducer has all of its input keys sorted.

Map outputs for the same key (such as "hamster") go to the same reducer, and are then combined together to form a single input record for the reducer.

Figure 5.3: MapReduce processing from Holmes (2012, p.8).

5.3.3 YARN

The architecture of YARN can be seen in Figure 5.4. YARN runs a single *ResourceManager*, which handles all the resources within a Hadoop cluster. The ResourceManager communicates with a number of *NodeManagers*, one on every DataNode. YARN refers to jobs as *applications*, and each application has an *ApplicationMaster* which communicates with both the ResourceManager and NodeManagers to negotiate resources for the tasks that make up its application. It is the addition of these ApplicationMasters in YARN



- Keeps track of live NodeManagers and available resources
- Allocates available resources to appropriate applications and tasks
- Monitors application masters

- Can submit any type of application supported by YARN

- Provides computational resources in form of containers
- Managers processes running in containers

- Coordinates the execution of all tasks within its application
- Asks for appropriate resource containers to run tasks

- Can run different types of tasks (also Application Masters)
- Has different sizes e.g. RAM, CPU

Figure 5.4: YARN architecture from Cloudera (2014).

that addresses the bottleneck issue in MapReduce. In MapReduce a single JobTracker was used to allocate and manage all the resources necessary for all jobs. In YARN however, each job - or ‘application’ in YARN terminology - has its own ApplicationMaster which does a similar job to that of MapReduce’s JobTracker.

When an application runs its ApplicationMaster makes a *ResourceRequest* to the ResourceManager’s *Scheduler* and the Scheduler then allocates a number of *containers* to the application, each container having a certain amount of resources such as CPU cores and memory. For example the Listing 5.1 shows the command needed to run an instance of *Spark-shell*, which is a command line interface to Spark. The command shown has a number of parameters, of these the parameters `--driver-memory 2g  --executor-memory 2g   --executor-cores 5 --num-executors 11` refer to YARN container parameters. Each container has a *driver* which has overall command of the container, and a number of *executors* which do the processing work. This example shows the driver being allocated 2 GB of memory (`--driver-memory 2g`), and the same amount to each executor (`--executor-memory 2g`), there being 11 such executors (`--num-executors 11`) and with a maximum of five on each machine (`--executor-cores 5`) (Vavilapalli et al., 2013).

List of Code 5.1: Command used to start Spark-shell.

```
/usr/hdp/2.6.3.0-235/Spark2/bin/Spark-shell --jars /home/ed/.ivy2/jars/stanford
    -corenlp-3.6.0-models.jar,/home/ed/.ivy2/jars/jersey-bundle-1.19.1.jar,/
    home/ed/Downloads/serde/json-serde-1.3.7-jar-with-dependencies.jar  --
    packages databricks:Spark-corenlp:0.2.0-s_2.11,edu.stanford.nlp:stanford-
    corenlp:3.6.0 \
--driver-memory 2g  --executor-memory 2g   --executor-cores 5 --num-executors
    11 --master yarn  --deploy-mode client --conf Spark.driver.maxResultSize=2g
```

The advent of YARN meant that Hadoop processing no longer was exclusively tied to MapReduce. Other alternatives for distributed processing that access data residing in HDFS, are now possible, one of which is Apache’s Spark. The following section discusses

Spark, its machine learning capabilities, and how it interacts with YARN.

5.4 Spark

Apache Spark arose out the work of the University of California, Berkeley, who in 2009, started a project to create an engine that could handle distributed data processing. One of the aims was to create a single system that could handle the different parts of big data processing stack, the loading, querying and machine learning processes all performed by the same software. They created a system with a programming model similar to MapReduce, but with the addition of distributed data structures known as *Resilient Distributed Datasets (RDD).* RDDs allow Spark to handle different kinds of processing on the same dataset. Spark performs transformations on the data lazily, that is it only transforms the data when it is required to return data to the requester, and will stack a series of transformations together and perform them all at once. Spark data structures are, by default, ephemeral, that is they are recreated each time they are used, however it is possible to persist the data in memory and this in-memory data sharing means Spark's performance often outweighs that of MapReduce as much as 100 times. Spark is also more suited to machine learning processing than MapReduce, since machine learning usually requires an iterative algorithm requiring a number of steps. For MapReduce to handle such an algorithm it needs to write data to HDFS, and this writing to disk takes up much of the time of the processing for a job. Spark's in-memory data structures reduce the amount of disk access required, thus speeding up processing significantly (Zaharia et al., 2016). In March 2015 version 1.3.0 of Spark was released with a new data structure, the *dataframe*, which can be thought of conceptually as a table with rows and columns, similar to SQL tables, but with performance optimisations for Spark (Spark, 2015b). The dataframe API is easier to optimise than the RDD one but lacks some of the functionality of the RDD, so a further API, the *dataset* was introduced in Spark 1.6.0 (Spark, 2016b), which aimed to combine the advantages of both the RDD and dataframe (Spark, 2016a). The dataframe and dataset were combined in Spark 2.0 (Damji, 2016). In order to handle machine learning processing Spark contains two libraries: *`mllib`* and *`ML`*. The original Spark machine learning library, `mllib` uses the RDD and includes a number of machine learning algorithms, including the classification models: Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Decision Tree (DT), Gradient Boosted Tree (GBT) and Artificial Neural Network (ANN). *`ML`* was introduced in Spark 1.2 (Spark, 2015a). *`ML`* introduced the use of the dataframe API in ML, and pipelines. In Spark version 2.0 `mllib` entered maintenance mode[6] and the `ML` package became the primary machine learning API for Spark, although, rather confusingly, its guide is known as the 'MLLib' guide (Spark, 2017b).

Spark provides a number of APIs to access its functionality, and these APIs allow the use of four different programming languages: Scala, Python, Java and R. This research utilised the Scala version of the API, since that language is a language that Spark is written in, and its Scala API is the richest of the three, with new functionality appearing in the Scala API first. An example of Scala programs using this API are given in Listings A.1 and A.2. A.1 shows the use of HIVEQL to access tweets data stored in Hive, and then the preprocessing of the extracted tweet textual and metadata. A.2 shows the use of ML's implementation of SVM, *`LinearSVC`* and the use of pipelines in ML. This program performs a number of manipulations on the data in the pipeline, before using it in the *`LinearSVC`* algorithm to produce a model, which is then assessed with a number of metrics, and used to predict whether the D1 tweets were racist or not.

Spark's architecture can be seen in Figure 5.5. Spark has for each application a single *driver* program that is created when a SparkContext is created. The driver communicates with the cluster manager.[7] The cluster manager manages a number of workers, each of which runs one or more executors.

[6] According to Xiangrui (2016) the placing of `mllib` into maintenance mode means:

- We do not accept new features in the RDD-based Spark.mllib package, unless they block implementing new features in the DataFrame-based Spark.ml package.
- We still accept bug fixes in the RDD-based API.
- We will add more features to the DataFrame-based API in the 2.x series to reach feature parity with the RDD-based API.
- Once we reach feature parity (possibly in Spark 2.2), we will deprecate the RDD-based API.
- We will remove the RDD-based API from the main Spark repo in Spark 3.0.

[7] For this research the cluster manager used was YARN, but there are other possibilities: Mesos can be used, and Spark also has its own built-in cluster manager.

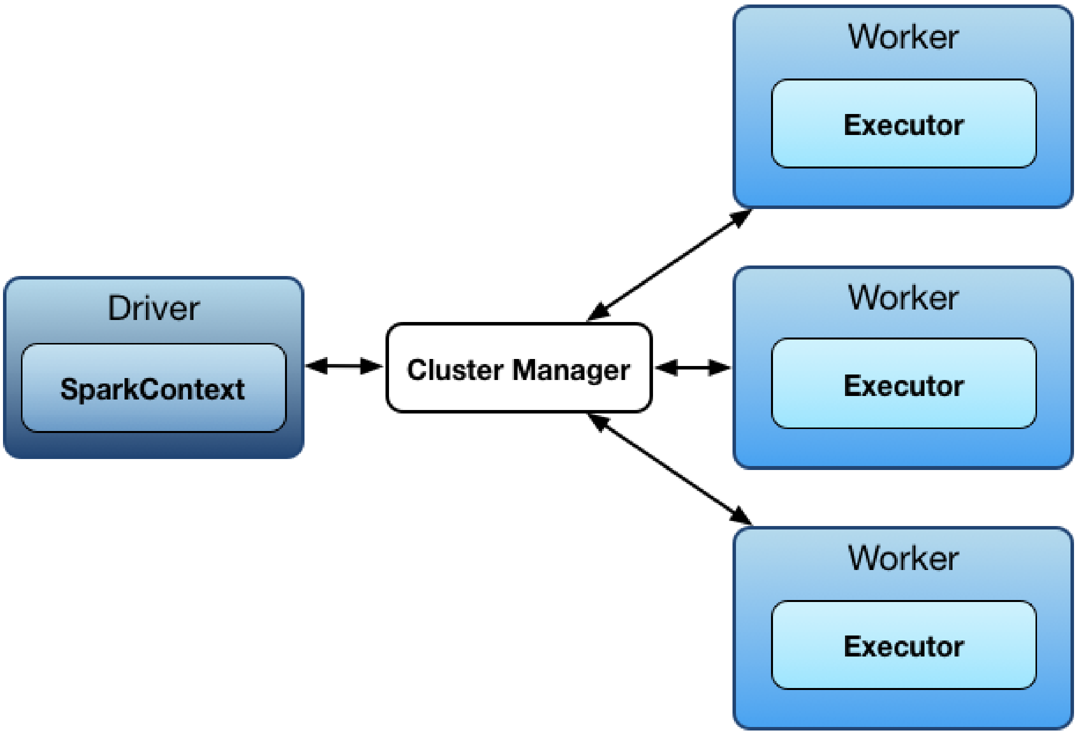


Figure 5.5: Spark architecture from Laskowski (2017, p.242).

A diagram representing the submission of a Spark job on YARN is shown in Figure 5.6.

Spark/YARN/Hadoop is the framework that was utilised for this research to store and manipulate large Twitter datasets. Now the nature of Twitter data is discussed.

5.5 Tweets

The 140 character messages that tweets contain are only a small part of their structure. Tweets are in fact complex JavaScript Object Notation (JSON)[8] structures (an example is given in Appendix A.3), which shows that tweets are far more complex and contain much more information than their 140 character text would suggest. Their structure is described in Twitter's Developer documents (Twitter, 2017). A tweet's structure contains

[8] For many explorative tasks JSON format is not ideal; The JSON format can be quite unwieldy, and it can be hard to gain insight from data, when viewed in this format. For example figure 5.7 is a screenshot of the file razakars1.txt opened in the integrated development environment Geany.

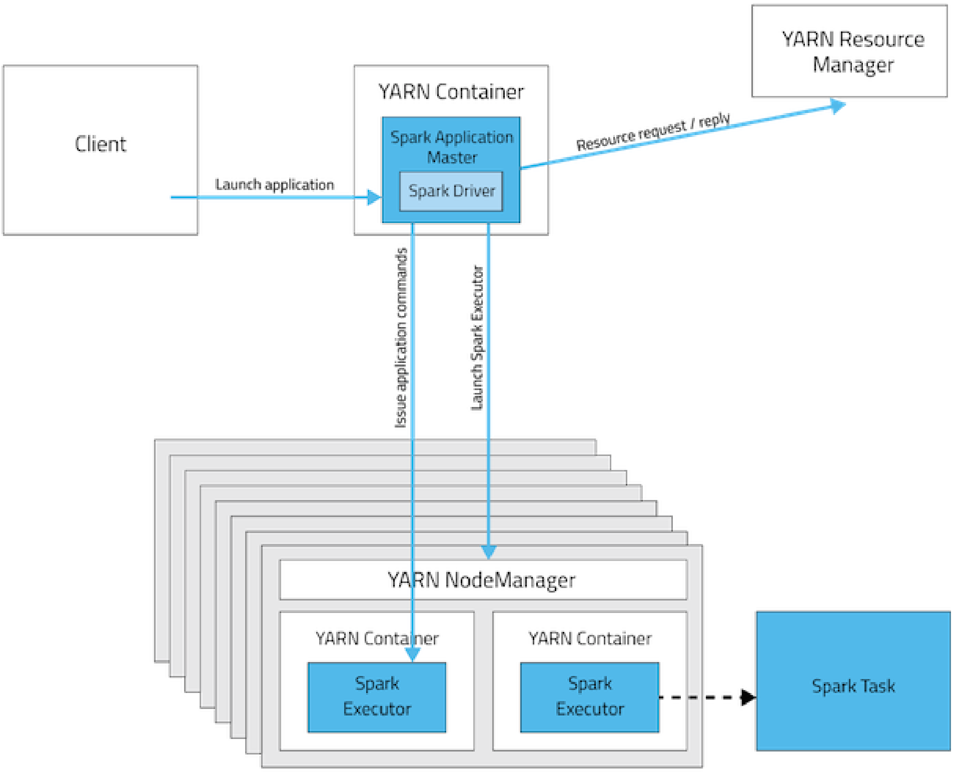
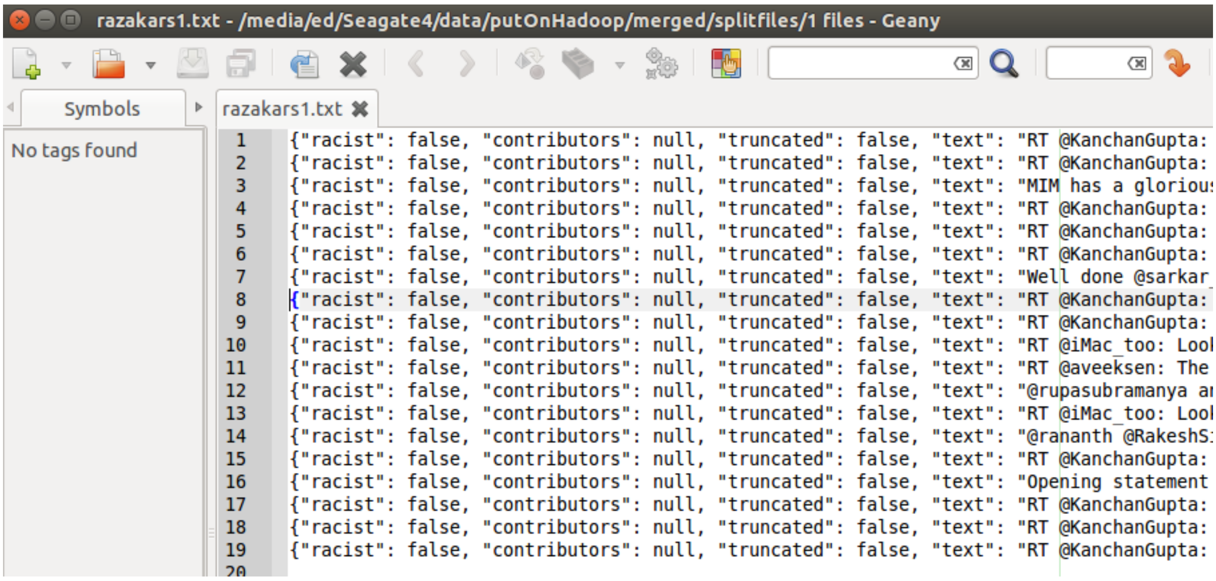


Figure 5.6: Submission of a Spark application to a YARN cluster, from Cloudera (2014).

a number of root level attributes such as *user*, some of which contain child objects such as *user.location*.


razakars1.txt - /media/ed/Seagate4/data/putOnHadoop/merged/splitfiles/1 files - Geany
Symbols
No tags found
razakars1.txt

```
{"racist": false, "contributors": null, "truncated": false, "text": "RT @KanchanGupta:
{"racist": false, "contributors": null, "truncated": false, "text": "RT @KanchanGupta:
{"racist": false, "contributors": null, "truncated": false, "text": "MIM has a gloriou
{"racist": false, "contributors": null, "truncated": false, "text": "RT @KanchanGupta:
{"racist": false, "contributors": null, "truncated": false, "text": "RT @KanchanGupta:
{"racist": false, "contributors": null, "truncated": false, "text": "RT @KanchanGupta:
{"racist": false, "contributors": null, "truncated": false, "text": "Well done @sarkar_
{"racist": false, "contributors": null, "truncated": false, "text": "RT @KanchanGupta:
{"racist": false, "contributors": null, "truncated": false, "text": "RT @KanchanGupta:
{"racist": false, "contributors": null, "truncated": false, "text": "RT @iMac_too: Loo
{"racist": false, "contributors": null, "truncated": false, "text": "RT @aveeksen: The
{"racist": false, "contributors": null, "truncated": false, "text": "@rupasubramanya a
{"racist": false, "contributors": null, "truncated": false, "text": "RT @iMac_too: Loo
{"racist": false, "contributors": null, "truncated": false, "text": "@rananth @RakeshS
{"racist": false, "contributors": null, "truncated": false, "text": "RT @KanchanGupta:
{"racist": false, "contributors": null, "truncated": false, "text": "Opening statement
{"racist": false, "contributors": null, "truncated": false, "text": "RT @KanchanGupta:
{"racist": false, "contributors": null, "truncated": false, "text": "RT @KanchanGupta:
{"racist": false, "contributors": null, "truncated": false, "text": "RT @KanchanGupta:
```

Figure 5.7: JSON file opened in Geany.

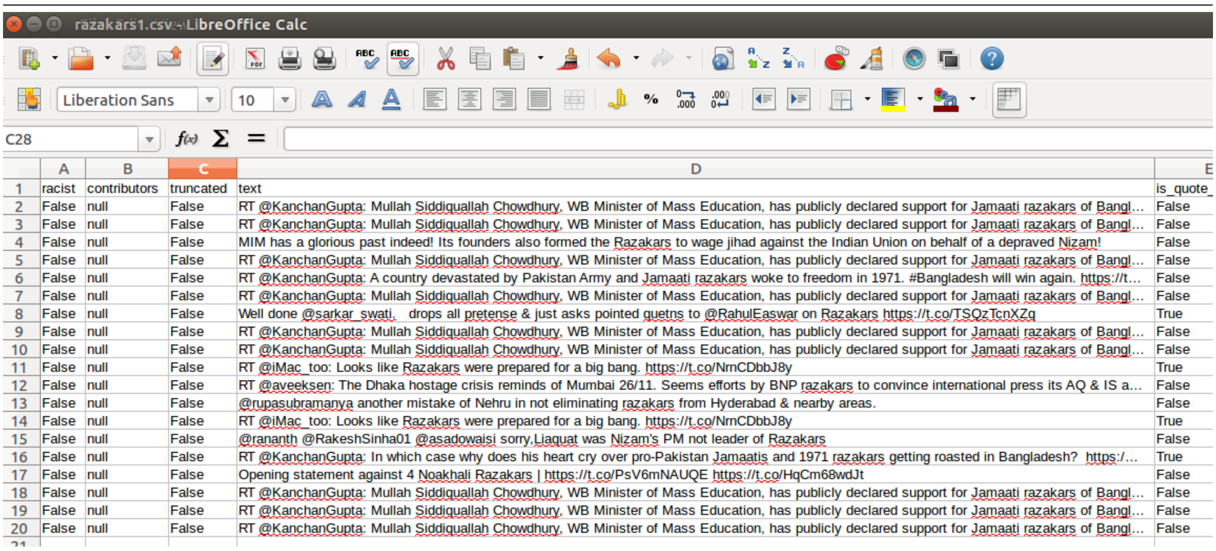
This view of the file, has field names and their contents side-by-side, separated by colons, which is an awkward way of presenting data.

For ease of viewing it can be useful to transform JSON to csv format using, for example, the website https://json-csv.com/. An example of a resultant csv file opened using LibreOffice Calc is shown in figure 5.8.

5.6 Sampling Tweets

As mentioned in Chapter 1 the use of Twitter data as providing an insight into people's opinions beliefs is seductive, since it is easy to think of it as representing a 'digital agora' and being representative of the 'public sphere'. It is also, of course, seductive from a methodological viewpoint, since Twitter data is public, and reasonably easy to access, either through Twitter's website, third party apps and websites such as TweetDeck, or with a little programming knowledge through its API. However, like all data sources, the use of Twitter is not without problems in relation to bias (Morstatter and Liu, 2017). There is an age bias in Twitter data, since the age profile of Twitter tends to skew towards younger age groups (ibid.). However that is not an issue for this research, since the aim of the analysis is knowledge about Twitter usage, and further generalisations are not relevant (although of course it is always of interest to be able to generalise findings).

Morstatter and Liu (ibid.) also noted that malicious users can introduce bias into Twitter data. There are two types of malicious users: bots, which are automated programs, that are used for spam, or to mass tweet, thus creating a topic to trend or altering the mass of tweets about a particular topic. The other kind of malicious users are humans, who may be performing similar actions to bots.



	A	B	C	D	E
1	racist	contributors	truncated	text	is_quote_
2	False	null	False	RT @KanchanGupta: Mullah Siddiquallah Chowdhury, WB Minister of Mass Education, has publicly declared support for Jamaati razakars of Bangl…	False
3	False	null	False	RT @KanchanGupta: Mullah Siddiquallah Chowdhury, WB Minister of Mass Education, has publicly declared support for Jamaati razakars of Bangl…	False
4	False	null	False	MIM has a glorious past indeed! Its founders also formed the Razakars to wage jihad against the Indian Union on behalf of a depraved Nizam!	False
5	False	null	False	RT @KanchanGupta: Mullah Siddiquallah Chowdhury, WB Minister of Mass Education, has publicly declared support for Jamaati razakars of Bangl…	False
6	False	null	False	RT @KanchanGupta: A country devastated by Pakistan Army and Jamaati razakars woke to freedom in 1971. #Bangladesh will win again. https://t…	False
7	False	null	False	RT @KanchanGupta: Mullah Siddiquallah Chowdhury, WB Minister of Mass Education, has publicly declared support for Jamaati razakars of Bangl…	False
8	False	null	False	Well done @sarkar_swati. drops all pretense & just asks pointed quetns to @RahulEaswar on Razakars https://t.co/TSQzTcnXZq	True
9	False	null	False	RT @KanchanGupta: Mullah Siddiquallah Chowdhury, WB Minister of Mass Education, has publicly declared support for Jamaati razakars of Bangl…	False
10	False	null	False	RT @KanchanGupta: Mullah Siddiquallah Chowdhury, WB Minister of Mass Education, has publicly declared support for Jamaati razakars of Bangl…	False
11	False	null	False	RT @iMac_too: Looks like Razakars were prepared for a big bang. https://t.co/NmCDbbJ8y	True
12	False	null	False	RT @aveeksen: The Dhaka hostage crisis reminds of Mumbai 26/11. Seems efforts by BNP razakars to convince international press its AQ & IS a…	False
13	False	null	False	@rupasubramanya another mistake of Nehru in not eliminating razakars from Hyderabad & nearby areas.	False
14	False	null	False	RT @iMac_too: Looks like Razakars were prepared for a big bang. https://t.co/NmCDbbJ8y	True
15	False	null	False	@rananth @RakeshSinha01 @asadowaisi sorry,Liaquat was Nizam's PM not leader of Razakars	False
16	False	null	False	RT @KanchanGupta: In which case why does his heart cry over pro-Pakistan Jamaatis and 1971 razakars getting roasted in Bangladesh? https:/…	True
17	False	null	False	Opening statement against 4 Noakhali Razakars | https://t.co/PsV6mNAUQE https://t.co/HqCm68wdJt	False
18	False	null	False	RT @KanchanGupta: Mullah Siddiquallah Chowdhury, WB Minister of Mass Education, has publicly declared support for Jamaati razakars of Bangl…	False
19	False	null	False	RT @KanchanGupta: Mullah Siddiquallah Chowdhury, WB Minister of Mass Education, has publicly declared support for Jamaati razakars of Bangl…	False
20	False	null	False	RT @KanchanGupta: Mullah Siddiquallah Chowdhury, WB Minister of Mass Education, has publicly declared support for Jamaati razakars of Bangl…	False

Figure 5.8: Calc displaying converted csv file.
This view of data is much more intuitive, with field names as headings of columns, and their data beneath the relevant column.

Twitter data can be obtained through three APIs: the *Streaming API*, the *Sample API*, and the *Firehose API*. The Firehose API gives access to all of the tweets on Twitter, but accessing this is expensive. As a result most researchers use one of the other two APIs, both of which provide 1% of the output of the Firehose. The main advantage of the streaming API is that it allows users to filter their sample, by the use of keywords. It has the disadvantage that its results are not a random sample.[9] The results returned by the sample API, *are* a random sample, however it is not possible to filter these results. For this research it was necessary to select data by keyword and so the Streaming API was used.

5.7 Collecting the Data

For this research tweets were collected over two time periods: from Wednesday, 25 May 2016 10:52:22 to Friday, 12 August 2016 17:03:57 a period of 79 days, and Thursday, 18 May 2017 12:11:59 to Wednesday, 21 June 2017 07:47:21 a period of 34 days. The two datasets were denoted D1 and D2 respectively. The D1 dataset was used in both the creation and validation of machine learning predictions of racist tweets, the grounded theory analysis of their content, and the creation of metrics aimed at spotting racist tweeters. The D2 dataset was used as validation in each of these tasks.

To determine which tweets to collect it is possible to use information such as retweet status, replied to and other metadata (for example, see Williams and Burnap (2015)) but the system created for this research, collected tweets that contained potentially racist keywords. The approach to collecting the data was to collect tweets based on keywords taken from the Wikipedia List of Ethnic Slurs page (Wikipedia, 2017). The lack of contextual information in such an approach can make keyword searching of tweets potentially lead to both false-positives and false-negatives (Nobata et al., 2016), however it was decided to take this approach, since this will collect tweets containing a particular type of comment: those that will be labelled as *personal racism* (see Section 3.4.6), although of course racist

[9] Morstatter and Liu (2017) noted that data collection through APIs is also potentially biased, since the Streaming API does not give a random sample.

speech that does not contain these racist keywords will be excluded. One alternative is to take context into account and perhaps look at Ngrams, but this is likely to lead to exclusion of certain tweets and the aim for this research was to include everything at the data collection stage that might be deemed to be racist. This approach might lead to the exclusion of disguised racist words such as 'ni66er' but such words can easily be added to the set of keywords, so it is expected that the results would generalise to other datasets, since the task really amounts to improving the classification based on a set of words.[10]

The data was collected using the Python program data.py (see Appendix A.4), that harvests tweet data from Twitter, and stores it in text files. To access Twitter via its API four variables are required: consumer key, consumer secret, access key and access secret. These variables apply to a particular Twitter account and their values, which are unique for a particular account, are generated by Twitter. The program data.py uses three different sets of these variables, depending on an input parameter being set to one, two or three. The reason for this is that Twitter limits the amount of requests to its API for each particular account, and using multiple accounts allows for more requests. The program creates an object based on Tweepy's[11] StreamListener class, which listens for tweets containing a keyword (or the hashtag version of the keyword) from a set of keywords. The program then writes the data for any tweet it finds to a text file for that particular keyword, for example if a tweet contains 'abbo' it is written to abbo.txt.

There are also limitations on the length of the list of keywords that can be sent to the Twitter API, so data.py has another parameter which takes the values a, b or c, and the value of this parameter determines which set of keywords are tracked. The data is stored in JavaScript Object Notation (JSON) format and an example is given in A.3. The data were collected by a laptop using a broadband connection, which on occasion, lost its connection to Twitter. As a result sometimes incomplete JSON records were recorded. To remove these incomplete JSON records from the data, the resulting text files of both D1 and D2, were processed using the program in A.5. This program reads through a text

[10] There are of course instances of racist tweets that will not contain any of the racist keywords, for example Kwok and Wang (2013) quote a tweet containing a racist joke: 'Why did Obama's great granddaddy cross the road? Because my great granddaddy tugged his neck chain in that direction'. Handling such tweets is discussed in Section 8.7

[11] Tweepy is a Python library for accessing the Twitter API.

file line by line, if a line is successfully read, then it is written to an output file, but if the line read throws an error, then the line is not written to the output file.

5.8 Datasets

The D1 dataset consisted of 83,994,885 tweets. These 84 million tweets were reduced to 41,260,026 when those without values for *utc_offset*[12] were excluded.[13] The D1 tweet data collected consists of 283 text files (listed in Table B.1), which range in size from 2.4 KB (jijjiboo.txt with one record) to 127.6 GB (nigga.txt with 27,733,830 records), which is in agreement with Silva et al. (2016)'s findings that 'nigga' is the most common hate term on Twitter. The counts of the files, when those without values for *utc_offset* were excluded, are given in Table B.2. The removal of *utc_offset* resulted in 169 files ranging in size from razakars.txt with eight records to nigga.txt with 15,432,550 records.

The D2 dataset consisted of 28,766,811 tweets. These 28 million tweets were reduced to 13,870,269 when those without values for *utc_offset* were excluded.

From the D1 dataset 84,000 tweets were randomly sampled for human annotation as to whether the tweets were racist or not. This process is described in Section 5.10.

How these data were stored is described in the next section.

[12] UTC is the SI measure of time. It stands for Coordinated Universal Time and from a practical viewpoint is the same as Greenwich Mean Time (Nelson et al., 2001). *utc_offset* is Twitter's data field that stores the time in seconds between the hour of *created_at* and the user's time zone. The setting of time zone by the user is voluntary and so *utc_offset* may not have a value.

[13] Tweets contain the attribute, *created_at*, that contains the UTC time that the tweet was created. When analyzing hour of day, UTC time is not sufficient, since if a tweet is sent from a country that is not using UTC time, then the actual time of the tweet within that country will be *created_at* plus or minus an offset determined by whichever time zone the country is in. It was decided to exclude tweets that did not contain values for *utc_offset* since, if that value is not set, the tweets could be originating from any country, and the actual time they were sent could be up to 12 hours different from that recorded in *created_at*. See Section 6.4.3.5 for further discussion of temporal features of tweets.

5.9 Storing the Data

As mentioned in 5.3 big data scale problems are suited to using Hadoop clusters.[14] Hadoop can be run on a single machine in *pseudo-distributed mode*, but this is only suitable for testing purposes, since Hadoop jobs add overhead to processing and these are likely to run slower than normal processing on a single machine. Hence Hadoop should be run on a cluster in *fully distributed mode.* For this research four machines were connected via a switch and a Hadoop cluster was installed on them.

The most widely used operating system for Hadoop is Linux; although it is possible to run Hadoop on Windows. As a result Ubuntu, (a Linux distro), version 14.04 was used for each machine's operating system.

There are a number of ways of implementing Hadoop. It can be installed on its own or as part of other software that is designed to make the administration and running of Hadoop easier. The three main vendors of this type of 'packaged Hadoop' are: Cloudera, Hortonworks and MapR. During this research Hadoop was installed on its own, and later as part of Hortonworks Data Platform (HDP).[15] Even with HDP Hadoop is difficult to install and run, and upgrading components can also be taxing.

The HDP cluster allowed distributed storage of data using HDFS. Data stored thus can be accessed using HDFS commands run via an HDFS script (Foundation, 2015). For example the following command:

```
hdfs dfs -ls /
```

lists the files under the HDFS root directory. HDFS's commands are fairly limited, so for this research Hadoop's data warehousing application Hive was used. Hive allows for SQL-like access to data, thus providing easy access and manipulation of the data.

[14] A cluster is a number of machines networked together.

[15] HDP was chosen since it is free, open source and has sophisticated Hadoop administration tools.

D1 and D2 were imported to Hive and stored as Optimised Row Columnar (ORC) tables, because ORC tables are more efficient than the default text tables[16] (Leverenz, 2017). One drawback of using the ORC format is that you cannot directly load text files into an ORC table. Instead text files must be loaded into an interim Hive text table which then can be dumped into an ORC file (rich, 2016).

This stored data was then annotated as discussed in the next section.

5.10 Annotating and Classifying the Data

The annotation of racist speech by humans, is a difficult task subject to high levels of subjectivity (Ross et al., 2017; Olteanu et al., 2017). For example Kwok and Wang (2013) looked at racist tweets directed against Blacks. They asked three students of the same age and gender but different races, to determine whether they thought each of a sample of 100 tweets were offensive, and how offensive on a numerical scale. They found only 33% agreement among the students.

Human classification of racist tweets is problematic since it is reliant on an individual's interpretation of small amount of text. Such interpretations will, of course, be subjective and will depend on a myriad of personal attributes, exposures and other factors. While subjectivity cannot be completely excluded, it can be reduced by utilizing a list of criteria that must be followed by any coder classifying text. For this research the rules followed by Waseem and Hovy (2016, p.89) were adapted.[17] The rules used were as follows:

A tweet is racist if it meets any of the following criteria:

1. Uses a racial slur.
2. Attacks an ethnic minority.

[16] The ORC format also allows updates of the data, whereas storing data in the default text format does not.

[17] Their rules were successfully applied to classify tweets containing hate speech. They are a good set of rules since they minimise ambiguity on the behalf of the annotator, they are inherently coherent, and when applied consistently by different annotators, should produce consistent annotations.

3. Seeks to silence an ethnic minority.
4. Criticises an ethnic minority, without a well founded argument.
5. Criticises an ethnic minority and uses a 'straw man' argument.
6. Blatantly misrepresents truth or seeks to distort views on an ethnic minority with unfounded claims.
7. Shows support of problematic hash tags. E.g. '#BanIslam', '#whoriental', '#whitegenocide'
8. Negatively stereotypes an ethnic minority.
9. Defends xenophobia.
10. Contains a screen name that is offensive, as per the previous criteria, the tweet is ambiguous (at best), and the tweet is on a topic that satisfies any of the above criteria.

The files were hand-annotated by six annotaters: ED, MH, AR, AM, JB, and GL although MH only did a small pilot sample that was not used in the final analysis. Each tweet's text was examined and determined to be racist or not. The Tweets were extracted randomly from the D1 files and given to annotaters in batches of 1,000 tweets. In total 84,000 tweets were classified by the annotaters.

To determine what the interrater reliability was at the beginning, middle and end of the annotation process, the same three datasets were given to each of the annotators. These datasets were 100 tweet samples from nigga.txt, nigger.txt and wop.txt. At the beginning of the annotation process the annotators were given the sample from nigga.txt, in the middle (after they had completed 30 hours annotation) they were given the sample from nigger.txt, and at the end they were given the sample from wop.txt. The annotation scores (either zero or one) were placed into SPSS and Krippendorff's α[18] was calculated for each of the three datasets, using De Swert's (2012) SPSS macro. Krippendorff's α was 0.60, 0.92, and 0.90 for the nigga.txt, nigger.txt and wop.txt annotations respectively. The α score for nigga.txt was low, albeit just in the acceptable range (ibid.), whereas the α score for both nigger.txt and wop.txt was high. Overall this shows good interrater

[18] Krippendorff's α is a measure that is commonly used in the literature to compare interrater reliability (Ross et al., 2017). It is a complex amalgam of other statistics that measures disagreement between raters (Krippendorff, 2004).

reliability. The difference between the score for nigga.txt and the scores for both nigger.txt and wop.txt, may be due to practice effects, that is the raters are becoming better at determining whether tweets are racist or not. Another possibility is the difficulty of determining whether tweets containing the word ‘nigga’ are racist or not, and the relative ease of the words ‘nigger’ and ‘wop’. There is more discussion on the difficulty of determining the meaning of the word ‘nigga’ in Section 3.4.5.

As can be seen in Table 7.1 the racist tweets were in a minority (approximately 4.9% of the hand annotated sample and 2.4% of the predicted data) and the implications of this are discussed in the next section.

5.11 Imbalanced Data

The data for this research exhibited *class imbalance*, that is the number of instances in each class is very unequal, in this case there were many more nonracist tweets than racist ones. This is a common occurrence in machine learning and is problematic since it is often the minority class that is of interest. This can mean predictions based on unbalanced data under-exaggerate the importance of the minority class. There are commonly three ways used in the literature to overcome this difficulty: under sampling of the majority class, over sampling of the minority class, or a hybrid method. Simple random over sampling was used for this research since this often outperforms under sampling with the caveat that it can lead to overfitting (Wallace et al., 2011). Details of the oversampling are given in Section 7.1.1.

The data was used in machine learning and grounded theory methods and the machine learning approach is discussed in the next section.

5.12 Machine Learning

This section discusses the use of machine learning approaches[19] and their evaluation using various metrics. Models were created using seven different algorithms: NB, LR, SVM, RF, DT, GBTs and ANN. These models were used to build 'racism detection models' on a dataset of tweets and a comparison of models is given in Section 7.1.4 along with a description of the Twitter-sourced dataset in Section 5.9. The models were compared against one another using standard metrics to determine the best model for determining racist tweets. They were also used to evaluate the efficacy of using tweet metadata as features in the predictive models. Further details are given in Section 6.4.

Whichever model is used it can be modified as discussed in the next section, in order to optimise certain metrics which are discussed in Section 5.12.2.

5.12.1 Model Tuning

Models have both *parameters* and *hyper-parameters*. Parameters are numerical aspects of the model that are not normally changed by human intervention, whereas hyper-parameters are set by the user. An example of a parameter for an SVM model would be the support vectors generated by the model, and an example of a hyper-parameter is the number of features used by the model. Hyper parameters can be modified to optimise a model. Optimization of models can be evaluated by examining their residuals, that is the difference between their predicted values and the corresponding actual values. Model residuals can be used to calculate various evaluation metrics, which are discussed in the next section.

[19] The machine learning algorithms were run consecutively within one Zeppelin paragraph or using the spark-shell CLI (see figure A.1 for an example program).

5.12.2 Evaluation of the Classifiers

In order to determine which is the ‘best’ model, the possible outcomes for the model need to be considered. For supervised binary (two possible classes) classification problem such as this, there are four possible outcomes:

- True Positive (TP) - data is correctly classified as positive.
- True Negative (TN) - data is correctly classified as negative.
- False Positive (FP) - data is incorrectly classified as positive.
- False Negative (FN) - data is incorrectly classified as negative.

These are summarized in Table 5.1.

Table 5.1: Possible outcomes for binary classification.

		Predicted class	
		1	0
Actual class	1	True positive	False negative
	0	False positive	True negative

For this research positive will mean ‘racist’, that is positive is being used to mean category of interest has been found for this piece of data. From these four outcomes, a number of metrics can be derived. The metrics that were used are described below.

5.12.2.1 Accuracy

This is simply the percentage of predictions that are correct. That is:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}.$$

At first sight this seems like a good measure of a classification process, however skewed data can highly influence accuracy. For example if a dataset contains 95% non-racist tweets then simply classifying every tweet as non-racist would give an equivalent accuracy of 95% (Spark, 2017a).

5.12.2.2 Precision

Precision is the ratio of true positives to all those classified as positives, whether they are correct or not, i.e.

$$Precision = \frac{TP}{TP+FP}.$$

So if ten observations are analysed and three are predicted correctly as positive, and two incorrectly as positive, then precision = 3/5 = 0.6.

5.12.2.3 Recall

Recall is the ratio of correctly classified positives divided by the total number that are actually positive, i.e.

$$Recall = \frac{TP}{TP+FN}.$$

So for the same example, assuming there is one false negative, then recall equals 3/4 = 0.75. The use of both precision and recall is advantageous when the two classes are not distributed evenly. This is certainly the case with the racist data as can be seen in Table 7.1, there being far fewer racist tweets (positive ones) than nonracist ones.

Precision and recall are concerned with the performance of positive classification, as opposed to the classification of negatives. Precision is a measure of the proportion of positives that are correct, and recall is a measure of the proportion of positive events that were correctly predicted.

5.12.2.4 Area Under the Precision-Recall Curve (AUPRC)

AUPRC is the area under the curve of precision against recall (Powers, 2011).

5.12.2.5 Area Under the Receiver Operating Characteristic curve (AUROC)

AUROC is the area under the graph of true positive rate versus false positive rate (ibid.).

5.12.2.6 F-score

F-score is the harmonic mean of precision and recall, which for precision, p and recall, r, is expressed as:

$$F_\beta = (1+\beta^2) * \frac{\mathrm{p} * \mathrm{r}}{(\beta^2 * \mathrm{p}) + \mathrm{r}},$$

where the importance of recall is β times the importance of precision (Goutte and Gaussier, 2005).

There is little discussion of the value of β within the racism prediction literature, for example Bermingham and Smeaton (2011) use F-score without discussing β, Kwok and Wang (2013) only use accuracy as a metric and Burnap et al. (2015) use $\beta = 1$ without comment. Since there is no guidance in the literature, it was decided to use the default values given by Spark, $\beta = 1$, giving equal weight to precision and recall, and $\beta = 0.5$ giving recall twice the weight of precision. This approach seems reasonable since $\beta = 1$ is the value used in much of the literature, and $\beta = 0.5$ since it seems better to have a higher recall that is a better detection rate, albeit at the expense of precision, that is giving a higher false-negative rate. As can be seen in Table 7.2 F-score for the different values of β follow similar patterns, as do AUPRC and AUROC so the value of β was not critical in model selection.

The problem with the evaluation metrics discussed above, is that they do not evaluate a classifier with respect to how well it performs with unseen data. Cross validation is an evaluation method that overcomes this issue, and is discussed in the next section.

5.12.2.7 K-fold Cross Validation

Cross validation involves the removal of part of the dataset, then the model is trained and finally the model is fitted against the removed data to evaluate how well it does on 'new' data. This can be performed by what is known as the *holdout method* which separates the data into a training set and a test set, and the classifier is run against only the training set data. The resultant model is used to predict the test set data, which was not used to create the model, and then its residuals can be examined to calculate metrics. This method is subject to high variance, since it is reliant on how the data is split into the two datasets. *K-fold cross-validation* improves on this method by splitting the data into K subsets, and on each of these the holdout method is run, each time one of the K subsets is used as the test set, and the other K-1 subsets are combined as the training data (Schneider, 1997). The data for this research was split in the ratio 70:30, that is 70% training and 30% test, and tenfold cross validation was used.

In the next section the qualitative analysis will be discussed.

5.13 Qualitative Analysis of the Tweet Data

in addition to the qualitative machine learning analysis of the data, a qualitative analysis was also performed. A preliminary analysis was first conducted, in order to gain an understanding of the underlying qualitative nature of the data. To do this words and Ngrams were counted in order to provide a summary of the vocabulary of the data. The most frequently occurring words were also examined visually using word clouds. Sentence structure was analysed using word trees, giving a visual representation of the structures of the frequently occurring sentences.

After this initial analysis a grounded theory analysis was performed. The aim of this was to extract thematic concepts from the data. These processes are discussed in the following sections.

There now follows a theoretical discussion of the qualitative analysis. The act of tweeting is about self-production, regular updates which Twitter encourages, become part of a user's identity (Murthy, 2012). Although the content of many tweets might be considered banal, even such banality can be regarded as having important meaning regarding a user's identity, for example, posts about what somebody had for breakfast can be interpreted as 'I exist' type messages (Bourdieu, 1984; Murthy, 2012). At first sight the data in this research, is anything but banal. Many of the tweets express rage and hatred. In order to determine whether this is an accurate picture of the popular racist tweets dataset as a whole, a better understanding of the data is required using qualitative methods. Since a deeper analysis requires some theoretical exploration of the nature of these Twitter conversations, a brief discussion of Goffman's work on the analysis of conversations and Habermas's idea of the public sphere will be given.

Murthy (2012) applied Goffman's (1981) framework for understanding conversation, to the digital realm. Goffman's (ibid.) analysis argues that 'talk' is made up of: *ritualization*, *participation framework*, and *embedding*. Ritualization, according to Goffman (ibid., p.2), is how the 'movements, looks, and vocal sounds' people use as gestures in interaction are acquired over their lifetimes, and how they have meanings. Participation framework refers to the various participants that are within range of 'talk' and their conduct which is subject to normative expectations of what is appropriate. Embedding refers to the fact that in conversation the words used may not be those of the speaker; individuals will quote others either directly or indirectly and the subject of their speech is not necessarily about them. Although Goffman was analysing direct talk, that is talk that is not mediated, Cetina (2009) extends the ideas of Goffman to mediated digital 'talk' (Murthy, 2012). Murthy (ibid.) further applies Goffman's ideas to the digital realm, this time to Twitter. For Murthy (ibid.) there is ritualization in the symbols and text used by Twitter users. These might include avatars, emoticons and text signifiers such as ellipses signifying pauses, or exclamation marks signifying surprise or dismay. This textual and graphic information takes the place of visual gestures. The perceptual range of off-line talk is usually fairly obvious to the talker, but this is not the case for Twitter, since somebody sending an original tweet does not know in whose feed it will appear. There are, however, participation statuses within Twitter, in relation to a tweet, some might retweet a tweet, others might respond directly to it, some might read it without response and others might simply ignore it.

Embedding is certainly a feature of Twitter, due to the mechanism of retweeting. Although retweets contain a reference to the original tweeter, this is often subconsciously or consciously ignored and thus the subject of the tweet has been altered in some people's perception (Murthy, 2012). Twitter is different in other ways when compared to off-line talk, tweets that are re-tweeted may be embedded in a different time frame and audience. That is a tweet that is retweeted five hours after the original tweet, may have a primary audience in a different time zone, which may have been completely unintended and unexpected by the original tweeter (ibid.).

The complexities of Twitter's participation framework can be analysed using Habermas's (Habermas, 1989; Habermas et al., 1974) notion of the 'public sphere', which is the fora that make up public opinion, open to all without restriction. The notion of Twitter as a public sphere can be seen in the work of Lotan et al. (2011), who examined tweets relating to the both the Tunisian and Egyptian uprisings in 2011. They wanted to understand the role of different kinds of social actors in spreading information during significant world events such as these. They collected 168,663 tweets with keywords '#sidibouzid' or 'tunisia' for the Tunisian uprising and 230,270 tweets with keywords 'egypt' or '#jan25' for the Egyptian one. From this data they selected 963 users that were the first to tweet or were retweeted at least 15 times, and classified these by hand as to their user type, such as mainstream media organisation, mainstream media employee, bloggers and so on. Examining the distributions of the tweets in terms of user type they found that most of the original tweets came from journalists and activists and most of the retweeting of these was performed by bloggers and activists. It is dangerous to read into any causal relationship between this tweeting activity and the actual revolution, since at the time there were very few Egyptian or Tunisian Twitter users, and it seems more likely that many of the tweets were sent by Western observers (Murthy, 2013).

However Fuchs (2017) argues that Twitter is not a public sphere as Habermas imagined one. Habermas used Marxist theory to determine that a correctly functioning public sphere would have amongst its key priorities the aim of equality, but Fuchs (ibid.) argues that Twitter has no such mission, being a private profit making organisation, and that its content may provoke social change and be on the side of, as he sees it, virtuous left realism, but on the other hand also contains right-wing extremism and other socially destructive ideas.

These theoretical constructs are useful in framing a qualitative analysis of Twitter. Now the data and methods used in the analysis will be discussed.

Although computer resources were used the qualitative analysis was more reliant on human intervention than the machine learning procedures. So the data for the qualitative analysis needed to be small enough for human analysis, and hence a sample was used. The data used in the qualitative analysis was the tweets that were flagged as racist i.e. they are in the dataset Pred R. From these tweets, it was decided that only those tweets that were influential would be analysed. To determine tweets that were influential, only those that had been retweeted at least once were included in the qualitative analysis of the tweet data.

5.13.1 Initial Qualitative Analysis

A popular method used to try and extract meaningful insight using qualitative data, is that of grounded theory, and this will be the primary tool used to analyse this dataset. However before this is utilised, it is useful to gain an understanding of the data's nature, including its vocabulary, its most frequently occurring words, bigrams and trigrams. While it is difficult to distil thousands of tweets into a summary that gives a true representation of the data, frequency analyses, word clouds and word trees are useful tools in attempting this.

An initial analysis of these data was performed using standard NLP modules provided by the Natural Language Toolkit (NLTK), which allows users to write programs in the programming language Python for NLP tasks (Bird et al., 2009). NLTK has a number of built in functions that make the processing of text a relatively straightforward task. NLTK was used to produce the Top 25 most common: words excluding stopwords, hashtags, bigrams and trigrams from the tweets in Pred R that had been retweeted more than once. The results are discussed in Section 7.3.1.

To attempt to get a 'visual understanding' of the data with regard to its popular words, it was uploaded into the software package Nvivo, which was used to generate *word clouds*. Word clouds are a popular way to summarise text by displaying words that commonly occur in a corpus. Words are displayed in a random 'cloud' and their font size is proportional to their frequency (Heimerl et al., 2014). Nvivo has five levels of grouping of words. The first level is exact matches only, the next level of grouping for Nvivo is for words with the same stem, then synonyms, then specialisations and finally generalisations. Analysis of the word clouds is given in Section 7.3.2.

Nvivo also provides the facility to produce *word trees*. A word tree is a visual representation of a word and the contexts in which it occurs. The word tree takes the word of interest as its central point and from this branches emanate, each branch being a particular context that the word was found in. A branch's font size is proportional to the frequency of the word or phrase that belongs to the branch. The length of the phrase in a branch can be altered; there is a trade-off between losing information and the graphic becoming too unwieldy. The default length of the phrase in a branch is five words and the word tree branches in this research were limited to five words. Analysis of the word trees is given in Section 7.3.3.

In addition to the frequency analyses, word clouds and word trees, prior to the thematic analysis of grounded theory, tweet examples were analysed discursively, to give an idea of the kinds of data making up the dataset. The results of this analysis are given in Section 7.3.4.

5.13.2 Grounded Theory Analysis

The analyses discussed so far have concentrated on statistical and visual summary measures of influence. These measures and images provide a good summary of the content of these data. However they lack theoretical insight into the content of the messages. While this is a good first step in understanding these data, the technique of grounded theory was used to provide a deeper theoretical understanding of the tweets. Grounded theory was originally formulated by Glaser and Strauss (1967). At the time of grounded theory's conception much research was aimed at supporting existing theories, rather than generating new ones. There was a move towards the use of quantitative methods which often only used qualitative data in preliminary studies. Glaser and Strauss (ibid.), wanted to challenge these types of research ideas, in particular they wanted to challenge:

> (1) the arbitrary division of theory and research; (2) the prevailing view of qualitative research as primarily a precursor to more 'rigorous' quantitative methods by claiming the legitimacy of qualitative work in its own right; (3) the belief that qualitative methods were impressionistic and unsystematic; (4) the separation of data collection and analysis phases of research; and (5) the assumption that qualitative research only produced descriptive case-studies rather than theory development (Charmaz and Belgrave, 2007, p.29).

To do this Glaser and Strauss (1967) created a method of qualitative research, consisting of a series of steps to be followed, which was in contrast to much of the qualitative research at the time, which was very much free-form and unsystematic (Charmaz and Belgrave, 2007). Later Strauss and Glaser diverged in their conceptualisations of grounded theory, and this divergence culminated in the publication of Strauss's conceptualisations of grounded theory in Strauss and Corbin (1990). Engward (2013, p.39) provides a summary of the differences between the two main conceptualisations of grounded theory and these are given in Table 5.2. While there are some differences in the two approaches, the similarities are perhaps more striking.

The original conceptualisation of grounded theory in Glaser and Strauss (1967), whilst

Table 5.2: Differences between Glaser and Strauss (1967) and Strauss and Corbin (1990), from Engward (2013, p.39).

Glaser and Strauss (1967)	**Strauss and Corbin (1990)**
Starts with a general idea of where to begin	Starts with a general idea of where to begin
Uses neutral questions	Uses structured questions
Development of conceptual theory	Conceptual description (description of situations)
Development of theoretical sensitivity (the ability to perceive variables and relationships) from immersion in data	Development of theoretical sensitivity from methods and tools
Theory is grounded in the data	Theory is interpreted by the observer
A basic social process should be identified	Basic social processes need not be identified
The researcher is passive, exhibiting disciplinary restraint	The researcher is active
Data reveals theory	Data is structured to reveal theory
Coding and continuous comparison of data enable patterns to emerge	Coding is defined by technique, leading to micro-analysis of data word by word
Uses two coding phases to develop concepts that expain the phenomena: simple (breaking data down into small segments and group into similarities that begin to describe patterns in the data) and substantive (open or selective choosing of a core category and relating other categories to it to expore emergent patterns)	Uses three types of coding: open (identifying, naming, categorising, describing phenomena), axial (the process of relating codes to each other) and selective (choosing a core category and relating other categories to it)
Regarded as the "true" grounded theory	Regarded as a form of qualitative data analysis rather than grounded theory

stressing systematic qualitative research, also allowed room for some freedom in the analysis, and this can be seen in the various re-conceptualisations of grounded theory. This freedom is appealing from a pragmatic viewpoint, so grounded theory sits well within the overall methodology of this research. The analysis of the tweets by grounded theory, in which concepts and categories emerge only from the tweets' textual data (Strauss and Corbin, 1990), provided a richer notion of the meanings within the tweets, and facilitated the uncovering of themes that are ignored by the numerical analysis. In addition, grounded theory provides the option of ultimately organising the emergent themes into theories by a coding process that defines categories and identifies groupings of categories. The identification of categories begins the process of definition with explication of the category properties, conditions under which the categories arise, and relations of the category to other categories. Finally axial coding identifies over-arching categories within the data.

An example of the use of grounded theory in analysing tweets is given in Roberts et al. (2017), who performed axial coding based on Strauss and Corbin (1990)'s grounded theory approach. They analysed 17,000 tweets related to the Woolwich murder of Fusilier Lee Rigby. They aimed to identify 'whether the content was extremist, expressing far-right/far-left/Islamist views; the presence of emotional attributes including fear, anger, shock, revulsion; and whether it could be identified with a particular group' (Roberts et al., 2017, p.7). They also took a subset of of 2,000 of these tweets which they analysed further to 'obtain insights into the complex interactional dynamics observed taking place between different ideological groups and positions (ibid., p.7).

During the grounded theory process, memos are kept during all stages of analysis to document any change in category/concept labels. Constant comparison techniques are employed throughout data analysis in order to fully explore differences and similarities between tweets.

The aim of the research was to develop a grounded theory of the content of the tweets predicted as racist. A sample of tweets selected from the Pred R was used. The tweets that had been retweeted at least once were sampled using a simple random sample of

2,000 tweets. Concepts were sought from the text. From the concepts categories were determined. These categories were themselves grouped into a few main categories. The analysis was continued until 'theoretical saturation' was reached. The analysis was a process of constant re-evaluation and concepts and categories constantly evolved during the analysis. Pictures and other graphics were available by following URIs from some tweets. The images that the tweets point to are a potentially a rich source of information, however for this analysis they were only used as clarification, when a tweets meaning was ambiguous and concepts could only be determined with the addition of image analysis.[20]

A key concept of grounded theory is theoretical sensitivity, which Strauss and Corbin (1990, p.41) define as 'an awareness of the subtleties of meaning of data'. Researchers using grounded theory methodologies need to be able to bring their expertise to analysing the data, while at the same time being sceptical about inferences drawn from that expertise, and to think creatively rather than exclusively drawing on their existing knowledge. So for this study an attempt was made to categorise creatively according to the data but also at the same time not to run away with the creative process and try and retain the reality of the tweets' messages. When analysing the data order effects had to be taken into account. It would be ideal if the conceptualising could have been done with sufficient time lag in order to reduce the influence of concepts derived from one tweet leading to the same concepts being fresh in the mind of the coder when looking at another tweet, but practically this was not possible. Similarly grounded theory requires an open mind when performing open coding. It requires expert knowledge and reflexivity at every step of the process. This is hard to do as it is a very time-consuming process and 'concept fatigue' soon sets in. As such no more than 200 tweets were coded in any one day. Karstedt (2001, p.285) warns, 'cultural comparisons often suffer from exaggerations of differences' and as such the researcher aimed to try and avoid seizing upon any minor differences

[20] The imagery was analysed to gain an understanding of the likely purposes behind these visual displays (Silverman, 2015). In order to understand the work done by this visual imagery, the semiotic analysis method of Barthes (1978) was used to process the visual data. This method involves trying to attain the underlying symbolic meaning of an image. To do this an attempt is made to avoid the obvious message an image portrays and not just take it at face value. The analysis is performed on images that are grouped into similar groupings, wherever possible, rather than on individual images. Then the analysis aims to identify the signs that constitute the images' sign system and what messages that system is conveying. The messages thus derived aided in determining the list of grounded theory concepts. Care needs to be taken with such an analysis of symbolic meanings, since for example, British researchers, may interpret an American symbol differently than the culturally specific meaning that was intended.

between tweets from different nations when determining concepts and categories from the data. To validate the researcher's success in this and other attempts at objectivity, an uninvolved person experienced in grounded theory was asked to independently code the same set of data and discuss discrepancies found, in order to seek independent validation of the findings. Since concepts and categories which are included in the final theories may not emerge until some way through the analysis, it was important to determine that final theories were checked against the earlier data. A post hoc frequency count of concepts was retrospectively used to assess support for the final theory across the sample. The results are discussed in Section 7.3.6.

Up to this point the methodology has discussed the identification and analysis of racist *tweets*. In the next section the focus changes to the identification and analysis of racist *accounts*.

5.14 Accounts

There are 39,082,290 accounts in D1[21] when utc_offset is not null, and of these 105,211 are linked with predicted racist tweets. So the predicted racist tweets were created by only 0.27% of the accounts. However despite being a small percentage 105,211 accounts are far too many to be handled without some form of automated identification as to which of these accounts are the most likely to warrant further investigation as possibly racist accounts.

The automatic identification of racist tweets is the first step in identifying problematic racist accounts. While it is important to identify the racist content, and this in itself is useful in order to be able to potentially remove or filter this content, an automated system must also be able to identify tweeters that might be racist. Of course racist tweeters might be of interest for a number of reasons, for example as a result of obscene or threatening content but the intention of this research was to identify the racist accounts that have the most influence over the network. There is a consensus view that within a social network,

[21] D1 is the dataset of 84 million tweets collected in summer 2016. For further details see Section 5.7.

levels of influence vary and some accounts are more influential than others in the spreading of information. There is also a consensus that the amount of attention given to a user's content on Twitter is determined by two factors: the status and popularity of the user and how their content propagates through the network (Romero et al., 2011).

In order to find measures suitable to identify such accounts, the standard metrics of influence[22] are unsuitable. These measures are aimed at finding which are the most

[22] There is no consensus on which measures and aspects of a network best identify influential accounts (Pei et al., 2017). The spread of information on social networks, usually takes the form of cascades: starting with a few 'seed' accounts, then it spreads from contact between accounts and may ultimately defuse 'virally' and reach virtually the whole community (Pei et al., 2014). Identification of the seed accounts was the focus of this research, since there is much research and focus on *influencers* that involves their identification and stresses their importance in *information cascades* (Pei et al., 2017). Hence, it was decided to focus less on the spreading of the data and instead to evaluate efforts on determining which of the tweeters were most likely to be from racist accounts.

The identification of influencers is performed either solely by analysis of aspects of a social network's topology (the structure and connections in the network) or by a topology analysis and assumptions about aspects of how information spreads dynamically in the network (ibid.). There are many measures that aim to determine a node's level of influence in spreading information in a network.The most common predictors are *degree*, *PageRank*, *betweenness centrality*, *closeness centrality*, *eigenvector centrality* and *k-core*. Degree is the number of direct connections that a node has in a social network, intuitively this seems like a good measure of influence, since seemingly the more connections you have the more chance there is that you will pass on information. As well as being intuitively appealing, degree is also trivial to calculate, however its main drawback is that it only takes into account immediate neighbours. Since most information spreading is in the form of cascades, highly influential accounts can occur at the periphery of a network, their messages being picked up by a few accounts and then amplified as the information spreads throughout the network, thus negating the usefulness of degree (ibid.).

PageRank is the algorithm developed by Google used to rank importance of World Wide Web pages, although it can also be used in a number of ranking situations such as node importance.

Betweenness centrality of a node is a number of shortest paths linking two other nodes that pass through it.

Formally a social network topology can be represented by a graph $G = (V, E)$, where, V represents the set of nodes, and E represents the links between nodes, or edges. the number of edges in the shortest path between v_i and v_j is $dist(v_i, v_j)$ that is the length of the geodesic between v_i and v_j. The degree of node, v_i is $d(v_i)$. Then betweeness centrality is defined as:

$$C_B(v_i) = \sum_{k=1}^{n} \sum_{j:j>k}^{n} \left[\frac{g_{kj}(v_i)}{g_{kj}} \right], v_i \epsilon V,$$

where g_{kj} is the number of shortest paths which connect v_k and v_j and $g_{kj}(v_i)$ is the number of these which pass through v_i (Ruhnau, 2000).

Closeness centrality is defined as:

$$C_C(v_i) = \sum_{j=1}^{n} dist(v_i, v_j), v_i \epsilon V.$$

Like other measures based on closeness, closeness centrality requires a large amount of computation to determine shortest parts, and so is not practical to use for BD scale data.

Eigenvector centrality is a measure based on the assumption that a node's influence is determined by their neighbouring nodes spreading capabilities (Pei et al., 2017). If the *adjacency matrix* is denoted as $A = (a_{v,t})$, i.e. $a_{v,t} = 1$ then if node v is linked to node t, and $a_{v,t} = 0$ otherwise, then the relative centrality score of node v can be defined as:

influential Twitter accounts, and so are all aimed at identifying prolific tweeters, or at least tweeters whose messages spread prolifically. While they could be useful in terms of racist research, it is perhaps unrealistic to treat racist accounts in the same way. It is a relatively easy task to monitor known potentially problematic accounts from far-right organisations such as the English Defence League, but this research aims to find accounts that are somewhat 'under the radar', so any indicators of interesting accounts for this research, will need to spot 'influential' accounts that have relatively little activity, yet high levels of racist activity. These measures are likely to be unsuitable for the racist accounts that are of interest, since these are relatively small-scale accounts, the maximum number of racist tweets sent from any single account was 256. Additionally, since it is desirable to avoid using API calls to Twitter,[23] much of this data is unavailable, for example whilst degree is obtainable from a tweet's followers_count, the other measures of centrality all require access to the graph connections within Twitter.

As a result it was necessary to develop new measures. It was decided that any measure should be created from just the tweets data, excluding API calls to Twitter

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t,$$

where $M(v)$ is a set of adjacent nodes of v and λ is a constant. This can be rewritten as

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x},$$

(Newman, 2008).

K-core is a measure that is determined for a node by iterative pruning of nodes. All nodes of degree one are given the k-core score of one, these are then removed and the process is repeated with the new nodes with degree one given a k-core score of two, and so on.

The seminal work on showing that k-core can be a better indicator than degree is not always a good indicator of influence is that of Kitsak et al. (2010). They investigated four networks: the friendship network between 3.4 million members of the LiveJournal.com community, the network of email contacts of University College London's Computer Science Department, the contact network of inpatients from hospitals in Sweden, and the network of actors who have co-starred in movies labeled by imdb.com as adult. They performed simulations using Susceptible-Infectious-Recovered (SIR) and Susceptible-Infectious-Susceptible (SIS) models. For SIR individuals become infected with probability, β, and once infected can recover with probability, μ, and become immune, which is the same for SIS but with the exception that an individual does not become immune, but becomes susceptible again with probability, λ. They they found k-score was a reliable indicator of influence, but the spreading from nodes of equal degree differed depending on where the note was situated, central nodes would cause different spreading patterns to peripheral nodes. Unlike degree, k-score takes into account connectedness of a account's connections. If a network is made up of a node of degree 6 connected to 6 nodes of degree one, the central node's k-score is 2 and it can only pass messages to 6 other accounts. However if another network is made up of a central node of degree 6, connected to 6 nodes of degree 5 (each of these connected to 5 single nodes) its k-score will be 3 and it can pass messages to 30 other accounts.

[23] Any API calls to Twitter will take time and thus slow a system and can cause problems with rate limits set by Twitter.

which can be time-consuming and expensive (Riquelme and González-Cantergiani, 2016). As a result this research avoided using network connections, so following and followers relationships were unavailable. This is not necessarily problematic to any great extent, since, as has been mentioned before, Twitter is an unusual social network, in that the follower-following relationship is complicated by the use of hashtags for searches. In fact it may be of more interest to investigate how racist messages are being seen by accounts which are not following the original tweeter, although on the other hand the following-follower relationship does give an insight into relationships on the network. However this research, in using just the text and metadata from the tweets, takes a more fluid approach, not relying on network structures to determine how racism is being disseminated.

Whilst accounts that send a high number of racist tweets might be of interest, it is possible that the number of retweets of tweets from racist accounts is more of interest since the act of retweeting suggest more interaction with the message, and also aids in its dissemination. The maximum number of retweeted tweets for any of the accounts that send multiple predicted racist tweets was 10. Obviously this measure, the number of racist tweets that were retweeted, is a possible indicator of a racist account. This factor could simply be tabulated to identify the 'most racist' accounts. However this does not take into account the scale of retweeting, and instead the average number of retweets per racist tweets might be a better metric. It needs to be determined how well these different measures identify the busy racist accounts and give less attention to the less busy racist accounts and nonracist accounts.

To clarify, in order to determine which potentially interesting accounts might warrant further investigation four measures were used: the total number of original tweets from an account, *oCount*, the number of distinct tweets that were retweeted from an account, *rDistinctCount*, the total number of retweets of an account's tweets, *rCount*, and the *retweetRatio* which is calculated as,

$$retweetRatio = \frac{rCount}{rDistinctCount},$$

(if a account has been retweeted three times, but of the three there are only two distinct tweets then their *retweetRatio* would be $3 \div 2 = 1.5$).[24]

These concepts can be clarified with the following example. Hypothetical data is given in Table 5.3.[25] The table shows a number of tweets with tweet IDs from 1 to 17. It shows an account's name, 'Jim' having sent a tweet with TweetID 1, Mark sent a tweet with TweetID 2 and so on. If a tweet is a retweet it also shows the account which sent the original tweet. *RO* denotes whether a tweet was an original (denoted as 'o') or a retweet (denoted as 'r') . If a tweet is a retweet then the column *OrigtweetID* is the original tweet's TweetID. It can be seen that both Jim, Mark and Dave sent only original tweets, Alice sent four retweets and one original, Bill sent one of each and Charles sent two of each. Table 5.4 shows the metrics *oCount*, *rDistinctCount*, *rCount*, and *retweetRatio*. It can be seen that Jim had one tweet retweeted twice so his *retweetRatio* equals $2 \div 1 = 2$. Mark on the other hand, had two of his tweets retreated, one was retweeted once and one was retweeted twice, his *retweetRatio* equals $3 \div 2 = 1.5$. This shows that although Mark was retweeted more times than Jim, but Jim had a higher *retweetRatio*, so the *retweetRatio* is a measure of the popularity of tweets of the account. To calculate these

Table 5.3: Example tweet data.

TweetID	**Account that sent tweet**	**retweet of account**	**RO**	**OrigtweetID**
1	Jim		o	
2	Mark		o	
3	Mark		o	
4	Mark		o	
5	Dave		o	
6	Dave		o	
7	Alice	Jim	r	1
8	Alice	Mark	r	3
9	Alice	Dave	r	5
10	Alice	Alice	r	9
11	Alice		o	
12	Bill	Mark	r	3
13	Bill		o	
14	Charles	Mark	r	4
15	Charles		o	
16	Charles		o	
17	Charles	Jim	r	1

[24] *retweetRatio* is useful to inspect data, but was not used in the machine learning process described later, since it is merely the ratio of two other variables.

[25] This table shows a simplified version of data that might be expected on a tweet, for example tweet IDs in reality are 18 digit numerical integers for example 850006245121695744.

Table 5.4: Example tweet expected results.

Account	oCount	rDistinctCount	rCount	retweetRatio
Jim	1	1	2	2
Mark	3	2 (3 and 4)	3 (3 twice and 4 once)	1.5
Dave	2	1	1	1
Alice	1	1	1	1
Bill	1	0	0	0
Charles	2	0	0	0

numbers is not quite as straightforward as it seems. Each tweet has an *account.id_str* field which corresponds to the account which created the tweet. The tweet may or may not be a retweet. If it is it will have a *retweeted_status* structure, and within this there is a field *retweeted_status.account.id_str* which corresponds to the account which created the original tweet that was retweeted.

In order to count the number of tweets an account created, the SQL in Listing 5.2 was used:

```sql
select account.id_str, account.screen_name,
sum(case when retweeted_status  is null then 1 else 0 end) as oCount,
sum(case when retweeted_status  is not null then 1 else 0 end) as rCount
from predictions_d where prediction=1 and utc_offset is not null
group by account.id_str;
```

List of Code 5.2: SQL used to count the number of tweets an account created.

This SQL aggregates by account, and counts original tweets, *oCount*, by testing whether *retweeted_status* is null. It also counts the accounts retweets, *rCount*, by testing whether *retweeted_status* is NOT null. However this code's calculation of *rCount* is the number of tweets the account has retweeted, NOT the number of times their original tweets have been retweeted, which is what is of interest. In order to count the number of retweets of the accounts' original tweets it is necessary to group by *retweeted_status.account.id_str* rather than *account.id_str*. To calculate the retweets count correctly the SQL in Listing

5.3 can be used:

```
select retweeted_status.account.id_str, retweeted_status.account.screen_name,
sum(1) as rCount,
sum(1) / count(distinct retweeted_status.id_str) as retweetRatio
from predictions_d where prediction=1 and utc_offset is not null and
   retweeted_status  is not null
group by retweeted_status.account.id_str, retweeted_status.account.screen_name;
```

List of Code 5.3: SQL used to count the number of retweets of an account's original tweets.

This SQL only looks at tweets that have a *retweeted_status*, and counts these as *rCount*. This is now the correct retweet count, since this SQL is grouping by *retweeted_status.account.id_str*, which is the account that created the original tweet, which this tweet has retweeted. This SQL also calculates *retweetRatio* which is *rCount* divided by the count of distinct tweet IDs that were retweeted for the particular account. The results from the two previous SQL statements could be combined for example by pasting into Excel, however practically an automated system should not require this level of interaction on the account's part. To avoid this the two SQL routines can be combined into one, as shown in Listing A.6.

5.14.1 Machine Learning and Accounts

. Twitter does not give details on how it flags problem accounts, merely stating that it uses behavioural analysis rather than just keyword searching. Without access to Twitter's software, this research created a system that uses machine learning techniques to spot racist accounts. While Twitter's system may or may not be efficient at doing a similar task, it is still useful to create a tool that law enforcement can use independently of Twitter, whose requirements for such a tool are likely to differ from those of law enforcement. To create such a tool the retweet data, already discussed in this chapter, was used in a machine learning analysis of the accounts. The tool created predicts whether an account is racist or not, using just the data available from tweets, without any other API calls to Twitter, which, as already noted, can be time-consuming and expensive.

To perform a machine learning analysis of the accounts, the following procedure was used. First, data for the analysis was taken from the retweet data from the accounts with 15 or more original tweets (there were 1,445 such accounts). From these 1,445 accounts 250 accounts were chosen at random and were determined to either be a racist account or not. A second rater performed the same analysis, and the interrater agreement was 0.87, with a third rater providing a deciding vote in the case of any disagreement. For each account the following variables were tabulated: *oCount*, *rCount*, *rDistinctCount*, *followers_count*, and *Vocabulary*, where *oCount*, *rCount* and *rDistinctCount* are as described above in this chapter, and Vocabulary is the set of distinct lemmatised words in an account's original tweets from the retweet dataset. These variables were used as input features in an SVM model. The model was trained on the data split 70:30 training to test data and run with tenfold cross validation.

A similar procedure was performed, with the results from the grounded theory analysis. The set of focused coding codes were used as categorical input data. These data in the form of sparse vector, where each tweet's vector had element equal to 1 if the code applied to it, and zero otherwise. The model was again trained on the data split 70:30 training to test data and run with tenfold cross validation.

Finally both sets of features were combined, the retweeted data and the grounded theory data both used as features in a machine learning analysis of the accounts.

The results of these analyses of the Accounts data are given in Section 7.4.

5.15 Chapter Summary

This chapter explores the use of mixed methods, that applying a combination of quantitative (ML) and qualitative (grounded theory) methods to answer the research questions posed. First a theoretical discussion of mixed methods is given. Then the use of machine learning is explored, including an explanation of the difference between supervised and unsupervised learning. Then the application of machine learning to textual data is discussed, including a section on how the data is preprocessed prior to input into an ML algorithm, along with the techniques employed to evaluate the output of these classifiers. Finally the qualitative analysis of the tweets is discussed, including various techniques, the main one being that of grounded theory.

This research uses both quantitative and qualitative methods, the qualitative methods chiefly being grounded theory, which is very much situated within the constructivist paradigm, and the quantitative method of machine learning which is within the positivist paradigm (Ponterotto, 2005). The solution to the use of methods with potentially conflicting theoretical bases, is to take a *mixed methods* approach, with such an approach following the paradigm of *pragmatism* (Feilzer, 2010). To do this machine learning routines were written with the viewpoint that it is possible, albeit imperfectly, to identify what ‘racist tweets’ and who ‘racist tweeters’ are. This postpositivist work was combined in a mixed methods approach, with an analysis, using grounded theory, of the identified tweets, and the concepts which arise from this analysis, were used to identify racist tweeters. The grounded theory portion of the work, is very much situated in constructivism, yet it also fits within the pragmatic paradigm. The racist tweets created by the postpositivist computer system research were analysed using constructivist grounded theory, and the results of this were input into more postpositivist computer programs. One of the main drivers behind the choice of method in this research is the type of data being analysed. Williams and Burnap (2015) state that criminology is in the infancy of using big data, and indeed there are very few publications that discuss the intersection of big data and criminological theory. The work that has been done in utilising big data sets for criminological research is unconvincing since often the amount of data analysed is relatively small, and therefore not big data, or the analytical techniques do not fully utilise the capabilities of Hadoop or other processing frameworks. Despite the lack of big data and criminology research, it is argued that so called 'digital sociology' ('the sociology of online networks, communities, and social media' (Murthy and Bowman, 2014, p.2)), has emerged as an important field in the area of quantitative sociology with a concomitant rise in the study of large datasets available from social networking sites. Despite the rise of this new field of sociological study, Chan and Bennett Moses (2016) argue that big data is seen by some as a threat to criminology , although they argue it is not, except for 'the use of machine learning procedures in predictive analysis is one area where establish ways of doing criminology may well be threatened' (ibid., p.34). Their concerns are that with big data causal relationships are no longer necessary, and that correlations are sufficient. They argue that it is possible to just continue to add more and more features as input to a machine learning algorithm and sooner or later it will be an excellent predictor. There is no need for theory, just an observable successful predictive result. This logic seems to ignore the process of feature selection, which is certainly guided by theory, and is often regarded as the most important part of machine learning (Kaushik, 2016). While it is true that if you include enough variables you will get a very good predictor for your data, chances are that it will be overfitting the data and will perform poorly in other situations. The Twitter data collected in this research is an example of big data, and the methods used to analyse this data are techniques associated with the handling of such data.

Big data systems need to be highly scalable so that further increases in volume, variety or velocity of data can be handled efficiently. Big data systems normally allow for scaling by being designed to allow parallelisation of their tasks. Hadoop is the most prevalent distributed parallel computing paradigm. Hadoop handles the parallelisation of any processing necessary and it also handles the management of any resources required in the processing. It also reduces the problems created by unreliable hardware and software, since Hadoop makes it very easy to add and remove resources. HDFS is a component of Hadoop that allows files to be stored in distributed storage and MapReduce is the original computational framework in Hadoop. YARN was introduced to replace MapReduce. YARN allows users of Hadoop to utilise different processing frameworks against their HDFS data, rather than just MapReduce (Vavilapalli et al., 2013). One such processing framework is that of Spark, which can handle loading and querying of data and machine learning processing. Spark allows data to persist in memory and this in-memory data sharing means Spark's performance often outweighs that of MapReduce as much as 100 times. Spark is also more suited to machine learning processing than MapReduce, since machine learning usually requires an iterative algorithm requiring a number of steps. For MapReduce to handle such an algorithm it needs to write data to HDFS, and this writing to disk takes up much of the time of the processing for a job. In order to handle machine learning processing Spark contains two libraries: `mllib` and `ML`.

Spark/YARN/Hadoop is the framework that was utilised for this research to store and manipulate large Twitter datasets.

The 140 character messages that tweets contain are only a small part of their structure. Tweets are in fact complex JavaScript Object Notation (JSON) structures.

Twitter data can be obtained through three APIs: the *Streaming API*, the *Sample API*, and the *Firehose API*. The Firehose API gives access to all of the tweets on Twitter, but accessing this is expensive. As a result most researchers use one of the other two APIs, both of which provide 1% of the output of the Firehose. The main advantage of the streaming API is that it allows users to filter their sample, by the use of keywords. It has the disadvantage that its results are not a random For this research it was necessary to select data by keyword and so the Streaming API was used.

For this research tweets were collected over two time periods: 79 days in Summer 2016, and 34 days in Summer 2017. The two datasets were denoted D1 and D2 respectively. The D1 dataset was used in both the creation and validation of machine learning predictions of racist tweets, the grounded theory analysis of their content, and the creation of metrics aimed at spotting racist tweeters. The D2 dataset was used as validation in each of these tasks. The approach to collecting the data was to collect tweets based on keywords taken from the Wikipedia List of Ethnic Slurs page. The lack of contextual information in such an approach can make keyword searching of tweets potentially lead to both false-positives and false-negatives (Nobata et al., 2016), however it was decided to take this approach, since this will collect tweets containing a particular type of comment:

those that will be labelled as *personal racism* , although of course racist speech that does not contain these racist keywords will be excluded. One alternative is to take context into account and perhaps look at Ngrams, but this is likely to lead to exclusion of certain tweets and the aim for this research was to include everything at the data collection stage that might be deemed to be racist. The data was collected using a Python program that harvests tweet data from Twitter, and stores it in text files. The D1 dataset consisted of 83,994,885 tweets. These 84 million tweets were reduced to 41,260,026 when those without values for *utc_offset*. The D2 dataset consisted of 28,766,811 tweets. These 28 million tweets were reduced to 13,870,269 when those without values for *utc_offset* were excluded.

These data was stored on four machines, which were connected in a cluster running a Hortonworks Data Platform (HDP) Hadoop distribution in *fully distributed mode* under Ubuntu. In this cluster D1 and D2 were imported to Hive and stored as Optimised Row Columnar (ORC) tables. 84,000 Tweets were randomly extracted from the D1 files and given to annotaters in batches of 1,000 tweets. The annotators followed rules to determine if the tweets were racist or not, adapted from Waseem and Hovy (2016, p.89).

To determine what the interrater reliability was at the beginning, middle and end of the annotation process, the same three datasets were given to each of the annotators. 100 tweet samples from nigga.txt, nigger.txt and wop.txt were given to each of the annotators, at the beginning, middle and end of the process. Krippendorff's α was calculated for each of the three datasets, giving 0.60, 0.92, and 0.90 for the nigga.txt, nigger.txt and wop.txt annotations respectively.

The data research exhibited *class imbalance*: there were many more nonracist tweets than racist ones. Simple random over sampling was used to correct this.

Supervised machine learning was used to classify the unlabelled tweets. Supervised machine learning involves a classifier being given a set of training data which is usually hand annotated using expert knowledge of a domain. The classifier then takes this training data and produces a predictive model that can calculate probabilities of text belonging to either class based on the features of the text. To determine whether the classifier is a good one, it is given a test dataset, which is usually a subset of the original training data, that is extracted before the training data is given to the classifier to produce the model. When the model is run on the test data the output that is produced can be compared with the actual hand coded values, and a series of metrics can be produced. For this research this consisted of using 84,000 randomly selected tweets, which were labelled as racist or not, and used as the training and test datasets input to machine learning algorithms in order to predict whether each of a big data scale amount of tweets were racist or not.

For supervised binary (two possible classes) classification problem such as this, there are four possible outcomes:

- True Positive (TP) - data is correctly classified as positive.
- True Negative (TN) - data is correctly classified as negative.
- False Positive (FP) - data is incorrectly classified as positive.
- False Negative (FN) - data is incorrectly classified as negative.

For this research positive will mean ‘racist’, that is positive is being used to mean category of interest has been found for this piece of data. From these four outcomes, a number of metrics can be derived, and accuracy, AUROC, AUPRC and F-score for $\beta = 1$ and $\beta = 0.5$ were used to compare the machine learning algorithms. Accuracy is the percentage of predictions that are correct. AUROC is the area under the graph of true positive rate versus false positive rate. The other metrics utilise precision and recall. Precision is the ratio of true positives to all those classified as positives, whether they are correct or not. Recall is the ratio of correctly classified positives divided by the total number that are actually positive. AUPRC is the area under the curve of precision against recall. F-score for β is the harmonic mean of precision and recall, where the importance of recall is β times the importance of precision. While these metrics are useful evaluators of the algorithms, they do not evaluate a classifier with respect to how well it performs with unseen data. Cross validation is an evaluation method that overcomes this issue. *K-fold cross-validation* splits the data into K subsets, the model is run K times, each of the K subsets is used as a test set, and the other K-1 subsets are combined as the training data. The data for this research was split in the ratio 70:30, that is 70% training and 30% test, and tenfold cross validation was used.

Goffman (1981) argues that 'talk' is made up of: *ritualization*, *participation framework*, and *embedding*. For Murthy (2012) there is ritualization in the symbols and text used by Twitter users. Embedding is certainly a feature of Twitter, due to the mechanism of retweeting. The complexities of Twitter's participation framework can be analysed using Habermas's (Habermas, 1989; Habermas et al., 1974) notion of the 'public sphere', which is the fora that make up public opinion, open to all without restriction. The notion of Twitter as a public sphere can be seen in the work of Lotan et al. (2011), who examined tweets relating to the both the Tunisian and Egyptian uprisings in 2011. However Fuchs (2017) argues that Twitter is not a public sphere as Habermas imagined one. Habermas used Marxist theory to determine that a correctly functioning public sphere would have amongst its key priorities the aim of equality, but Fuchs (ibid.) argues that Twitter has no such mission, being a private profit making organisation and that its content may provoke social change and be on the side of, as he sees it, virtuous left realism, but on the other hand also contains right-wing extremism and other socially destructive ideas.

The data for the qualitative analysis needed to be small enough for human analysis, and hence a sample was used. The data used in the qualitative analysis was the tweets that were flagged as racist i.e. they are in the dataset Pred R. From these tweets, it was decided that only those tweets that were influential would be analysed. To determine tweets that were influential, only those that had been retweeted at least once were included in the qualitative analysis of the tweet data. NLTK was used to produce the Top 25 most common: words excluding stopwords, hashtags, bigrams and trigrams from this dataset. Nvivo, which was used to generate *word clouds* and *word trees.*

Tweet examples were analysed discursively, to give an idea of the kinds of data making up the dataset. Grounded theory was then used to provide a deeper theoretical understanding of a 2,000 tweet sample. The aim was to develop a grounded theory of the content of the tweets predicted as racist. Concepts were sought from the text. From the concepts categories were determined. These categories were themselves grouped into a few main categories. The analysis was continued until 'theoretical saturation' was reached.

The accounts linked to the predicted racist tweets were too numerous (there were 105,211 such accounts) to be handled without some form of automated identification. The aim was to identify the racist accounts that have the most influence over the network. In order to find measures suitable to identify such accounts, the standard metrics of influence were unsuitable since these measures identify relatively prolific accounts, whereas the ones of interest for this research were not prolific. It was also desirable to use measures that avoided API calls to Twitter, which can be time-consuming and expensive. Three measures were used: the total number of original tweets from an account, *oCount*, the number of distinct tweets that were retweeted from an account, *rDistinctCount*, and the total number of retweets of an account's tweets.

Twitter does not give details on how it flags problem accounts, merely stating that it uses behavioural analysis rather than just keyword searching. It is useful to create a tool that law enforcement can use independently of Twitter, whose requirements for such a tool are likely to differ from those of law enforcement. To create such a tool the retweet data was used in a machine learning analysis of the accounts. To perform a machine learning analysis of the accounts, the following procedure was used. First, data for the analysis was taken from the retweet data from the accounts with 15 or more original tweets (there were 1,445 such accounts). From these 1,445 accounts 250 accounts were chosen at random and were determined to either be a racist account or not. A second rater performed the same analysis, and the interrater agreement was 0.87, with a third rater providing a deciding vote in the case of any disagreement. For each account the following variables were tabulated: *oCount*, *rCount*, *rDistinctCount*, *followers_count*, and *Vocabulary*, where *oCount*, *rCount* and *rDistinctCount* are as described above in this chapter, and Vocabulary is the set of distinct lemmatised words in an account's original tweets from the retweet dataset. These variables were used as input features in an SVM model. The model was trained on the data split 70:30 training to test data and run with tenfold cross validation. A similar procedure was performed, with the results from the grounded theory analysis. The set of focused coding codes were used as categorical input data. These data in the form of sparse vector, where each tweet's vector had element equal to 1 if the code applied to it, and zero otherwise. The model was again trained on the data split 70:30 training to test data and run with tenfold cross validation. Finally both sets of features were combined, the retweeted data and the grounded theory data both used as features in a machine learning analysis of the accounts.

The next chapter provides a discussion of the relationship between big data and criminology and gives further details on the theory of machine learning.

Chapter 6

Big Data and Machine Learning

This chapter provides a discussion of the relationship between big data and criminology and gives further details on the theory of machine learning. There are sections on *features*, that is the dimensions of the data used as input to prediction algorithms. Text, user, geographical and temporal features are discussed. Then Spark extraction and transformation methods are explored. This is followed by section on the different algorithms utilised in this research, along with the techniques employed to evaluate the output of these classifiers.

6.1 Use of Big Data in Criminology

Williams and Burnap (2015) state that criminology is in the infancy of using big data, and indeed there are very few publications that discuss the intersection of big data and criminological theory. While there are a number of papers regarding big data and investigation or crime reduction (for example: Pramanik et al. (2016), Goswami et al. (2016)), there is a paucity of big data and criminology work, because using Hadoop[1] is not easy (Woodie, 2017a; Woodie, 2017b; Agarwal, 2017) and the work that has been done in util-

[1] Hadoop is the most prevalent distributed computing paradigm that is used to handle big data. For further discussion of Hadoop see Section 5.3.

ising big data sets for criminological research is unconvincing. Some of it can be criticised for not really using big data. For example, Williams and Burnap (2015) collected tweets containing the hashtag #Woolwich for one month after the murder of Lee Rigby in the terrorist attack in Woolwich, London in 2013. They collected 427,330 tweets. If it is assumed that each tweet is 4K in size these tweets would be 427,330 *4K = 1,709,320K = 1.7GB, not a large data set by contemporary standards and it is unsure whether their description of their work as big data is accurate.

Another example is that of O'Brien et al. (2015) who analysed Boston's constituent relationship management (CRM) system to develop measures of 'broken windows' i.e. indicators of social decay. They framed their work as big data research. Their data were 365,729 requests for service, again a relatively small amount of data that does not require large scale processing. They do not discuss how their work and data should be classified as big data.

Another criticism of the work utilising big data in criminology, is that the analytical techniques do not fully utilise the capabilities of Hadoop or other processing frameworks. For example, Jain and Bhatnagar (2016) plotted Indian crime rates using Hadoop and the data flow language, Pig. However they were running Hadoop in pseudo-distributed mode, that is on a single machine, negating any benefits Hadoop has in big data processing.[2]

Despite the lack of big data and criminology research, it is argued that so called 'digital sociology' ('the sociology of online networks, communities, and social media' (Murthy and Bowman, 2014, p.2)), has emerged as an important field in the area of quantitative sociology with a concomitant rise in the study of large datasets available from social networking sites. Despite the rise of this new field of sociological study, Chan and Bennett Moses (2016) argue that many see big data as a challenge to criminology. They discuss whether the veracity of claims made by certain theorists, that big data is challenging the use of theory in sociology, and that the possibility to include all data within an analysis, due to computing systems able to handle very large amounts of data, means traditional forms of analysis such as sampling and correlation are no longer necessary, indeed they

[2] See Section 5.3 for a discussion of this.

are obsolete. They contend that big data is not a threat to criminology except for 'the use of machine learning procedures in predictive analysis is one area where establish ways of doing criminology may well be threatened' (ibid., p.34). Their concerns are that with big data causal relationships are no longer necessary, and that correlations are sufficient. They argue that it is possible to just continue to add more and more features as input to a machine learning algorithm and sooner or later it will be an excellent predictor. There is no need for theory, just an observable successful predictive result. This logic seems to ignore the process of feature selection, which is certainly guided by theory, and is often regarded as the most important part of machine learning (Kaushik, 2016). While it is true that if you include enough variables you will get a very good predictor for your data, chances are that it will be overfitting[3] the data and will perform poorly in other situations.

At the same time as the rise in digital sociology, there has been increases in computing data storage and processing speeds, yet there are still challenges to handling big data scale data. Large-scale big data solutions are expensive and are usually not viable for individual researchers (Murthy and Bowman, 2014). Murthy and Bowman (ibid., p.6) examined the use of a 'small-scale' big data system noting that,

> most information available about the design and implementation of distributed storage and retrieval systems focuses on large, multi-node systems, which are likely overkill for most academic social research case scenarios.

Their aim was to create a Hadoop system costing no more than \$5,000 (or £3,000 at their time of writing). They noted the cost effectiveness of such a system, since comparable systems using traditional RDBMS databases, instead of Hadoop, could cost \$50,000, at their time of writing. They noted the ability of Hadoop to process large datasets using several nodes, yet surprisingly created a system with only a single node. For software they used Cloudera's version of Hadoop, claiming that this implementation of Hadoop is 'the most relevant to social data projects' (p.5), although they give no support for this claim. Their aim was to store one year's worth of tweets, from the 1% publicly available sample

[3] Overfitting occurs when a machine learning algorithm works very well for a dataset due to the algorithm using a large number of the dataset's features as input. The model created is based on a large amount of information for that dataset, so it fits it well, but does not generalise to other data.

stream, since they felt their system would not be able to handle the Twitter Firehose.[4] To ingest the data they used Apache Flume which allows connection to a number of data sources and pushing of the retrieved data into a number of data sinks.[5] Their data was stored in HDFS and they used Hive to query it, since it performs very well in data loading and range queries.[6]

This brief review illustrates the sparsity and paucity of criminology and big data research. As mentioned utilising big data successfully is not easy, and there follows a discussion of what is meant by big data and how Hadoop handles it, showing some of the technical issues that need to be overcome to successfully apply big data techniques.

6.2 This Research and Big Data

The Twitter data collected in this research is an example of big data, and the methods used to analyse this data are techniques associated with the handling of such data. Now what is meant by 'big data' and how it can be handled methodologically will be considered.

As mentioned in Section 1, the aim of this research, is to automatically identify racist tweets and tweeters by a combination of machine learning and other computer-based techniques. At the heart of these techniques is the manipulation of data, which requires computer processing and storage capabilities. How the data for this research was processed and stored was a key consideration for the research methodology. The data for this research was data collected from Twitter, using keyword searching of the public streaming API which provides a non-random sample of 1% of tweets, which Twitter provides free of charge (see Section 5.6 for details). Even with this small sample percentage a significant amount of data (hundreds of gigabytes) was collected. Since this work could be applied to the entire Twitter dataset[7], it was sensible to create a system that potentially

[4] Twitter's Firehose is an API which emits all tweets. See Section 5.6 for a discussion of Twitter's data retrieval APIs.

[5] For this PhD research Flume was not used, since at the time of data collection the Hadoop system was not ready to store the data.

[6] See Section 5.3 for a discussion of this.

[7] See 8.7 for a discussion on this.

could handle the very large amounts of data Twitter generates on a daily basis. As a result it was decided to take a *big data* approach to this issue, as such an approach has, as one of its underlying tenets, the ability to scale in order to handle large amounts of data. There are three typical aspects of handling this kind of data that are potentially difficult without any big data processing ability. These are the size of the data, the speed with which the data arrives and the ability to perform intensive processing on the data. The size of the data is not really an issue for this research, since all the data can be stored on a single 5 TB drive. Even so, processing this amount of data on a single machine can be problematic, so the ability to spread the data across multiple machines that big data techniques allow is beneficial. The speed with which the data arrives is also not problematic for this research since the data is stored and processed at a later date. This, however, would become an issue if the system was changed to handle live, streaming data [8], and then the various Hadoop[9] software components that allow the easy ingestion of such data, would be beneficial. The main requirement for this research which makes a big data and Hadoop approach vital, is the requirement to perform intensive processing on the data. With a single 'standard' machine even straightforward processing on big data amounts of data is difficult. For example editing of text files becomes problematic when their size exceeds 2 GB, and doing simple transformations, for example the removal of malformed records from the data that was performed for this research, can take several hours on a single machine. For this reason alone a big data approach was a necessity.

Big data ideas and techniques arose from the work of large IT corporations: Google, Yahoo, Facebook and others, and their need to process large amounts of data. The increase in the amount of data generation over the last decade or so has been phenomenal, and it is not unusual to see companies dealing with petabytes of data or more. WPformers.com (2017) estimates that Google handles around 10 to 15 exabytes of data, although they incorrectly use a 1,024 multiple. To gain some understanding of why this is the case, a brief discussion of how computers store data will now be given.

[8] See 8.7 for a discussion on this.

[9] See Section 5.3 for a discussion of Hadoop.

6.2.1 Computer Data Storage

Computers logically store data as binary digits or *bits*, that is digits that can take the values zero or one only. Traditionally bits are grouped into multiples of eight, which are known as *bytes*. Also traditionally, there has been confusion regarding the terminology relating to multiples of bytes. For example, a *kilobyte* has referred to both 1,000 and 1,024 (since traditionally computer storage terminology referred to mutliples of 2^{10}, and 1,024 is 2^{10}). Table 6.1 shows some of this terminology and the difference between the Système International d'Unités (SI) and International Electrotechnical Commission (IEC) systems for referring to computer storage. As can be seen in the table, the SI units begin with the standard SI prefixes, kilo-, mega-, giga-, tera-, peta-, exa-, zetta- and yotta- denoting 10^3, 10^6, 10^9, 10^{12}, 10^{15}, 10^{18}, 10^{21} and 10^{24} bytes respectively. In contrast the IEC units prefixes are: kibi-, mebi-, gibi-, tebi-, pebi-, exbi-, zebi- and yobi- and they denote 2^{10}, 2^{20}, 2^{30}, 2^{40}, 2^{50}, 2^{60}, 2^{70}, and 2^{80} bytes respectively. The table also contains another column, 'number of TB/TiB' which shows the number of terabytes or tebibytes - depending on whether SI or IEC units are being used - that each byte size term is equivalent to. It can be helpful to think of this as the equivalent *laptops* if it is assumed a laptop is equivalent to 1 TB. Then, if Google is handling 10 to 15 EB of data, then storing that amount of data would require the equivalent of 10 to 15 million laptops. Table 6.1 also has a 'Difference between IEC and SI in TB' column, which calculates the difference between each of the SI and IEC equivalents in terms of terabytes (i.e. by dividing by one thousand million). This shows that at the exabyte/exbibyte level the difference is equivalent to over 150,000 laptops, i.e. a very considerable difference indeed.

With regards to this research, the data can be housed on a single 5 TB drive. By the end of the research, including datasets D1 and D2, and various transformations (for example removing 'bad records' from the original data) and backups the 'data' folder on this drive was 2.91 TB in size. This was only part of the storage required for this research, as there were many other programs, spreadsheets, Word documents and so on, but even this folder alone could not be stored by a standard laptop in 2018, since most laptop hard drives are around 1 TB in size. So while the size of the datasets D1 and D2 might not be considered as 'big data size' datasets, their handling certainly requires additional

Table 6.1: Computer storage terminology for both IEC and SI nomenclatures.

	Term	**Power**	**Number of Bytes as decimals**	**Number of TB/TiB**	**Difference between IEC and SI in TB**
SI	Kilobyte (KB)	10^3	1,000	0.000000001	
	Megabyte (MB)	10^6	1,000,000	0.000001	
	Gigabyte(GB)	10^9	1,000,000,000	0.001	
	Terabyte(TB)	10^{12}	1,000,000,000,000	1	
	Petabyte (PB)	10^{15}	1,000,000,000,000,000	1,000	
	Exabyte (EB)	10^{18}	1,000,000,000,000,000,000	1,000,000	
	Zettabyte (ZB)	10^{21}	1,000,000,000,000,000,000,000	1,000,000,000	
	Yottabyte (YB)	10^{24}	1,000,000,000,000,000,000,000,000	1,000,000,000,000	
IEC	Kibibyte (KiB)	2^{10}	1,024	0.0000000009	0.000000000024
	Mebibyte (MiB)	2^{20}	1,048,576	0.000001	0.000000049
	Gibibyte(GiB)	2^{30}	1,073,741,824	0.001	0.000074
	Tebibyte(TiB)	2^{40}	1,099,511,627,776	1	0.0995
	Pebibyte (PiB)	2^{50}	1,125,899,906,842,620	1,024	126
	Exbibyte (EiB)	2^{60}	1,152,921,504,606,840,000	1,048,576	152,922
	Zebibyte (ZiB)	2^{70}	1,180,591,620,717,410,000,000	1,073,741,824	180,591,621
	Yobibyte (YiB)	2^{80}	1,208,925,819,614,620,000,000,000	1,099,511,627,776	208,925,819,615

capabilities than a single laptop can provide.

Informally these kinds of extremely large levels of data storage, and their concomitant processing, are what is usually referred to as ‘big data’. The following section discusses more formal definitions of big data.

6.3 Definitions of Big Data

Big data is often described in terms of a number of V’s, originally these being three V’s: volume, variety and velocity.[10] Volume refers to the amount of data, velocity is the rate at which data is created or processed, and variety refers to the fact that data can come from disparate sources with different formats and types of data. This type of definition of big data is a feature-based definition, focusing on technical issues such as how data is stored or processed. Other feature-based definitions such as that of Kitchin (2014) go beyond the three V’s (Chan and Bennett Moses, 2016). Kitchin (2014) notes that there is a wide range of big data definitions, from what he calls the ‘trite’ definition of a dataset that is too big to be handled by a single machine, to ‘sophisticated ontological assessments that tease out inherent characteristics’ and the examples he gives are of Boyd and Crawford (2012) and Mayer-Schönberger and Cukier (2013). His survey of the literature shows common features of big data to be (italics are the authors):

> *exhaustive* in scope, striving to capture entire populations or systems (n = all), or at least much larger sample sizes than would be employed in traditional, small data studies; fine-grained in *resolution*, aiming to be as detailed as possible, and uniquely *indexical* in identification; *relational* in nature, containing common fields that enable the conjoining of different datasets; *flexible*, holding the traits of extensionality (can add new fields easily) and *scalable* (can expand in size rapidly) (Kitchin, 2014, p68).

[10] There are many attempts to add V’s and the most V’s that were found by the author is seven: volume, velocity, variety, variability, veracity, visualization, and value (DeVan, 2016).

Chan and Bennett Moses (2016) argue that feature-based definitions, such as these, are problematic since they relate to the current state of computing, and so are therefore constantly changing, since innovations and progressions in processing capability and storage mean that what currently is difficult to process or handle using 'traditional' techniques may not be so in the near future. Chan and Bennett Moses argue that big data can be seen as having other dimensions beyond that of technology. For Chan and Bennett Moses (ibid.) and Boyd and Crawford (2012) big data is more than merely a technological construct, it exists in the interaction between technology, culture and scholarship, that is the reference to big data often includes an interaction between three things: the technological aspect of handling large datasets, the analysis of these datasets and the mythology that suggests that big data provides an objective 'truth', that it is better than traditional techniques. While it might be the case that there is a mythology around big data, it does not seem useful to include this within a definition of big data. The idea of Chan and Bennett Moses that the term 'big data' has become inextricably linked with this mythology, is debatable, and even if it is true, it is much more fruitful to narrow down what is meant by big data, instead of expanding it to include subjective views of its efficacy. For this reason this research employs a definition[11] of big data that is purely technologically based.

In discussing big data, Mansour (2016, p.1) notes other uses of the phrase which muddy the waters, when he notes that,

> increasingly it is used to refer to social networking websites and the enormous quantities of personal information, posts, and networking activities contained therein.

This conflation of big data with social network data seems to be fairly rife in the literature. For example in their study Williams and Burnap (2015), the authors collected approximately half a million tweets, which they termed 'big data'. However they did not discuss what they understand by big data, nor do they explain what makes this relatively small amount of data a 'sample of "big data"' (ibid., p.225). They note their use of ma-

[11] This definition is given at the end of this section.

chine learning classifiers and Java WEKA machine learning libraries, but do not give any methodological details regarding scalability.

None of the definitions so far have made much of the fact that big data requires a considerable amount of *processing power*. This research concerns performing computationally intensive processing on a large dataset, and this is a requirement for much big data work. Although the most important V of big data is often thought to be the volume of the data, in some instances this is less important than the processing required to handle the data. When the machine learning routines are run on a large dataset, *D1*[12], of tweets collected for this research, the D1 dataset is split between machines by Hadoop[13], and each machine performs processing on its portion of the dataset (further details of this are given later in this chapter). The storage (i.e. its volume) of the D1 dataset could reasonably be handled by a single machine, but the *processing of it* could not. Because of this, definitions that give some focus to processing are now discussed.

Ward and Barker (2013, p.2) review various definitions of big data and, drawing on these, formulate their own:

> Big data is a term describing the storage and analysis of large and/or complex data sets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning.

While this definition does address processing, it has become rather dated since MapReduce is now just one technology that runs on YARN a more generalised parallel framework, and MapReduce is often seen as being superseded by Spark another technology that runs on YARN.

Another definition was provided by De Mauro et al. (2015, p.103), who looked at the frequency of keywords related to big data research. They provide a consensus definition:

[12] For details see Section 5.7.

[13] For further discussion of Hadoop see Section 5.3.

> Big data represents the information assets characterised by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value.

This definition also addresses processing with its reference to 'analytical methods', but is problematic in that it suggests the data must be big and varied, thus excluding huge homogeneous datasets as being defined as big data. It also excludes relatively low volume datasets that require lots of computing power for analysis. Both these definitions are consensus definitions. Consensus definitions, such as theirs, have their appeal but they assume the consensus is correct or useful. Simply because a researcher deems their work to be big data it is not necessarily so, it may be labelled as such because of the cachet of big data research.

So the definitional landscape of big data is indeed complex, however since the underlying theoretical framework of this research is pragmatism, a definition of based on practical aspects of big data is appropriate. A definition needs to highlight the nature of both the data and its processing. Such a definition, that reflects how big data is different to any other analysis with respect to processing and size of data, can succinctly be expressed as:

> big data is data that requires parallel processing either because of its size or analysis requirements.

6.4 Machine Learning

The main aim of this research is to provide an automated racist detection system for Twitter. To do this machine learning techniques were used. Machine learning is a process which uses data to inform a computer system, which is then able to predict something based on this data. There are many ways to achieve any particular machine learning task. The task here is one of text handling, so any machine learning system must include some form of preprocessing of the text. The preprocessing of textual data is discussed in Section 6.4.2. This preprocessing includes steps such as tokenisation of the data. Then the machine learning system requires features to be chosen, that is a subset of variables that provide a good predictive output need to be determined. Machine learning also requires the choice of an algorithm, which determines how the machine learning generates predictions based on input data. The choice of models is discussed in Section 6.4.9.

This processing is very intensive and so the distributed processing framework Spark was used. Spark provides methods to both extract and transform data and these are listed in Sections 6.4.7 and 6.4.8. The process of choosing how to preprocess the data, which features to use, which model to select, and which spark extraction and transformation methods to use is an iterative process, and requires a large number of tests and retests to be made, whilst examining output evaluation metrics for the different combinations of these factors and these processes are described in the following sections.

This section discusses the use of machine learning approaches[14] and their evaluation using various metrics. Models were created using seven different algorithms: NB, LR, SVM, RF, DT, GBTs and ANN. These models were used to build 'racism detection models' on a dataset of tweets and a comparison of models is given in Section 7.1.4 along with a description of the Twitter-sourced dataset in Section 5.9. The models were compared against one another using standard metrics to determine the best model for determining racist tweets. They were also used to evaluate the efficacy of using tweet metadata as features in the predictive models. Limitations of the techniques used for racism determination in tweets are presented.

Machine Learning is enabling machines to learn without programming them explicitly (Samuel, 1959) usually in order to facilitate 'the automated detection of meaningful patterns in data' (Shalev-Shwartz and Ben-David, 2014, p.xv). Machine learning is the field of study concerned with making machines learn and detect patterns experientially. The aim is for a machine learning algorithm to learn from data and adapt its actions to make more accurate predictions as a result (Marsland, 2015). Machine learning allows for

[14] The machine learning algorithms were run consecutively within one Zeppelin paragraph or using the spark-shell CLI (see figure A.1 for an example program).

the analysis of datasets that are too large for humans to analyse feasibly. It enables the detection of patterns invisible to humans and aids in the reduction of subjectivity in the analysis of data (Pentreath, 2015, p.39). Machine learning involves inductive inference, that is examples related to a phenomenon are extracted from data and these are used as input to an algorithm which aids in inferring a general model.

For machine learning the data is normally in the form of attribute vectors which map to known classes,[15] and an algorithm is chosen that can map these vectors to classes, and which is also able to do this for data that is as yet unseen (Podgorelec et al., 2002). The attribute data along with its mapped classes is usually then split into *training* and *test* datasets. The training dataset is used by a machine learning algorithm to inductively generate a model which maps attributes to classes, and the test dataset is used to generate metrics which can be used to evaluate the efficacy of the model (ibid.).

More formally Japkowicz and Shah (2011, p.24) describe a general model of learning as having three components:

> 1. an instance space X from which random vectors $\mathbf{x} \in \mathbb{R}^n$ can be drawn independently according to some fixed but unknown distribution,
> 2. a label $y \in Y$ for every vector $\mathbf{x}$ according to some fixed but unknown conditional distribution. In the more general setting, y need not be scalar but can annotate the example $\mathbf{x}$ with a set of values in the form of a vector $\mathbf{y}$, and
> 3. a learning algorithm A that can implement a set of functions f from some function class F over the instance space.
>
> Given these three components, the problem of learning is that of choosing the best classifier from the given set of functions that can most closely approximate the labels of the vectors. This classifier is generally selected based on a training set S of m training examples drawn according to X with their respective labels. Each tuple of a vector $\mathbf{x}$ and its label y can be represented by $\mathbf{z} = (\mathbf{x}, y)$ which

[15] For classification problems machine learning requires an input dataset that maps attribute data to classes that have been determined by some process (often human annotating).

> can be assumed to be drawn independently from a joint distribution D.
>
> [...]
>
> When the learning algorithm has access to the labels y for each example $\mathbf{x}$ in the training set, the learning is referred to as *supervised learning*; the term *unsupervised learning* is used otherwise.

This research uses a supervised classifier machine learning process.[16]

6.4.1 Machine Learning and Text

This research follows the lead of most of the literature, which classifies tweets into two categories, either containing racist speech or not (Malmasi and Zampieri, 2017b). A minority of researchers also use the 'don't know' category, which they then discard.

The classification of tweets as racist or not is an example of Text Categorization that is 'the activity of labelling natural language texts with thematic categories from a predefined set' (Sebastiani, 2002, p.1). Until the late 80s this process was dominated by the knowledge engineering approach which involved the manual creation of a set of rules to determine which category a document would be classified in. Such rules are derived from the knowledge of experts. This paradigm gave way to the machine learning approach and this still holds sway. This approach uses a set of pre-classified documents which a classifier uses as input in order to learn to automatically classify new documents (ibid.).

Text Classification is performed by a classifier, which is an algorithm that takes a set of features of textual data and uses them to predict which class the text belongs to. Classifiers can be chosen from a number of algorithmic families as discussed in section 6.4.9 each having benefits and drawbacks. Supervised learning is used to do this: a classifier is given a set of training data which is usually hand annotated using expert knowledge of a domain. For example, a classifier could be given a set of tweets which have been determined to be racist or not by human intervention. The classifier then takes

[16] For a more in-depth mathematical discussion of classification and machine learning in general, see, for example, Watt et al. (2016).

this training data and produces a predictive model that can calculate probabilities of text belonging to either class based on the features of the text. To determine whether the classifier is a good one, it is given a test dataset, which is usually a subset of the original training data, that is extracted before the training data is given to the classifier to produce the model. When the model is run on the test data the output that is produced can be compared with the actual hand coded values, and a series of metrics can be produced (see Section 5.12.2). For this research this consisted of using 84,000 randomly selected tweets, (as discussed in section 5.8), which were labelled as racist or not, and used as the training and test datasets input to machine learning algorithms in order to predict whether each of a big data scale amount of tweets were racist or not.

Prior to performing the machine learning, the data needed to be preprocessed so that it was in a suitable format for input to the machine learning routine. This is discussed in the next section.

6.4.2 Pre-processing the Data

Analysis of textual data is normally performed using NLP techniques. NLP is the process of extracting the meaning (or semantics) of a text (Kao and Poteet, 2007) usually via the use of automated computer systems (Kumar, 2011). It normally encompasses information retrieval (e.g. tweet text extraction and parsing), text classification (e.g. labelling text tokens as particular parts of speech), text clustering (e.g. storing various forms of the verb 'to go' together) and entity, event and relation extraction (e.g. an entity might be 'member of gang'). To do this NLP utilises known language attributes such as parts of speech and grammatical structures along with lexicons of words, their meanings and their grammatical properties, grammatical rules, lists of things that exist and their relationships ('ontologies of entities and actions') and lists of synonyms and/or abbreviations (Kao and Poteet, 2007) . NLP can consist of a variety of approaches, but often the standard NLP approach contains the following preliminary steps: tokenisation which is the splitting up of a string of text into its constituent parts, that is words and punctuation, which are known as tokens and normalisation which is the process that standardises text, so for example all capitalisation may be removed (Alonso Alemany and Carrascosa, 2011; Bird et al., 2009).

A major challenge for NLP handling of text is the informal and non-standard nature of much User Generated Content (UGC), which can include abbreviations, misspellings and neologisms (a newly coined turn of phrase, an example of which would be 'lol', Gouws et al., 2011b). The 'SMS language' provides a number of challenges to solutions attempting to automate its handling, due to the wide variety of orthography (spelling) and grammar involved in SMS authorship. Often words are spelled phonetically in SMS (e.g. 'shud' instead of 'should'), vowels are removed (e.g. 'txt' instead of 'text'), numbers are used phonetically (e.g. 'gr8' instead of 'great'), acronyms are used in place of words (e.g. 'lol' instead of 'laugh out loud') or combinations of these and other non-standard orthographic forms are used (Kobus et al., 2008). Tweets are very similar to SMS in that they are both short messages and prone to the non-standardisation of UGC (Ritter et al., 2011). A widely used approach to handle such text is that of 'normalization' which attempts to standardise any text discovered (Melero et al., 2012). The standardisation of text might simply ensure that all text is in lower case, but might also include more complex processing such as stemming or lemmatization. Stemming is a process that reduces words to their root or 'stem.' So for example 'grabbing' might be stemmed to 'grab'. Different stemmers have different rules and do not produce standard results (Bird et al., 2009). Lemmatization is the process of reducing words to their base form, similar to stemming (Bird et al., 2009) but often more sophisticated in that lemmatization might replace 'automobile' with 'car' (although this depends on the rules contained in the lemmatising process).

Racist speech detection is related to Sentiment Analysis (SA) but differs in that SA can have a graduated scale whereas tweets are binary: either racist speech or not and SA is interested in both positive and negative data whereas negative racist speech has no information regarding racist speech (Djuric et al., 2015). It was decided to use a sentiment approach and, for preprocessing, explore the use of Bag of Word (BOW) and Ngram[17] combined with hashingTF, TF-IDF, stemming and removal of stopwords as these are

[17] See Section 6.4.3.2 for what is meant by BOW and Ngram.

commonly used in the literature.[18]

The attributes or characteristics of the input data are known as *features*. The creation and selection of features are often said to be the most important parts of ML, even more important than the choice of algorithms and hardware on which to run them (Kaushik, 2016). A discussion of features used in this research is given in the next section.

6.4.3 Features

The feature extraction and transformation process used is illustrated in Figure 6.1. First the text is tokenised, then stopwords are removed then URIs, digits, mixtures of alphanumerics and digits, whitespace, tokens of length two or less and emojis are removed. Then if the machine learning algorithm is SVM or Bayes the data is transformed via hashingTF followed by TF-IDF; for other algorithms the data is transformed using Word2Vec.[19] Finally for all algorithms Vector Assembler organises the transformed features into a vector.

[18] There are many other SA techniques noted by Fortuna (2017) in his systematic literature review, including: distance metric, profanity windows, part of speech, lexical syntactic feature-based, rule based approaches, participant vocabulary consistency, template based strategy, word sense disambiguation techniques, typed dependencies, topic classification, deep learning, named entity recognition, topic extraction, word sense disambiguation, techniques to check polarity, frequencies of personal pronouns in the first and second person, the presence of emoticons, and capital letters, further characteristics of the message such as hashtags, mentions, retweets, number of tags, terms used in the tags, number of notes (reblog and like count) and link to multimedia content such as image, video or audio attached to the post. Although it was not practical to examine all of these techniques in relation to predicting racist tweets, a number of the message characteristics were utilised and this is discussed further in Section 6.4.3.

[19] Tests were run for each model to determine whether to use TF-IDF or Word2Vec. For brevity results are not given here, but it was found that Bayes and SVM performed best with TF-IDF and the other algorithms performed best with Word2Vec.

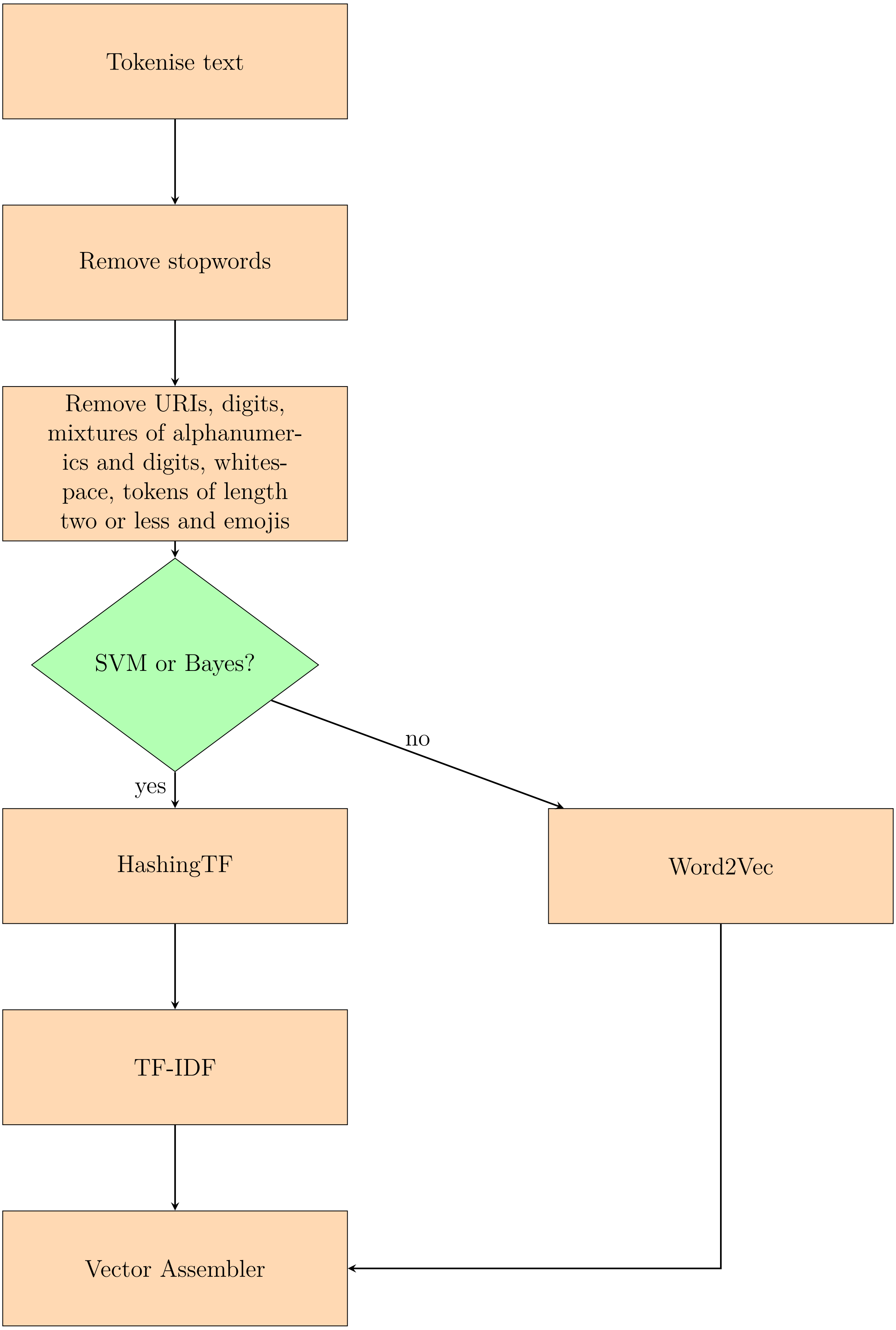

Tokenise text
Remove stopwords
Remove URIs, digits, mixtures of alphanumerics and digits, whitespace, tokens of length two or less and emojis
SVM or Bayes?
no
yes
HashingTF
Word2Vec
TF-IDF
Vector Assembler

Figure 6.1: Flowchart of feature extraction and transformation.

Most of the research on hate speech detection uses features based only on text ((Chen et al., 2012; Bartlett et al., 2014; Hosseinmardi et al., 2015; Gitari et al., 2015; Dinakar et al., 2012; Burnap and Williams, 2014; Djuric et al., 2015; Xiang et al., 2012; Burnap et al., 2015; Xu et al., 2012; Tulkens et al., 2016; Mehdad and Tetreault, 2016; Davidson et al., 2017; Malmasi and Zampieri, 2017b; Fortuna, 2017; Badjatiya et al., 2017; Park and Fung, 2017; Pitsilis et al., 2018). This is perhaps unsurprising since most of these researchers are trying to classify data based on textual input. However this ignores the considerable amount of information available from most textual data sources, for example the rich amount of metadata available on a tweet. Not all researchers ignore this data, as noted in Schmidt and Wiegand (2017) who surveyed the use of metadata in hate speech detection. They noted that a variety of such information had been used in the literature. These will be discussed in the following sections.

The use of features in the automatic detection of racist speech, can be categorised into the use of four types of features: user features, text features, geographical features and temporal features. Each of these will be discussed in turn.

6.4.3.1 Text Features

Dadvar et al. (2013) found that the use of profanity is indicative of future likelihood of hate speech tweeting but the number of replies to a post was found to be ineffective for classification (Schmidt and Wiegand, 2017, p.5) Dadvar et al. (2013) analysed the effect of user context in the automated detection of cyberbullying. They annotated YouTube comments as either bullying or non-bullying. The feature sets they used were based on content, cyber bullying and user information. The content based features contained: number of profane words, number of first and second person pronouns, profanity windows (whether profanity follows a second person pronoun within the size of the window), the number of emoticons and the ratio of capital letters (this was used as an indicator of shouting within the comment). The cyber bullying based features included a number of cyber bullying words and the length of the comment. The user based features included the history of users activities and the age of the users. Others, such as the number of comments associated to a post, have been shown to have conflicting results between different researchers.

Following Dadvar et al. (2013) this research used the following textual features:

1. The number of profane words in the tweet, based on a dictionary normalised by the total number of words in the tweet. A dictionary of 349 profane words was used (noswearing.com, 2018). This feature was denoted by *PROFANE_TWEET*.
2. The number of first and second person pronouns in the tweet. This feature was denoted by *PRONOUN_TWEET*.
3. Profanity windows of size 2 to 5 that is whether one of the pronouns was within 2 to 5 words of a profane word (Dadvar et al. (2013) used profanity windows which only considered pronoun followed by profane word). This feature was denoted by *PROFANEWIN_TWEET*.
4. A count of the number of emoticons in the tweet normalised by the tweet's length. This feature was denoted by *EMOTICONS_TWEET*.
5. The ratio of capital letters to a tweet's length. This feature was denoted by *CAPS_TWEET*.
6. The number of racist slurs in the tweet divided by the tweet's length. This feature was denoted by *RACIST_TWEET*.
7. The length of the tweet. This feature was denoted by *LENGTH_TWEET*.

   A similar set of features for a user's twitter account was also used:
8. The average number of profane words for an account. This feature was denoted by *PROFANE_ACCOUNT*.
9. The average number of first and second person pronouns for an account. This feature was denoted by *PRONOUN_ACCOUNT*.
10. The average of the number of emoticons for an account normalised by the tweet's length. This feature was denoted by *EMOTICONS_ACCOUNT*.
11. The average ratio of capital letters to a tweet's length for an account. This feature was denoted by *CAPS_ACCOUNT*.
12. The average number of racist slurs for an account divided by the tweet's length. This feature was denoted by *RACIST_ACCOUNT*.

13. The average length of a tweet for an account. This feature was denoted by *LENGTH_ACCOUNT*.

All these textual features are based on the content of the tweet. Another consideration regarding the text is how to treat tokens of the text which will be discussed in the next section.

6.4.3.2 BOW and Ngrams

Tokens of text can be treated entirely separately, which is known as a Bag of Words (BOW) approach, or they can be treated as Ngrams, where N is the number of consecutive words to keep together and treat as a single unit of input.[20] Experiments were performed to see if using BOW or Ngrams as features affected performance of the model. The Ngrams for a tweet used were: bigrams, trigrams and the set of Ngrams from length 1 to 5 (this will be referred to as 'N5') as per Burnap and Williams (2016). N5 was found to perform better than BOW. The full results are given in Section 7.1.3.

6.4.3.3 User Features

All of Dadvar et al.'s (2013) feature sets were textual except for the inclusion of user's age in the user-based features. How they determined a user's age is unclear as they give no details, and it is not available from Twitter's API. The number of posts by a user and the average of the total number of replies per follower were found to be ineffective for classification (Schmidt and Wiegand, 2017, p.5).

Dadvar et al. (2012) explored the use of gender as a feature in machine learning classification of cyber bullying, since research shows males and females use language to bully in different ways. Rather than using gender directly as a feature, they used four types of features: profane words, second person pronouns, other personal pronouns and TF-IDF

[20] For example the tweet *Spurs are going to win* would have 1 to 5grams: {Spurs, are, going, to, win, Spurs are, are going, going to, to win, Spurs are going, are going to, going to win, Spurs are going to, are going to win, Spurs are going to win}.

values, and split their dataset by gender, with 64% male, 34% female, the rest unknown. They found that gender is correlated with the likelihood of hate speech tweeting, males being much more likely to tweet hate speech.

In their analysis of sexism and racism on Twitter Waseem and Hovy (2016) compared bigrams to 4grams against a feature set with the addition of gender of the tweeter and another with gender and location. To determine gender they searched for names and other possible gender indicators such as pronouns, honorifics, and gender specific nouns within the user's profile data on a tweet and compare this to a database of nine million male and female names. They found very little difference in performance with the addition of the gender or gender and location features.

It is not possible to get the age and gender of an account/user directly from Twitter's API (ibid.), instead Waseem and Hovy (ibid.)'s method was used to approximate gender. Names, pronouns, honorifics, and gender specific nouns within the user's profile data on a tweet were searched for and compared to Kantrowitz's (2016) database of male and female names. This feature was denoted by *GENDER*.

6.4.3.4 Geographical Features

Other than its text a tweet contains four geographical fields: *coordinates*, *user.location*, *place*, and *geo*. For a tweet to include coordinates or geo information a user must have opted in to include it. The field geo has been deprecated, that is it has become obsolete and has been replaced by coordinates, although it still exists and all the tweets in D1 that had coordinates data, also had geo data.

Determing geolocation of a tweet is a difficult task as there is very little explicit location information in a tweet, and users are required to opt in to geo-tag their tweets, and around 2% or less do this (Waseem and Hovy, 2016). The data for this research contained an even smaller percentage of geo-tagged tweets: only 0.24% contained geo data.

Some of the methods that have been tried to determine tweet location in the hate speech literature will now be discussed. Mondal et al. (2017) looked at the relationship between geography and hate speech on two social networks: Whisper and Twitter. To do this they only used data from Whisper due to the low amount of geographic information in the Twitter data. Since they only looked at English-language Whispers, they found English-language countries to be the highest proportion of hate speech originators in the data: the United States, Canada and the UK accounting for 80%, 7% and 5% of the hate speech respectively. They then divided the categories of hate speech within these three countries, and this showed that hate based on behaviour was the dominant category followed by race in the United States and physical attributes in Canada and the UK. Hate based on sexual orientation was higher in the UK than in the US and Canada (14% versus 8% and 7%). They also examined the data with respect to target of the hate, in the US the number one target was black people, whereas in both Canada and the UK it was obese people. Hate of white people was sixth in both the US and Canada but did not make the top 10 in the UK. The UK hatred of people of different sexual orientations was again illustrated, as gay people were the second most targeted category whereas they were fifth in both the US and Canada. Both 'stupid' people and 'fake' people ranked highly, although fake people were only sixth highest in the UK. In the UK religious people were ninth most targeted but were not in the top 10 for either US or Canada. In addition they explored the hate volume within each US state. They corrected for different usage levels of Whisper between states by dividing the count of hate speech within a state by the total number of Whisper messages that originated from that state. They found that Western states were more prone to physical hate speech whereas southern states were more likely to hate people's race or sexual orientation.

Chaudhry (2015) investigated racist tweets in Canada between June and August 2013. He used HootSuite to search for specific racist terms in certain major Canadian cities and retrieved 776 such tweets, finding that cities with the most number of Aboriginal residents use the term 'natives' the most in their racist speech whereas those with the highest black population used the term 'nigger' the most, although these results were based on a very small sample and only a small set of racist speech words.

Hasanuzzaman et al. (2017) also utilised demographic information of tweeters, using age, gender and location embeddings in their features. In order to determine the geography of the tweet they followed Chen and Neill's (2014) method to determine the geographical location of the tweet. They first searched the text message for a location, then if they did not find one they used the *geo* information from a tweet, and finally try to extract geographical information from the user's profile. The geographical information they found was only at the country level. They found that the demographic variables did provide some improvements but these were dependent on the type of textual data used: bigram, trigram and word2vec feature sets showing an improvement but Ngrams (1 to 4grams) performed worse with the addition of the demographic variables.

In exploring the geography of a tweet, Waseem and Hovy (2016) noted that only around 2% of Twitter users knowingly provide geographical information. Instead they searched name and username for any indicators of geographical location. They then mapped any location found to its time zone. If no location was found they used time zone directly as a geographical feature, following the lead of Gouws et al. (2011a) who also used time zone as an geographical indicator. They found very little difference in performance with the addition of the gender or gender and location features, and they found that location actually negatively affected F1-score.

The percentage of users with geo information in the data for this research was only 201,996 out of 83,994,885, that is 0.24% which is a considerably lower percentage than the approximately 2% on average of Twitter users that disclose their location (Waseem and Hovy, 2016). It could be hypothesised that those using ethnic slurs in their tweets are less likely to disclose their locations since they do not want to be identified, so a smaller percentage of racist tweets that include geo-tags might be expected, but the majority of the data was likely to be nonracist, so it is not clear why there was such a small percentage of geo-tagged tweets in the whole sample collected for this research.

From the literature the text of a tweet is often used as an indicator of the tweet's geographical location. From the data collected for this research it was found that the text content was a poor predictor of location for some cities and countries but a good one for others. For example, within the data's locations London, Paris, New York and Manchester were used almost exclusively in tweets not originating from those cities or even within the countries in which those cities reside. However Nigeria and Delhi were used in tweets originating from that country and city respectively. It was the more ubiquitous locations that were unlikely to be representative of their tweet's origin, so it was decided to not use the tweet text for geographical information, unless there were no other geographical indicators in a tweet.

Table 6.3 gives counts for coordinates, user.location, place and geo for both D1 and the annotated sample.[21] For D1 (N=83,994,885) it can be seen that coordinates and geo have the same count, of 201,996 which is 0.24% of the data. The value for place was a higher at 2,156,689, around 2.6% of the data, and location was the most commonly found geographical feature with a count of 50,191,127, which is 60% of the data. For the annotated sample (N=84,000) it can be seen that coordinates and geo have the count, of 551 which is 0.66% of the data, a greater percentage than for D1, but still lower than the 2% figure reported in the literature. The value for place was a little higher at 2,420, nearly 3% of the data, and location was the most commonly found geographical feature with 51,639, which is 61% of the data. Between the two datasets the proportion of tweets that contain geo/coordinates information was greater in the smaller dataset, but the proportion of tweets with the other three measures was roughly the same. So from this data a sensible strategy to geo-locate a tweet might appear to be to rely on location first, then place and finally coordinates. This strategy however is not without problems, which can be seen if the data is examined. Place gives geographical coordinates of the form:

Table 6.2: Counts for coordinates, user.location, place and geo for D1 and annotated sample.

Dataset	**coordinates**	**location**	**place**	**geo**
D1	201,996	50,191,127	2,156,689	201,996
Annotated Sample	551	51,639	2,420	551

```
{"country":"United Kingdom","country_code":"GB","full_name":
"Aylesbury, England","bounding_box":{"coordinates":
```

[21] This is the sample annotated by the researchers, further details are given in Section 5.10.

```
[[["-0.858287","51.792766"],["-0.858287","51.838575"],

["-0.772721","51.838575"],["-0.772721","51.792766"]]],"type"

:"Polygon"},"place_type":"city","name":"Aylesbury",

"attributes":,"id":"5461b744712914a7","url":

"https://api.twitter.com/1.1/geo/id/5461b744712914a7.json"}
```

which gives country, city and bounding latitudes and longitudes, values generated by Twitter, which allow for unambiguous geographical information to be retrieved. Coordinates also gives Twitter generated unambiguous geographical information, for example: "coordinates":["-97.72453","30.26163"],"type":"Point". User.location data, on the other hand, is free-form and input by the user. Google provides an API call, *geocode* which can resolve a location from such free-form text.[22] The geocode API call is very good at finding locations from free-form data, however it does have its drawbacks. It only allows:

> 2,500 free requests per day, calculated as the sum of client-side and server-side queries. 50 requests per second, calculated as the sum of client-side and server-side queries (Google, 2017).

Although Google does allow up to 100,000 requests per day, at the rate of $0.50 USD per 1,000 additional requests, this was not financially viable for this research. Additionally 50 requests per day is slow when dealing with millions of tweets. The locations from the annotated sample dataset were geocoded using this API over several days, although interestingly more than 2,500 requests were often obtained on a day. While this strategy was viable for the small dataset, it was not possible to use geocode for D1. The other problem with using user.location data is the free-form nature of the text. Many of the locations are nonsense, or not decodeable by the geocode API. For example all of the following locations were encountered in the annotated sample dataset: SecondLife Solaris

[22] The API returns a result containing the following: latitude, longitude, accuracy, formatted_address, address_type,and status, where accuracy represents the scale of the location from country down to premise, formatted_address is a human readable form of the address, address_type is a more informative version of and a replacement for accuracy and status is an indicator of the success, or not, of the API call, ‘OK’ denoting a successful call.

Island, TheSecret, 901, HighwayAllDay, RIPPaul ♡, TeamGod Philippians 4:13, and ProudMuslim. Ultimately 17,256 locations were able to be decoded by geocode, some examples are given in Table 6.3, which are the first 20 user.locations geocoded from the annotated tweets sample.

From Table 6.3 it can be seen that Google correctly geocodes for example 'Minnesota, USA' as an administrative area, 'UK' as the country 'United Kingdom' and 'Welford Road, Leicester' (which is a rugby stadium in Leicester) as a point of interest with address 'Aylestone Rd, Leicester LE2 7TR'. It does on occasion seem to have a US bias, with '574' coded as 'FL-574, Florida, USA' and 'Global' coded as '2560 Anthem Village Dr #160, Henderson, NV 89052'. It also geocodes '221B Baker Street', which is presumably a reference to the fictional address of Sherlock Holmes, as' 221b Baker St, Buena Vista, GA 31803, USA'. It, unsurprisingly, fails to geocode 'main: @letmeknowmp3' but it also struggles with 'not so Great Britain.' which it decodes as 'Great Western Dockyard, Gas Ferry Rd, Bristol BS1 6TY' which is the site of the ship SS Great Britain, and 'cheshire xx' becomes '110 Great Russell St, Fitzrovia, London WC1B 3NA', which is the Cheshire Hotel. Even in this small sample, of these 20 addresses 12 are undoubtedly correct, but the remaining eight are all likely to be wrong. So from this brief look at the geographical data on a tweet, it can be concluded that the data is either likely to be missing and possibly unreliable. As a result it was decided to follow the lead of Gouws et al. (2011a) and Waseem and Hovy (2016) who both used time zone as a geographical indicator with information from the geographical data on a tweet used, when time zone is missing. From the annotated sample data the feature *GEO_TWEET* was created using the following rules:

1. If a tweet contains a timezone field, use that. If not go to 2.
2. If a tweet contains a coordinates field, convert its latitude and longitude to a country using geonames. [23] Then convert country to timezone, otherwise go to 3.
3. If a tweet contains a places field, convert its latitude and longitude to a country using geonames. Then convert country to timezone, otherwise go to 4.
4. Convert user.location to a country using geocode. Then convert country to timezone.

[23] For details on how to use geonames, see `https://stackoverflow.com/questions/14334970/convert-latitude-and-longitude-coordinates-to-country-name-in-r`

Table 6.3: Examples of user.location and their corresponding geocode results.

user.location	**geocode result**
Minnesota, USA	46.729553,-94.6858998,administrative_area_level_1,"Minnesota, USA","administrative_area_level_1,political",OK
San Antonio, Texas	29.4241219,-98.4936282,locality,"San Antonio, TX, USA","locality,political",OK
Multicultural Center	40.196562,-85.4080424,premise,"Multicultural Center, 325 N McKinley Ave, Muncie, IN 47303, USA",premise,OK
UK	55.378051,-3.435973,country,United Kingdom,"country,political",OK
Dublin, OH	40.0992294,-83.1140771,locality,"Dublin, OH, USA","locality,political",OK
Welford Road, Leicester	52.6242254,-1.1330856,establishment,"Aylestone Rd, Leicester LE2 7TR, UK","establishment,point_of_interest,stadium",OK
574	27.9813981,-82.3213571,route,"FL-574, Florida, USA",route,OK
Worthing, England	50.81787,-0.372882,locality,"Worthing, UK","locality,political",OK
Grand Blanc, MI	42.9275277,-83.6299518,locality,"Grand Blanc, MI 48439, USA","locality,political",OK
Sheffield, England	53.381129,-1.470085,locality,"Sheffield, UK","locality,political",OK
niagara university	43.1365152,-79.0353197,establishment,"5795 Lewiston Rd, Niagara Univ, NY 14109, USA","establishment,point_of_interest,university",OK
Taco Bell	33.638714,-112.2260202,establishment,"7714 W Bell Rd, Glendale, AZ 85308, USA","establishment,food,meal_takeaway,point_of_interest,restaurant",OK
Winfield, WV	38.5331448,-81.8934675,locality,"Winfield, WV 25213, USA","locality,political",OK
Chattanooga, TN	35.0456297,-85.3096801,locality,"Chattanooga, TN, USA","locality,political",OK
main: @letmeknowmp3	„„,ZERO_RESULTS
New York, N.Y.	40.7127837,-74.0059413,locality,"New York, NY, USA","locality,political",OK
not so Great Britain.	51.4491712,-2.6084058,establishment,"Great Western Dockyard, Gas Ferry Rd, Bristol BS1 6TY, UK","establishment,museum,point_of_interest",OK
cheshire xx	51.517503,-0.12956,establishment,"110 Great Russell St, Fitzrovia, London WC1B 3NA, UK","establishment,lodging,point_of_interest",OK
221B Baker Street	32.3236754,-84.5200332,street_address,"221b Baker St, Buena Vista, GA 31803, USA",street_address,OK

If this fails, or there is no user.location go to 5.

5. Search for geographical information in a tweet's text. Lookup country and convert to time zone.

This method provided a reasonable amount of geo-data: 48% of tweets were geocoded using the time zone field, and each of the other steps provided approximately 5% more geocoding, with a total of 68% of tweets geocoded.

6.4.3.5 Temporal Features

It is perhaps surprising that geographical, rather than temporal data, is more often used in analyses, since the geographical information on a tweet is prone to inaccuracy, as discussed above. The temporal information on a tweet is far more prevalent; every tweet has at least timestamp information, something which cannot be said to be true for geographical information. Also the analysis of temporal data and crime, lends itself to an exploration of RAT in keeping with the literature on cybercrime and RAT (see for example Leukfeldt and Yar (2016)). Ratcliffe (2010) is one of the many papers analysing the geography of crime but the authors also consider the temporal aspects and note the relative paucity of research when compared to geographical aspects. The temporal aspect is certainly important in the study of online crime, indeed the geographical aspect might be argued to be irrelevant since the 'offences' occur in cyberspace or at least less relevant since the distance between the 'target' and 'offender' is negligible in cyberspace thus not an impediment to the commission of a criminal act, whereas geography is an important factor in determining whether 'real-world' criminal acts are committed.

The interaction of time of day and criminal events has largely been restricted to the examination of the timings of real-world crime and its use in crime science and criminal investigation (Newton et al., 2014, for example). There is very little analysis of temporal features with respect to tweets in the literature. However Myers and Leskovec (2014) noted the 'bursty' nature of Twitter output, that is Twitter activity is often greater around events likely to invoke strong public reaction. For example, when Fusilier Lee Rigby was murdered Williams and Burnap (2015) noted an increase in tweets containing hate speech. The data collected for this research has been aggregated to avoid the influence of such triggering events.

The one work that does examined temporal features is Vosoughi et al. (2016) who collected 18 million geocoded tweets over three years. From these they sampled 3,000 tweets which were hand annotated as containing either positive or negative sentiment. They used these as input features into a NB classifier. They examined three temporal variables: hour of day, day of week and month. They found that the tweets were evenly divided over month, hour and day, and they found no significant difference in amount of tweets for any of the three measures of time. They found that as the week progressed from Monday to Saturday the positivity of the tweets rose, with a drop on Sunday, and a significant drop to Monday. They also found the hourly distribution was most positive between the hours of 10 AM and midday, with a gradual drop to 8 PM, then a rise to midnight followed by a drop to 3 AM, then a gradual rise to 10 AM.

In order to include temporal features a number of fields in a tweet need to be considered as well as what time period is of interest. It was decided that hour of day, and day of week are likely to be of interest, since, according to RAT, the routine activities of racist tweeters might lead to tweets at similar times of day or days of week. There was insufficient data to use month or year as a feature.

Tweets contain the attribute, *created_at*, that contains the UTC time that the tweet was created. When analyzing hour of day, UTC time is not sufficient, since if a tweet is sent from a country that is not using UTC time, then the actual time of the tweet within that country will be *created_at* plus or minus an offset determined by whichever time zone the country is in. There are two fields in a tweet with time zone information: *utc_offset* and *time_zone*. *time_zone* is set by the user and *utc_offset* is generated from this as the time in seconds that *time_zone* differs from UTC. These are potentially inaccurate or may not have been set at all (Twitter, 2017). Of the 83,994,885 tweets collected for this research, 41,260,026 were found to have time zone values i.e. 49.1%. So to calculate time of the tweet the following formula was used:

$$created\_at \pm (utc\_offset/3,600).$$

From this the hour, and day can be extracted. These features were denoted by *HOUR_TWEET* and *DAY_TWEET* respectively.

Each of the features discussed were added to N5 and run with a SVM model with tenfold cross validation (for explanations of these terms, see later in this chapter). The model was also run with no additional features and this was denoted by BASELINE.

Only HOUR_TWEET was an improvement on the BASELINE. DAY_TWEET was the next best performing additional feature, and so only these two featues were used in further analysis. The full results are given in Section 7.1.2.

In order to prepare the features Spark's methods were used. These are discussed in the next section.

6.4.4 Spark Methods

Spark has a number of methods that can be used to handle the preparation of features for input into a model. These are broadly grouped into: extraction, transformation, selection and local sensitive hashing (LSH) methods.[24] The Spark extraction methods used in this research are: Tokenizer,[25] TF-IDF, HashingTF and Word2Vec. The Spark transformation methods used in this research are: StopWordsRemover, StringIndexer, VectorAssembler and IndextoString. Also two User-Defined Functions (UDF): removeRegexUDF and ngramCreator are used as is a Stanford core NLP lemmatization transformer. These methods will be discussed below.

[24] Spark's selection methods perform various transformations of features that were not necessary for this research. Spark's LSH methods mostly apply to clustering algorithms and clustering was not performed in this research. Neither Spark's selection or LSH methods will be discussed further.

[25] Spark lists tokeniser as a transformation method, but it extracts from the raw data, and does not scale, convert or modify features.

6.4.4.1 Tokenizer

Tokenizer is a Spark transformer that tokenises text, splitting it on white space by default. For BOW Tokenizer is used to split the *text* field of a tweet into tokens.

6.4.4.2 ngramCreator

If Ngrams are used instead of BOW, then ngramCreator is used instead of Tokenizer. Use of ngramCreator can be seen in Listing A.1.

6.4.5 removeRegexUDF

It was decided to remove emoticons to standardise the text since they have only marginal value in classification tasks, and additionally to convert hashtags to their underlying words by removing the hash character since these also have marginal utility in classification (Agarwal et al., 2011). *removeRegexUDF* is a user-defined function written in Scala that performs these tasks.

6.4.6 Stanford Core NLP Lemmatization

To lemmatise the data the Stanford core-NLP version 3.6.0 (Manning et al., 2014) was used via a wrapper created by Databricks (Databricks, 2016). The lemmatisation process was considerably time-consuming, increasing the time taken for algorithms to run from the order of a few minutes, to many hours. As a result a program was written to perform a lemmatisation process once and write the results to a hive table. Then subsequent runs of algorithms used the lemma look up table and this significantly reduced processing time.

6.4.7 Spark Extraction Methods

The following methods act on either single tokens (BOW created by Tokenizer) or Ngrams (from 1 to 5grams created by ngramCreator).

6.4.7.1 HashingTF

Prior to calculating TF-IDF, the raw data must be converted to a vector form that TF-IDF can use. HashingTF is one such converter algorithm that transforms the data by taking the tokenised input words, converting these words to numbers by hashing them. To index the words hashingTF hashes each term using the Murmurhash 3 function and uses the generated hashes as indexes. This is a fast and efficient method of indexing, but potentially may lead to collisions if two hashes are calculated to be the same, although this is highly unlikely (Chen, 2015).

6.4.7.2 Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF is a measure of how important a word is within a corpus of documents (Ramos, 2003). It is calculated by first calculating two components: Term Frequency (TF) and Inverse Document Frequency (IDF). TF is a measure of how often a term occurs in a document, usually this is calculated by counting the number of times a word appears in a document and dividing by the total number of terms in the document. This normalisation is necessary since otherwise long documents would usually have higher TF scores than shorter ones. IDF is a measure of how important the term is within a set of documents. Frequently occurring words are unlikely to be important since these will be mostly stop-words (such as 'the'), so IDF gives more weight to rare words (tfidf, 2018). Spark (Spark, 2018b) calculates IDF of a term, t in a corpus D as,

$$IDF(t,D) = \log\left[\frac{|D|+1}{DF(t,D)+1}\right],$$

where $|D|$ is the total number of documents in the corpus, and $DF(t, D)$ is the number of documents that contain term t. TF-IDF is simply the product of TF and IDF.

6.4.7.3 Word2Vec

Word2Vec is an alternative method (versus TF-IDF) of representing words in a vector form. It represents similar words as close together within the vector space and so captures the semantic meaning of text, something which TF-IDF fails to do. Spark utilises the skipgram model for Word2Vec. This uses a simple neural network with one hidden layer, that aims to generate vectors that represent words with similar words close together (Apache.org, 2018).

6.4.8 Spark Transformation Methods

6.4.8.1 StopWordsRemover

Twitter sentiment analysis needs to overcome the problem of noisy data. One method aimed at reducing noise in the data is to remove *stopwords*, that is words that are not significant and provide little information in a message (Saif et al., 2014). Stopword lists are somewhat arbitrary and it has been questioned whether a pre-determined list of stopwords is ideal for short message research (ibid.). However extraction of stopwords from a list will be used, since much of the literature performs this step in text preprocessing, for example Agarwal et al. (2011). Spark has a built-in StopWordsRemover function which uses the file english.txt stored in Spark's online repository (Spark, 2017c). This file contains 181 stop words. However, for this research, the more comprehensive stopwords list containing 667 words, from ranks.nl was used, since this has been used successfully in other classification research (for example, see Neumayer (2006)).

6.4.8.2 StringIndexer

StringIndexer converts categorical data into a numerical vector representation of indices. The indices start at zero and are labelled by frequency, the most commonly occurring label is given the value zero and so on (Apache.org, 2018). This transformer is used on the racist column to convert the labels, true and false, to 0 and 1 respectively.

6.4.8.3 LabelIndexer

LabelIndexer indexes the label column.

6.4.8.4 VectorAssembler

VectorAssembler transforms a set of columns into a single column vector (ibid.). In this research it is used to combine the hour of day and day of week vectors with the text vectors, so that they can be used as a combined input features vector.

6.4.8.5 IndextoString

After a model is run, it generates a column of label indices, which can be mapped back to the original string labels using IndextoString (ibid.).

All of the methods discussed above act on the data either prior to, or after, the machine learning algorithm runs. The actual machine learning algorithms that were run are discussed in the next section.

6.4.9 Machine Learning Algorithms

There is considerable variability in performance of machine learning algorithms, and the choice of algorithm for a particular problem needs to be carefully considered (Caruana and Niculescu-Mizil, 2006). Fortuna (2017) performed a systematic literature review of automatic hate speech detection and found that the most commonly used algorithms were: Support Vector Machines, Random Forests and Decision Trees. However other algorithms were also evaluated as part of this research, since the historical selection of machine learning algorithms is likely to be partly determined by computing resources. The adoption of Spark and Hadoop has made the use of computationally intensive algorithms like Artificial Neural Networks (ANN) more viable, so it was of interest to determine their performance against more 'traditional' machine learning algorithms. A brief discussion of the theoretical bases of NB, LR, SVM, RF, DT, GBTs and ANN now follows.

6.4.9.1 Naive Bayes

NB is a method based on *Bayes' Theorem* which can be expressed for events A and B thus:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}.$$

This equation states that the probability of A given B, is equal to the probability of B given A, multiplied by the probability of A, divided by the probability of B. P(A) and P(B) are known as the *prior probabilities*, that is they are the probabilities of the events occurring prior to any information on how the events affect one another. P(A|B) is known as the *posterior probability* since it is the probability of A occurring conditioned on the event B. P(B|A) is known as the *likelihood* (Mohammed et al., 2016). If data is represented as a vector, $X = (x_1, x_2, ..., x_n)$, (that is the data has n features), then Bayes' theorem can be used to calculate the posterior probability, $P(c_j \mid X)$ which is the probability that the data is from class c_j, one of K possible classes given X. This probability is calculated as:

$$P(c_j \mid X) = \frac{P(X \mid c_j)\, P(c_j)}{P(X)}.$$

The *Naive* Bayes classification model is so-called because using it assumes that the probability of classification to class c_j given some feature x_i is independent of all the other features. In other words the probability of interest is

$$P(X \mid c_j) = P(x_1 \mid c_j) * P(x_2 \mid c_j) * ... * P(x_n \mid c_j),$$

which is straightforward to calculate. This assumption of independence is thought to be unlikely in most cases, yet the NB classifier has been shown to perform well in many situations (Marsland, 2015).

6.4.9.2 Logistic Regression

Logistic regression is the application of linear regression to classification problems, using the logistic function. It is based on the linear regression model that assumes an outcome z is related to n attributes, x_1, ..., x_n and a series of weights β_0, ..., β_n by the linear equation:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n.$$

The logistic function is a mathematical function that maps its input into the interval 0 to 1, without actually reaching either 0 or 1. When applied to this linear regression case, it is of the form:

$$P(C_0|x_1,...,x_n) = \frac{1}{1+e^{-z}},$$

with $P(C_0)$ being the probability of class 0. So the outcome of logistic regression is actually a probability, and this is translated to a class for a binomial situation, by the default of probabilities of greater than or equal to 0.5 giving an output of class 1, otherwise class 0 (Sammut and Webb, 2017).

6.4.9.3 Support Vector Machines

SVMs are a class of machine learning models that can be applied to classification and regression problems, first developed by Vladimir Vapnick (Vapnik, 2013), SVMs provide one optimal way to separate two classes. In Figure 6.2 the left-hand graph shows a situation where two classes can be separated by an infinite number of straight lines (Kuhn and Johnson, 2013).

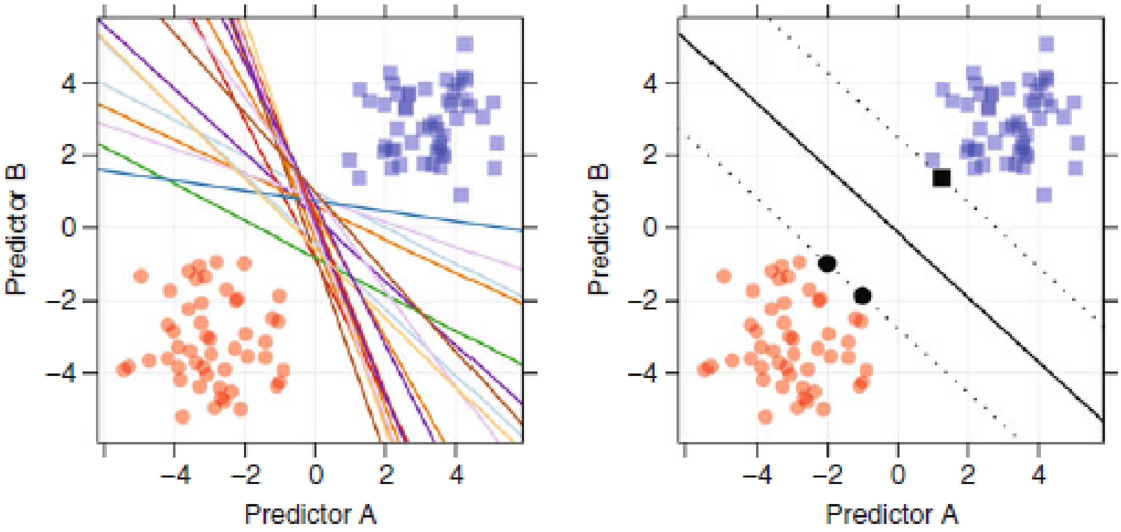


Figure 6.2: SVM linear maximium margin classifier, from Kuhn and Johnson (2013, p.344).

Since this is a classification task the two classes must be separated optimally but metrics such as accuracy are ineffective since they would be the same for each linear separator. Instead SVM uses a new metric known as the *margin*. The margin is the distance from the separating line to the closest point in each of the two groups as can be seen in the right-hand graph, where the margin is the distance from the solid black line to the dotted lines. The solid black line represents the *linear maximum margin classifier* which is the line that maximises the margin (ibid.). As a result the performance of SVMs is independent of the number of features (Joachims, 1998). The margin is determined by the points that are closest to the separation line, which in the right-hand diagram are signified as black circles or black squares, and these points are known as *support vectors*, and are sufficient to determine which class any new data is classified as (Kuhn and Johnson, 2013).[26]

6.4.9.4 Decision Tree

DTs handle classification problems by breaking them down into a series of decisions. The first decision is performed at the root of the tree and then there are a series of branches, at each of which another decision is made, eventually leading to the final classification at the tree's leaf. A DT is designed to classify an object into one of two or more classes by asking a series of questions concerning the object's attributes. In an optimal DT at each stage the question asked must be one that provides the most information, which is determined by measuring *entropy*,[27] the decision with the highest entropy should be

[26] A straight line is the boundary in two dimensions, but more generally SVM attempts to classify data by separating it by a *hyperplane* that is an n -1 dimensional surface within an n-dimensional space. This is the case for two features of the data, but of course there may be many more features, and SVM classifies n-dimensional data using what is known as the *kernel trick*. In the original 2-D space the distance between points was of interest, but for n-dimensional space these distances vectors can be multiplied by some kernel function, K which maximise the distance in n-dimensional space. Two-dimensional data may not be linearly classifiable but when mapped to 3 or more dimensions there may be a separating hyperplane that maximise separation between classes, as can be seen in Figure 6.3. In this representation no straight line can be drawn between the blue circles and red crosses but when another dimension is added the new visualisation shows a green plane which maximises the separation between the blue circles and red crosses. For this research linear kernels were chosen for the SVM as these are found to be just as effective as polynomial, RBF or sigmoid kernels (Zhang and Lee, 2003) but simpler.

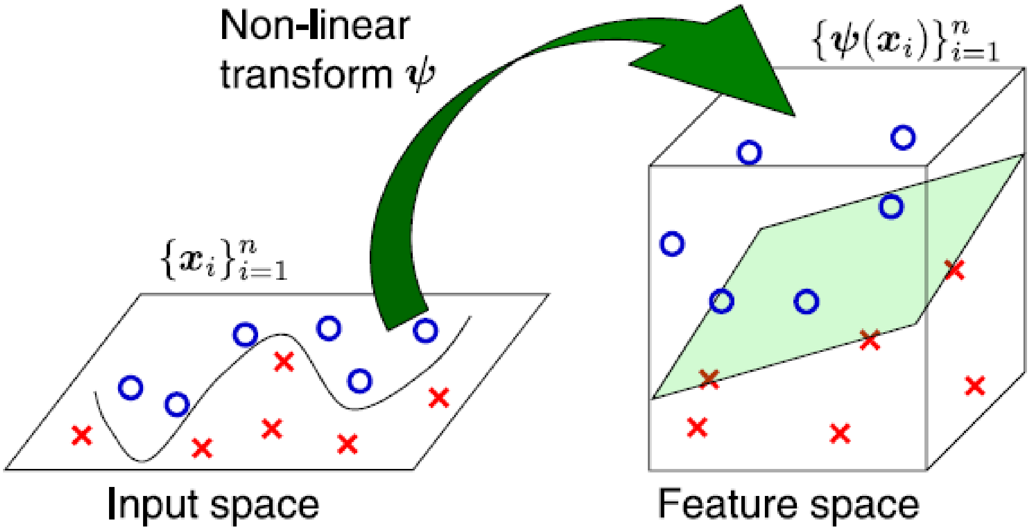


Figure 6.3: Nonlinearization of support vector machine by kernel trick, from ibid., p.311.

[27] Informally, entropy can be thought of as a measure of disorder. Each node on the DT is optimal when it provides the maximum amount of information. How much information the node produces can be

asked closest to the root (Marsland, 2015).

6.4.9.5 Random Forests

A RF classifier is based on the idea that by combining DTs an improved classifier is created as long as there is enough variety between the trees. An element of randomness has to be added to the trees and this is done by what is known as *bagging*: random bootstrap samples are selected from the dataset and a different one is given to each tree. In addition to this each tree is given a random subset of the features, and the tree can only pick from that subset. For an optimal RF trees are added until the error stops decreasing. The classification result is determined by majority vote of the set of trees (ibid.).

6.4.9.6 Gradient Boosted Tree

GBT is an ensemble classifier that comprises a set of regression trees that act on a sample of data. Figure 6.4 shows a series of regression trees making up a GBT classifier.

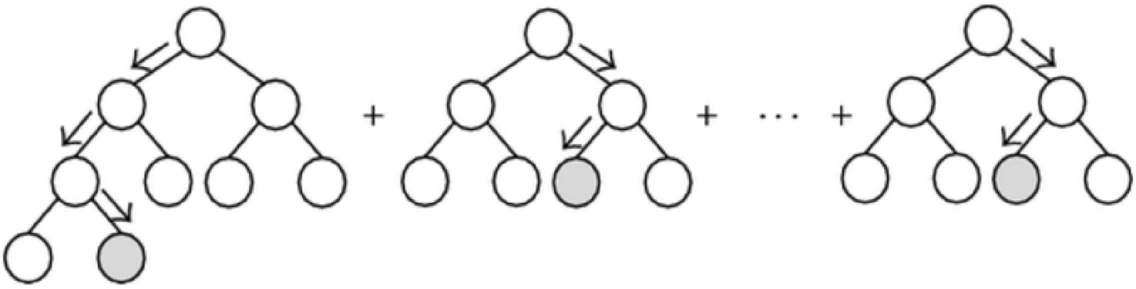


Figure 6.4: Gradient boosted decision tree ensemble, from Shin (2015, p.2012).

At the root level each regression tree provides two possible outcomes, and each node has binary outcomes at lower levels. Each of these binary nodes splits are on a feature at a specific value. The arrows in the figure indicate possible paths through each of the trees

calculated as: the entropy prior to the node minus the entropy after the node. The mathematical details of calculating entropy are not important to this research, but can be found in, for example, Safavian and Landgrebe (1991).

and the shaded circles are the final responses for each tree. The sum of the responses is the score for the particular sample (Ye et al., 2009). GBT comprises two parts: gradient descent (or ascent) and boosting. Gradient descent is an algorithm that attempts to optimise a function in a series of steps. This requires a predetermined precision to be set, and at each step the algorithm determines whether the functions minimum (or maximum for ascent) has been reached to the level of precision required (Mohri et al., 2012).

The other part of GBT is boosting, which converts the ensemble of gradient algorithms (weak learners) into a single algorithm (a strong learner) (Friedman, 2001). This aggregation of learners means that GBT is less prone to overfitting but is very sensitive to noise and is computationally expensive (Ye et al., 2009; Long and Servedio, 2010).

6.4.9.7 Artificial Neural Network

An Artificial Neural Network (ANN) is an algorithm that models the behaviour of the brain, consisting of a series of *neurons* connected in layers. Spark implements a feedforward ANN in its Multilayer Perceptron Classifier (MLPC). Spark's MLPC has a series of node layers with each layer's output becoming the input for the next layer. Each node in a higher layer is connected to each node in the next layer, as can be seen in Figure 6.5 (Kordos, 2005).

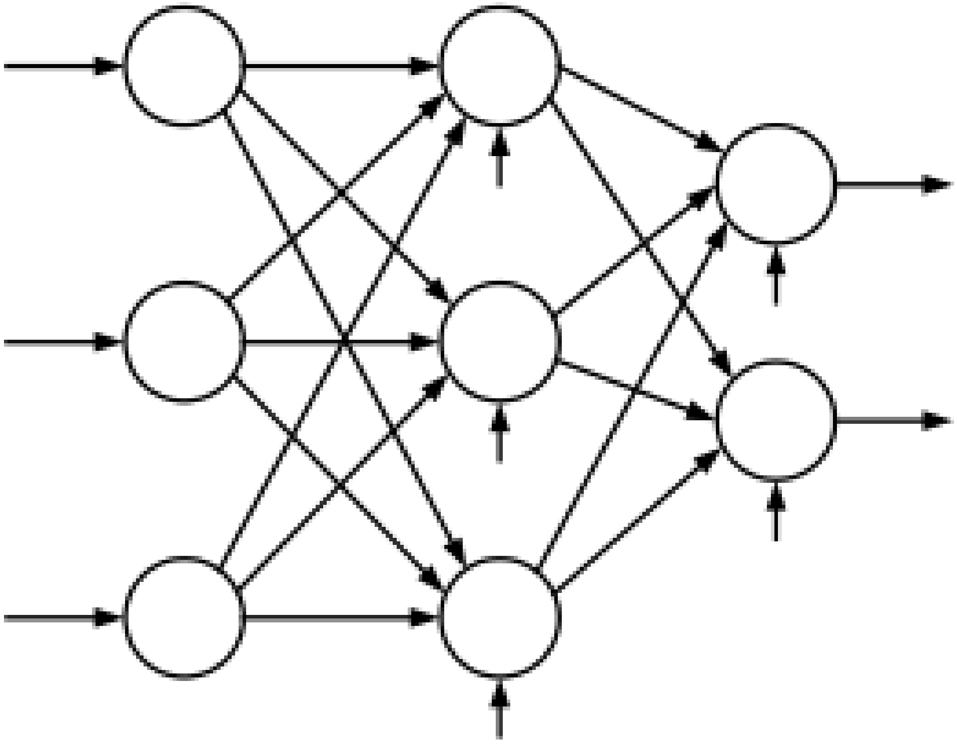


Figure 6.5: MLPC network, from Kordos (2005).

Each neuron has a *net function* and an *activation function.* The net function determines the weights of inputs to a neuron, each neuron also has a *bias* with the weight, w_0 as represented in Figure 6.6.

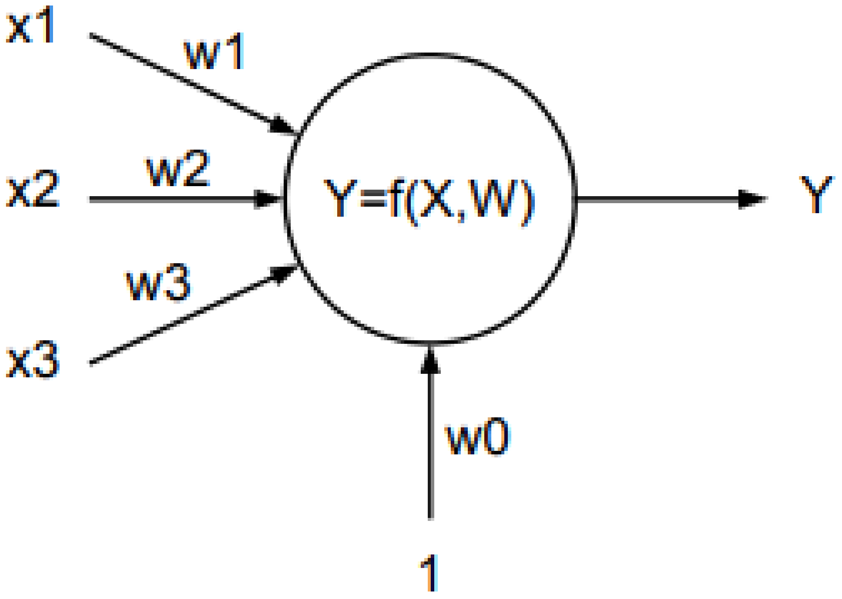


Figure 6.6: Neuron model, from Kordos (2005).

The activation function determines the output from the neuron (Kordos, 2005). Spark uses the logistic function for intermediate layers, and the softmax function for the output layer. It uses logistic loss function for optimization and L-BFGS as an optimization routine (Spark, 2018a).

Whichever model is used it can be modified as discussed in the next section, in order to optimise certain metrics which are discussed in Section 5.12.2.

6.5 Chapter Summary

This chapter provides a discussion of the relationship between big data and criminology and gives further details on the theory of machine learning. There are sections on *features*, that is the dimensions of the data used as input to predict. Text, user, geographical and temporal features are discussed. Then Spark extraction and transformation methods are explored. This is followed by section on the different algorithms utilised in this research.

Big data is often described in terms of a number of V's, originally these being three V's: volume, variety and velocity. Volume refers to the amount of data, velocity is the rate at which data is created or processed, and variety refers to the fact that data can come from disparate sources with different formats and types of data. This type of definition of big data is a feature-based definition, focusing on technical issues such as how data is stored or processed. Other feature-based definitions go beyond the three V's (Chan and Bennett Moses, 2016). Chan and Bennett Moses (ibid.) argue that feature-based definition are problematic since they relate to the current state of computing and what currently is difficult to process or handle using 'traditional' techniques may not be so in the near future. For Chan and Bennett Moses (ibid.) and Boyd and Crawford (2012) big data is more than merely a technological construct, it often includes an interaction between three things: the technological aspect of handling large datasets, the analysis of these datasets and the mythology that suggests that big data provides an objective 'truth', that it is better than traditional techniques. While it might be the case that there is a mythology around big data, it does not seem useful to include this within a definition of big data. The idea of Chan and Bennett Moses that the term 'big data' has become inextricably linked with this mythology, is debatable, and even if it is true, it is much more fruitful to narrow down what is meant by big data, instead of expanding it to include subjective views of its efficacy. The conflation of big data with social network data seems to be fairly rife in the literature and should be avoided. None of the definitions so far have made much of the fact that big data requires a considerable amount of *processing power*. The storage of the data for this research could reasonably be handled by a single machine, but the *processing of it* could not. Since the underlying methodological framework of this research is pragmatism, a definition of based on practical aspects of big data is appropriate. A definition needs to highlight the nature of both the data and its processing. Such a definition, that reflects how big data is different to any other analysis with respect to processing and size of data, can succinctly be expressed as:

> big data is data that requires parallel processing either because of its size or analysis requirements.

The attributes or characteristics of the input data are known as *features*. The creation and selection of features are often said to be the most important parts of ML, even more important than the choice of algorithms and hardware on which to run them (Kaushik, 2016). Following Dadvar et al. (2013) this research used a number of textual features of a tweet, for example the number of profane words in a tweet, and also a set of textual features of a user's twitter account, such as the average length of a tweet for an account. All these textual features are based on the content of the tweet. Another consideration regarding the text is how to treat tokens of the text. Tokens of text can be treated entirely separately, which is known as a Bag of Words (BOW) approach, or they can be treated as Ngrams, where N is the number of consecutive words to keep together and treat as a single unit of input. Experiments were performed to see if using BOW or Ngrams as features affected performance of the model.

Gender was used as a user based feature and for geographical data a feature was created using the time zone of the tweet, or if that was unavailable a tweet's various geographical fields were used. For temporal features, hour and day of the tweet's creation were used. Of all these features only the hour was an improvement on no features and day was the next best. As a result only these two features were used in further analysis.

The features we used as input to the following algorithms: Naive Bayes (NB), Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), Decision Tree (DT), Gradient Boosted Tree (GBT) and Artificial Neural Network (ANN). In order to determine which is the 'best' model, the possible outcomes for the model need to be considered.

The next chapter presents the results of the research.

Chapter 7

Results

This chapter presents results for the machine learning procedures and qualitative analysis. For the machine learning results first the effects of varying oversampling fraction of negatives are given. This is followed by a discussion of text, user, geographical and temporal features and their efficacy as input to the machine learning algorithms. Features are further considered in both discussing whether text should be treated as Ngrams or BOW and whether hour, hour+day or neither should be used as additional features. Then seven different algorithms are compared by analysing their metrics.

The temporal aspects of the tweets are further discussed in relation to the different datasets including both the input and predicted data. Then the qualitative results are discussed, first summary data from an NLTK analysis is given, followed by analysis of word clouds and word trees of the data. Then a brief discursive analysis is given followed by a grounded theory analysis of the data.

Finally accounts are discussed and results of machine learning processes aimed at predicting which accounts are racist, using metrics from the accounts, from the grounded theory analysis and a combination of these.

The relevant datasets these processes used were: the hand coded nonracist tweets, designated as *Inp NR*; the hand coded racist tweets, designated as *Inp R*; the predicted nonracist tweets, designated as *Pred NR*,[1] and the predicted racist tweets, designated as *Pred R*. These designations are given in Table 7.1.

Table 7.1: Description of the four datasets: Inp NR, Inp R, Pred NR and Pred R.

Data source	**Racist**	**Number of tweets**	**Dataset designator**
Hand-annotated sample	No	80,040	Inp NR
Hand-annotated sample	Yes	3,960	Inp R
Predicted by machine learning	No	40,203,196	Pred NR
Predicted by machine learning	Yes	972,830	Pred R

7.1 Machine Learning Results

With respect to the machine learning results in this chapter, the algorithms were run on the hand annotated sample from the D1 dataset split in the ratio 70:30, that is 70% training and 30% test. Due to the data's imbalance (see Section 5.11), in order to find a ratio of positives to negatives that would give optimum results, it was necessary to reduce the ratio of positives to negatives, and so the results of varying the fraction of negatives is now discussed.

7.1.1 Choosing Fraction of Negatives

As well as the model/features/preprocessing choices, oversampling was necessary (as discussed in Section 5.11) and this affects the models. The effects of varying oversampling was investigated by examining the output for different *fraction of negatives*.[2] To determine the optimum fraction of negatives the SVM & Lemma+Time+Ngrams model was run[3] with different values of this ratio and its metrics: Accuracy, AUPRC, AUROC, F-score for $\beta = 1$ and $\beta = 0.5$ are given in Figure 7.1 and Table B.3 for fraction of negatives

[1] *Pred NR* and *Pred R* were created by machine learning process.

[2] Fraction of negatives is the proportion of negatives in relation to the proportion of positives being one, i.e. if the proportion of negatives is 0.4, then the ratio of positives to negatives is 1:0.4, and the fraction of negatives would be $0.4 \div 1.4 = 0.286$.

[3] The SVM & Lemma+Time+Ngrams model was chosen as it performed best overall (see Section 7.1.4). Of course there is a circular relationship between the optimum model and optimum value of hyper-parameters, but other models were found to exhibit similar results with respect to fraction of negatives, although these results are excluded for brevity.

from 0.09 to 0.5, (these values correspond to ratios of positives: negatives with positives equal to 1 and negatives ranging from 0.1 to 1.0.). These show that as the fraction of negatives is increased from 0.09 to 0.5 accuracy shows a steady increase after an initial decrease. In contrast the other metrics: AUPRC, AUROC and F-score for $\beta = 1$ and 0 shown an initial increase and then a steady decrease. For the first two values of fraction of negative: 0.091 and 0.167 all metrics are within a small range and there is little variability. F-score for $\beta = 1$ has the lowest values (0.851 and 0.856) and accuracy the highest (0.898 and 0.909). Accuracy is the largest value for all values of fraction of negatives and F-score for $\beta = 1$ is the smallest value for all values of fraction of negatives. As the value of fraction of negatives increases the metrics diverge with accuracy increasing and the other four measures decreasing. As the value of fraction of negatives increases AUROC shows a nearly linear slight decrease whereas AUPRC and the F-scores for $\beta = 1$ and 0.5 show a more varied descent.

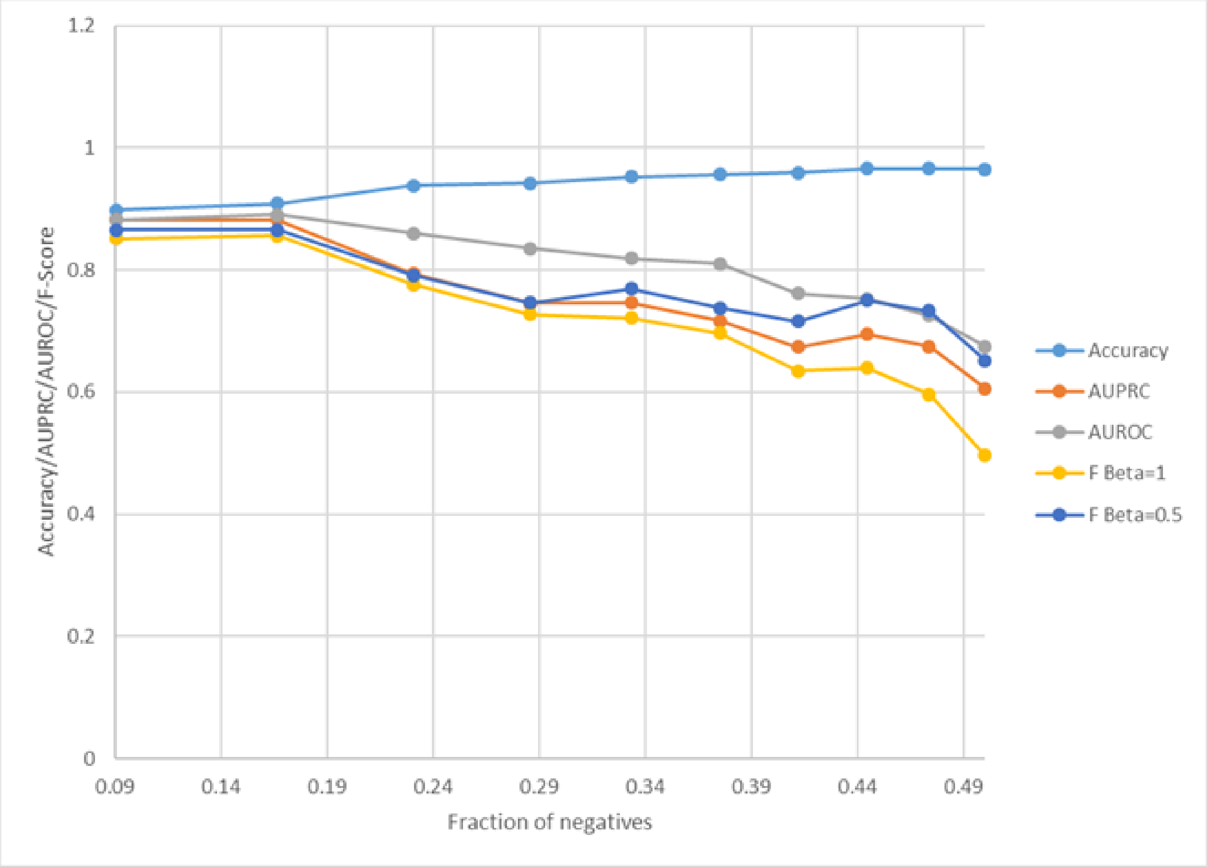


Figure 7.1: Plot of Accuracy, AUPRC, AUROC, F-score for $\beta = 1$ and 0 versus fraction of negatives from 0.09 to 0.5.

More fine grained investigation is given in Figure 7.2 and Table B.3, in which metrics, Accuracy, AUPRC, AUROC, and F-score for $\beta = 1$ and 0 are plotted against fraction of negatives from 0.009 to 0.231 (These values correspond to ratios of positives: negatives and positives equal to 1 and negatives ranging from 0.01 to 0.3.). This shows that as the oversampling ratio is increased from 0.009 to 0.3 accuracy increases steadily after an initial drop. In contrast the other metrics: AUPRC, AUROC and F-score for $\beta = 1$ and 0 shown an initial increase and then a steady decrease.

Although accuracy is important for a classifier the other four metrics are often regarded by the literature as more important (Guyon and Elisseeff, 2003). Since in this instance accuracy does not vary a great deal (minimum of 0.896 to a maximum of 0.938) it seems prudent to select the negative fraction based more on the other three metrics. Hence the optimum value of fraction of negatives was chosen to maximise AUPRC, AUROC and F-scores whilst retaining a high value for accuracy. Hence the over sampling fraction of negatives of 0.0385 (equivalent to a ratio of positives to negatives of 1 to 0.4)

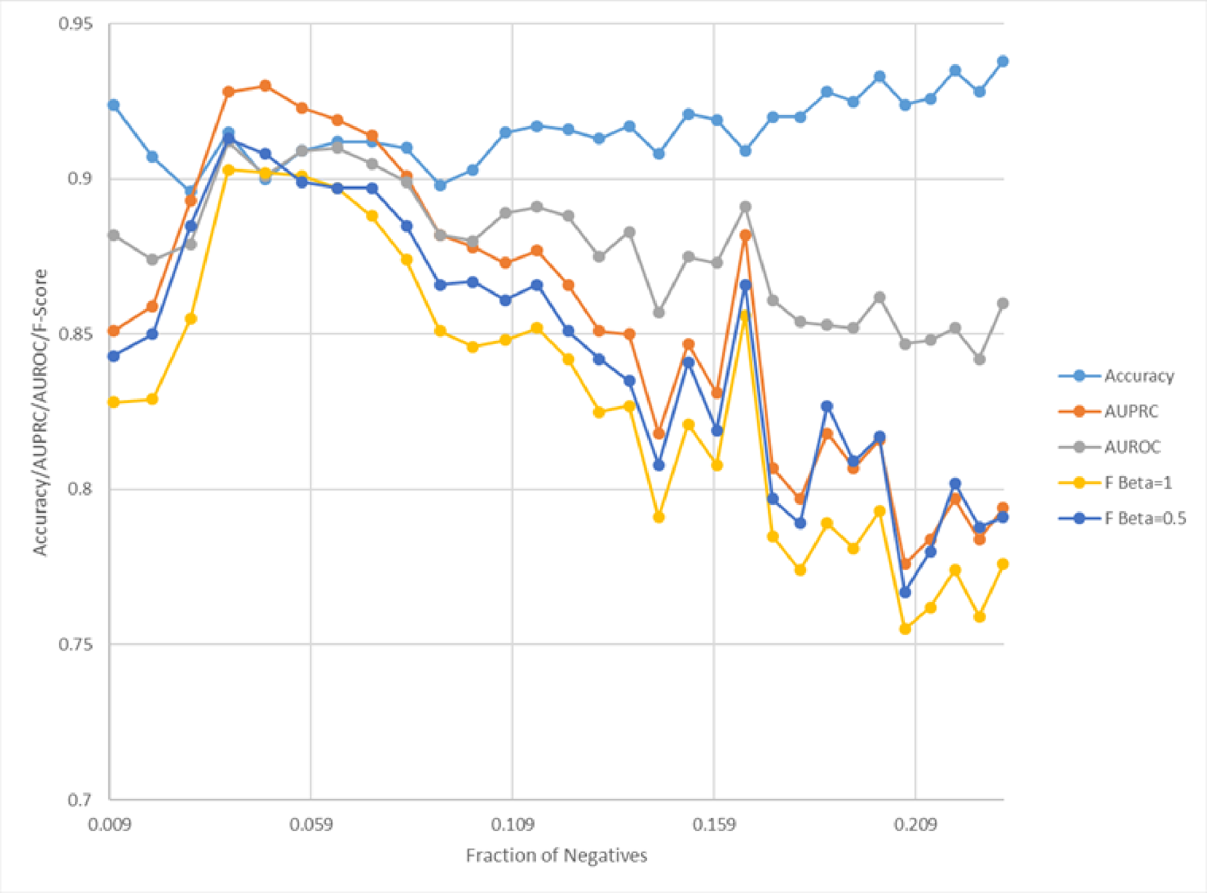


Figure 7.2: Plot of Accuracy, AUPRC, AUROC, F-score for $\beta = 1$ and 0 versus fraction of negatives 0.009 to 0.231.

was chosen. This value gives an AUPRC measure of 0.928 which is slightly lower than the highest value of 0.93, but it gives the maximum value of AUROC of 0.912, and F-scores for $\beta = 1.0$ and 0.5 of 0.903 and 0.913 respectively.

7.1.2 Text, User, Geographical and Temporal Features

Table 7.2: Metrics for different features with SVM and N5.

Add Features	**Acc**	**AUPRC**	**AUROC**	**F-Score**	
				$\beta = 1$	$\beta = 0.5$
PROFANE_TWEET	0.843	0.764	0.764	0.720	0.744
PRONOUN_TWEET	0.707	0.709	0.692	0.644	0.761
PROFANEWIN_TWEET	0.785	0.708	0.667	0.668	0.667
EMOTICONS_TWEET	0.655	0.698	0.680	0.685	0.683
CAPS_TWEET	0.606	0.699	0.645	0.637	0.64
RACIST_TWEET	0.778	0.704	0.714	0.702	0.709
LENGTH_TWEET	0.701	0.667	0.683	0.601	0.644
PROFANE_ACCOUNT	0.795	0.692	0.681	0.77	0.733
PRONOUN_ACCOUNT	0.695	0.658	0.688	0.673	0.678
EMOTICONS_ACCOUNT	0.70	0.666	0.671	0.606	0.645
CAPS_ACCOUNT	0.701	0.683	0.688	0.682	0.685
RACIST_ACCOUNT	0.757	0.728	0.722	0.719	0.711
LENGTH_ACCOUNT	0.758	0.757	0.736	0.	0.718
GENDER	0.821	0.764	0.788	0.724	0.727
GEO_TWEET	0.84	0.771	0.796	0.721	0.731
DAY_TWEET	0.887	0.777	0.831	0.736	0.746
HOUR_TWEET	0.925	0.864	0.870	0.816	0.860
BASELINE	0.925	0.859	0.884	0.835	0.854

Each of the features discussed in Section 6.4.3 were added to N5 and run with a SVM model with tenfold cross validation. The metrics are given in Table 7.2. The model was also run with no additional features and this was denoted by BASELINE in the table. As can be seen from the table only HOUR_TWEET was an improvement on the BASELINE. DAY_TWEET was the next best performing additional feature, and these two features were used in further analysis. The rest performed poorly and so were not used further.

It was disappointing that BASELINE was almost the most successful feature set, with most of the additional features reducing the performance significantly. There were not the improvements seen by and Dadvar et al. (2012) and Dadvar et al. (2013), which were the two works most influential in determining which features to experiment with. However in Dadvar et al. (ibid.) feature sets were compared against a baseline which was not BOW or Ngrams, but instead 'content-based features ', such as 'the number of profane words in the comment' and 'the normalised number of first and second person pronouns in the comment'. So it might be the case that there are other feature sets perform well in relation to this baseline, since it only includes a subset of the available information for analysis. This raises the question of why they did not use all the textual content as a baseline, and illustrates, the possibility at least, of how the efficacy of features can be distorted by the selection of what they are compared against. For this research the baseline was N5, which includes all of the text in a tweet, meaning that any feature efficacy comparisons are valid.

7.1.3 Ngrams vs BOW and Hour and Day

Table 7.3[4] shows metrics for an SVM model with different feature sets. The SVM model was run a total of 12 times. There were four different word features: BOW, Bi, Tri and N5 corresponding to bag of words features, bigram features, trigram features and Ngrams from 1 to 5 combined. For each of the different word features, the model was run three times with the following additional features: hour and day as additional features, hour as an additional feature and neither additional feature. For each combination there are five metrics given in the table. The metrics were: accuracy (Acc), AUPRC, AUROC and F-score for $\beta = 0.5$ and 1. The lowest accuracy of 0.761 was for trigrams with hour and the best with 0.925 was for N5 for both hour and no additional features. AUPRC ranged from 0.499 (performing worse than chance!), for trigrams and hour and day, to a maximum value of 0.864 for N5 with hour.

[4] Results in this table and Table 7.4 are the aggregate of 10 runs for each. The value reported is the mean value over these 10 runs, standard deviation is not reported to simplify the table. Standard deviations were low: in the range 0.01 to 0.05, and they were all fairly similar between and within models.

It can be seen from the table that the best combination overall was N5 with either hour or neither as an additional feature. For these two feature sets all of their metrics were better than any of the other feature sets. For these two feature sets the accuracy was the same with 0.925, and for AUPRC the hour feature was slightly better: 0.864 versus 0.859. For AUROC the situation was reversed with values 0.870 for hour and 0.884 without it. For F-score with $\beta = 1$ the hour feature performed worse with 0.816 versus 0.835 for no additional features. For F-score with $\beta = 0.5$ again the reverse was true, with 0.860 for hour and 0.854 without it.

If additional features are examined, Trigram features performed the worst, particularly in terms of F-score for which all their values were below 0.5, compared with all the other features having F-scores above 0.5, apart from bigrams with hour which had a value of 0.498. The relationship between the addition of hour and hour and day and type of word feature is complex. Overall for BOW hour outperforms hour and day and neither, for bigrams neither is best and this is also the case for trigrams. For N5 the addition of hour performs best. The addition of day and hour together as features was detrimental for all of the metrics for the best performing feature sets. It was beneficial for the BOW metrics: accuracy, AUPRC and F-score with $\beta = 0.5$, when compared with either hour or neither additional features. It was also beneficial for all of the metrics for bigrams when compared with hour as an additional feature, but was detrimental for all metrics with no additional features. Its only other positive effect was a slight improvement for

Table 7.3: Metrics for SVM for BOW, bigrams, trigrams and N5.

Model	**Add Features**	**Acc**	**AUPRC**	**AUROC**	**F-Score**	
					$\boldsymbol{\beta = 1}$	$\boldsymbol{\beta = 0.5}$
SVM	BOW+Hour+Day	0.885	0.781	0.796	0.721	0.791
SVM	BOW+Hour	0.884	0.771	0.808	0.726	0.768
SVM	BOW	0.880	0.767	0.821	0.733	0.747
SVM	Bi+Hour+Day	0.793	0.585	0.679	0.515	0.552
SVM	Bi+Hour	0.787	0.561	0.674	0.498	0.519
SVM	Bi	0.800	0.610	0.699	0.547	0.575
SVM	Tri+Hour+Day	0.762	0.499	0.629	0.424	0.453
SVM	Tri+Hour	0.761	0.516	0.631	0.434	0.472
SVM	Tri	0.769	0.504	0.641	0.438	0.455
SVM	N5+Hour+Day	0.882	0.771	0.827	0.740	0.748
SVM	N5+Hour	0.925	0.864	0.870	0.816	0.860
SVM	N5	0.925	0.859	0.884	0.835	0.854

trigrams compared with hour as an additional feature: 0.762 versus 0.761. The benefits or lack thereof of the addition of hour on its own is a complex picture. As a feature for BOW for accuracy, AUPRC and F-score with $\beta = 0.5$ it was beneficial when compared with just BOW, and detrimental when compared with the addition of hour and day. For AUROC and F-score with $\beta = 1$ the reverse was true, that is it had a negative effect when compared with just BOW but a positive effect when compared with the addition of hour and day. For bigrams hour as an additional feature was detrimental for all metrics in all cases. For trigrams for accuracy it was detrimental in both cases. AUPRC however had a positive effect in both cases as it did for F-score with $\beta = 0.5$. For AUROC and F-score with $\beta = 1$ it was positive compared to both hour and day as additional features and negative when compared with no additional features. For N5 for accuracy it was an improvement on both hour and day as additional features but was equally as effective as no additional features. For AUPRC and F-score with $\beta = 0.5$ it was beneficial in both cases. For AUROC and F-score with $\beta = 1$, the same effects was seen as for BOW, bigrams and trigrams, that is it was positive compared to both hour and day as additional features and negative when compared with no additional features.

As mentioned in chapter 2 the results of the performance of other researchers' machine learning systems with respect to Ngrams versus BOW were mixed. In the literature tokenisation of the text consisted of treating it as BOW or Ngrams (bigrams, trigrams or 5grams). Of the researchers that compared these, the majority found that 5grams worked best, although this was not unanimous. So while this research agreed that 5grams worked best, this should not be taken as carte blanche to use only 5grams, since in some research its use has not been as beneficial as the use of BOW.

As a result when algorithms were compared they were run with N5, N5 plus hour and N5 plus hour and day as their input feature sets. The comparison of algorithms using these feature sets is given in the next section.

Table 7.4: Metrics for the seven algorithms with N5.

Model	**Add Features**	**Acc**	**AUPRC**	**AUROC**	**F-Score**	
					$\boldsymbol{\beta = 1}$	$\boldsymbol{\beta = 0.5}$
NB	Hour+Day	0.848	0.691	0.713	0.584	0.695
NB	Hour	0.850	0.692	0.712	0.582	0.697
NB		0.848	0.689	0.723	0.599	0.691
LR	Hour+Day	0.857	0.571	0.500	0.250	0.172
LR	Hour	0.868	0.566	0.500	0.233	0.160
LR		0.865	0.567	0.500	0.238	0.163
SVM	Hour+Day	0.882	0.771	0.827	0.740	0.748
SVM	Hour	0.925	0.864	0.870	0.816	0.860
SVM		0.925	0.859	0.884	0.835	0.854
RF	Hour+Day	0.781	0.607	0.554	0.200	0.374
RF	Hour	0.791	0.571	0.555	0.206	0.374
RF		0.789	0.639	0.555	0.198	0.378
DT	Hour+Day	0.782	0.625	0.539	0.145	0.295
DT	Hour	0.790	0.643	0.538	0.141	0.292
DT		0.789	0.650	0.545	0.165	0.330
GBT	Hour+Day	0.789	0.512	0.600	0.352	0.469
GBT	Hour	0.794	0.535	0.599	0.347	0.484
GBT		0.744	0.474	0.605	0.388	0.423
ANN	Hour+Day	0.779	0.593	0.537	0.143	0.288
ANN	Hour	0.775	0.647	0.535	0.130	0.272
ANN		0.769	0.410	0.529	0.141	0.251

7.1.4 Algorithms

The algorithms were tested with three different feature sets as input: text, text+hour and text+hour+day, giving 7*3 = 21 different combinations of algorithms, features and preprocessing. Each of these algorithm/feature/preprocessing combinations were run and their metrics are given in Table 7.4. Table 7.4 shows the results for seven different algorithms: NB, LR, SVM, RF, DT, GBT and ANN (N5 was used as the word feature for each). These models were optimised using K-fold cross validation as discussed in Section 5.12.2.7 with K=10.[5] For each of the different models, the model was cross validates three times, that is once with each of the following additional features: hour and day as additional features, hour as an additional feature and neither additional feature. For each combination there are five metrics given in the table. The metrics are: accuracy, AUPRC, AUROC and F-score for β = 0.5 and 1. The model that performed best overall

[5] Tenfold cross validation was used since it has been shown to work well when selecting a classifier from a set of classifiers (Kohavi, 1995).

was SVM, which had the maximum value for all of the metrics. The second best model overall was NB; it had the second best overall AUPRC, AUROC, and F-scores for both $\beta = 1$ and $\beta = 0.5$, although LR outperformed it in accuracy. These three models were the only three with accuracy above 0.8. The accuracies for RF, DT and GBT were very similar. Compared to RF and DT, GBT gave slightly higher accuracies for the addition of hour and day and just hour, but slightly worse for the addition of neither. ANN was the worst performing model in terms of accuracy.

For AUPRC SVM was best, followed by NB although NB's AUPRC was considerably lower than for SVM. DT was next best in terms of AUPRC and GBT's AUPRC overall was the lowest, except for ANN with no additional features which had the value 0.410, meaning it performed worse than would be expected by chance. For AUROC SVM was again the best, and again NB was a distant second, although NB's values were all above 0.7. The third best performer, in terms of AUROC, was GBT, and the worst was LR with all values of 0.5.

For F-score with $\beta = 1$ once more SVM was the best and NB second best, albeit considerably worse than SVM. All the other models performed poorly with regard to this metric, in particular DT with values between 0.141 and 0.165 and ANN with values between 0.13 and 0.143.

Finally, for F-score with $\beta = 0.5$ it was the same story of SVM being best and NB being a distant second. All the other models scored less than 0.5 for this metric, GBT being the best of these with values between 0.423 and 0.484, and LR the worst with values ranging from 0.16 to 0.172.

This illustrates the overall efficacy of the different models, now the effect of the addition of hour and day and just hour as additional features will be considered in detail. For the NB model accuracy is best with the addition of hour, this is also true for AUPRC, although it is only a slight improvement in both cases. For NB AUROC is best with no additional features and this is also true for F-score with $\beta = 1$ but for F-score with $\beta = 0.5$ hour as an additional feature performs best.

For the LR model hour as an additional feature is best for accuracy but for AUPRC, and F-scores with both $\beta = 1$ and $\beta = 0.5$ the hour and day combination of additional features works best, albeit still very poorly for the latter two. For LR all the AUROC values are the same, 0.5.

For the SVM model hour as an additional feature is best for accuracy, for F-score with $\beta = 0.5$ and AUPRC. For the other metrics, AUROC and F-score with $\beta = 1$ no additional features is the best combination.

For RF the addition of hour as a feature works best for accuracy and F-score with $\beta = 1$. For this model the addition of no additional features gives the best result for AUPRC and F-score with $\beta = 0.5$. For this same model AUROC is virtually the same for all three feature sets.

For the DT model hour as an additional feature is best for accuracy, but no additional features is best for the other four metrics. For the GBT model hour as an additional feature was best for accuracy, AUPRC and F-score with $\beta = 0.5$. Similarly to the SVM model, for the other metrics, AUROC and F-score with $\beta = 1$, no additional features is the best combination.

For the ANN model the combination of hour and day as additional feature worked best for accuracy, AUROC and F-scores with both $\beta = 1$ and $\beta = 0.5$. Its other metric, AUPRC, was maximised with hour as an additional feature.

From these results it can be seen that SVM performed well not just in terms of accuracy, but also precision and recall. This means it is doing a good job of predicting true positives and true negatives correctly compared with all predictions (its accuracy is high), it is doing a good job of predicting true positives compared with all those that it classes as positive (its precision is high) and it is doing a good job of predicting true positives compared with all those that are actually positive (its recall is high).

NB also had reasonable accuracy, precision and recall, but not as good as SVM. LR had good accuracy but poor precision and recall, and the other models: RF, DT, GBT and ANN all had reasonable accuracy but poor precision and recall.

SVM is clearly the best model in this situation and so was used in the prediction stage. The other decision which needed to be made for the decision stage, is which of the sets of additional features should be used, either none, or the addition of hour on its own, or the addition of hour and day. For SVM this decision is not straightforward other than the rejection of hour and day, which does not perform optimally for any of the metrics. For the other two feature sets accuracy is identical, and the addition of hour is optimal for AUPRC and F-score with $\beta = 0.5$, whereas no additional features is optimal for AUROC and F-score with $\beta = 1$. For class imbalance situations AUPRC is regarded as a more reliable indicator of a model's performance than AUROC (Davis and Goadrich, 2006). The tweet data is an imbalanced class situation, since it is highly skewed in terms of negative i.e. nonracist tweets, so in order to distinguish between the two feature sets, the efficacy of AUPRC for imbalanced data became the tiebreaker, and it was decided to use the feature set with the addition of hour, since that gave the optimal value for AUPRC for the SVM model.

The efficacy of SVM was broadly in line with the existing literature. As mentioned in Chapter 2. The comparison of algorithms in previous research had given mixed results, some researchers finding SVM to be the optimum algorithm, whereas others found NB worked better. Those that used ANN, generally found that these worked well, although they were not necessarily compared with SVM. For this research SVM was significantly better than NB and also better than ANN. This is perhaps unsurprising since SVM is often noted in the literature as performing very well for text classification tasks, whereas ANNs are more often related to tasks such as visual recognition.

7.2 Temporal Results

The best performing model, that of SVM with hour as an additional feature, was run on the D1 dataset minus the hand annotated sample. This generated a predicted values dataset which was analysed as two separate datasets: one containing predicted racist tweets (denoted as Pred R) and the other containing predicted non-racist tweets (denoted as Pred NR). The graphs in Figure 7.3, Figure 7.4, Figure 7.5 and Figure 7.6, are plots of tweets by hour of day as percentage of total for that day, split by day, for Inp NR, Inp R, Pred NR, and Pred R respectively. Both the Inp NR and Inp R data show much more variability than Pred NR and Pred R, especially Inp R. This is to be expected because of their much smaller sample size. Inp R in particular shows a lot of variability both between hours and between days.

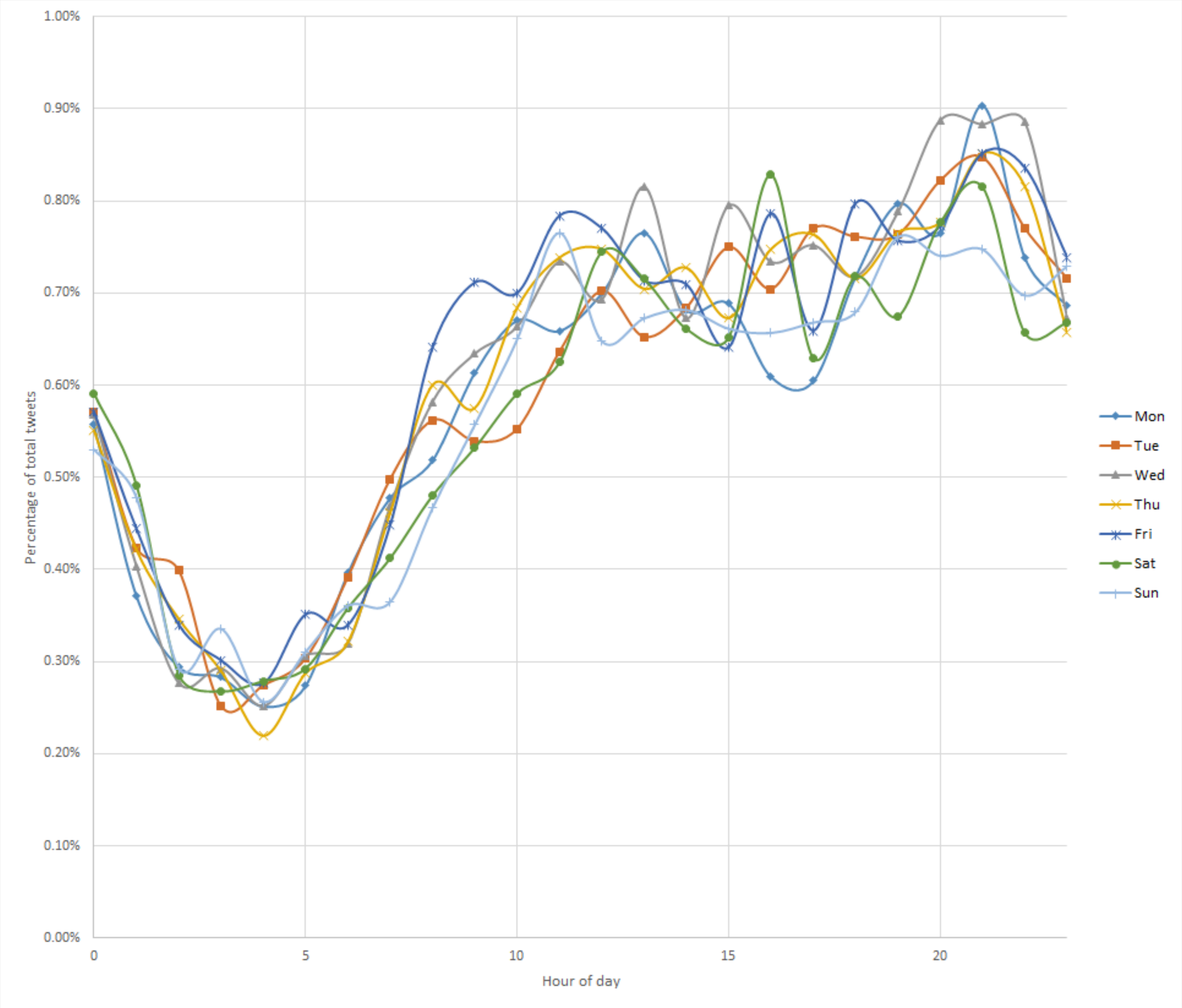


Figure 7.3: Tweets by hour of day as percentage of total for Inp NR.

The trend with respect to hourly percentages of tweets these graphs generally follow is for the number of tweets to decrease from around midnight to 4 or 5 in the morning, and then gradually increase to around 10 AM. The period from 10 AM to approximately

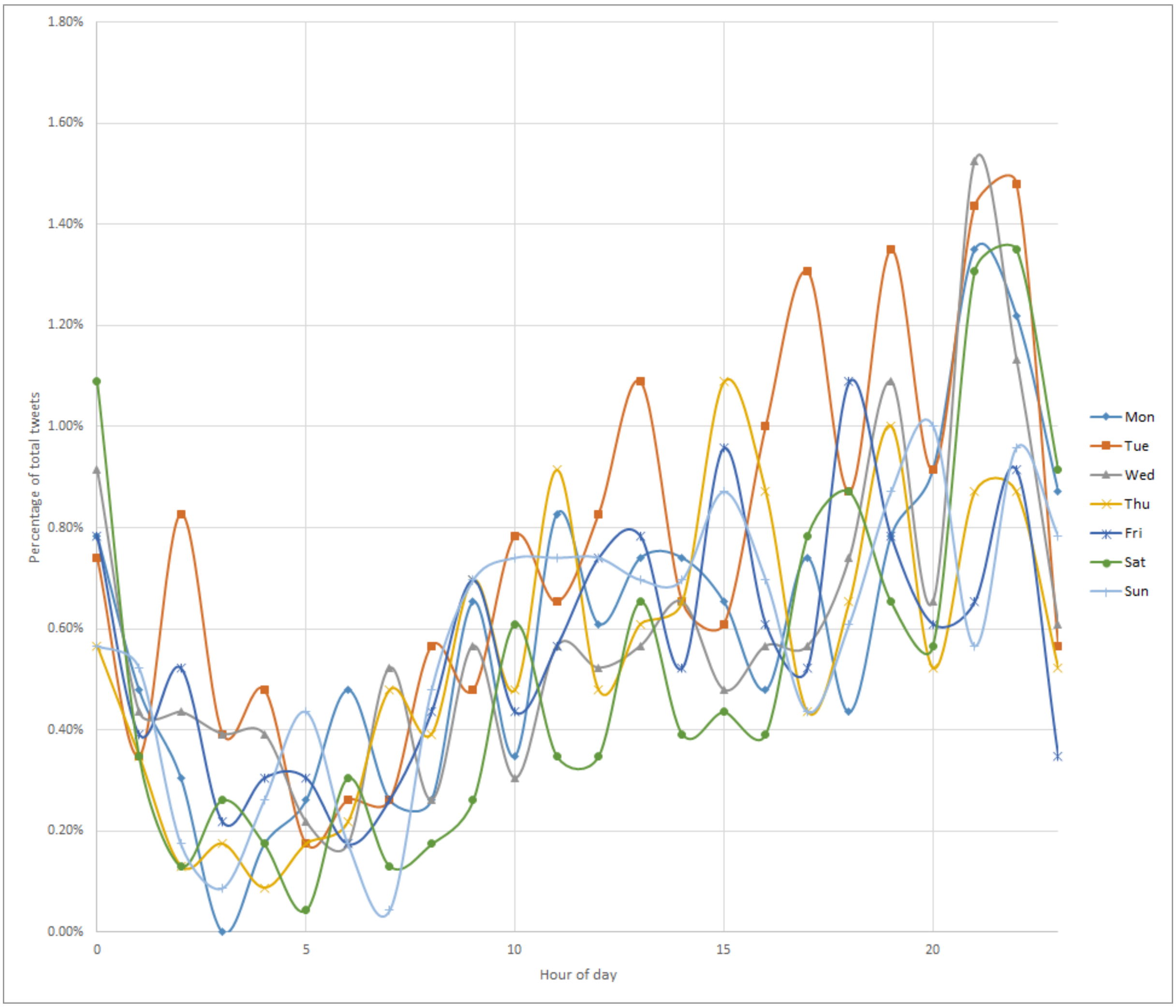


Figure 7.4: Tweets by hour of day as percentage of total for Inp R.

8 AM is more variable for both Inp NR and Inp R; there is a lot of variability and any trend is unclear. For Pred NR there is a slight downward trend until 4 PM, and for Pred R there is an upward trend until 4 PM and then a slight reduction to 5 PM. For each of these four graphs there is then an increase to a peak of 9 or 10 PM followed by reduction towards the early-morning minimum.

With respect to hourly percentages of tweets *by day*, the data from both Inp NR and Inp R is inconclusive, except that surprisingly for Inp R Friday is the day with the lowest percentage of tweets, and its late-night peak is also less pronounced than the other days except Sunday and Thursday. For Pred NR Saturday and Sunday have the fewest tweets for much of the day and their late-night peaks are less pronounced than the other days. The other days of the week show reasonably similar patterns throughout the day, although

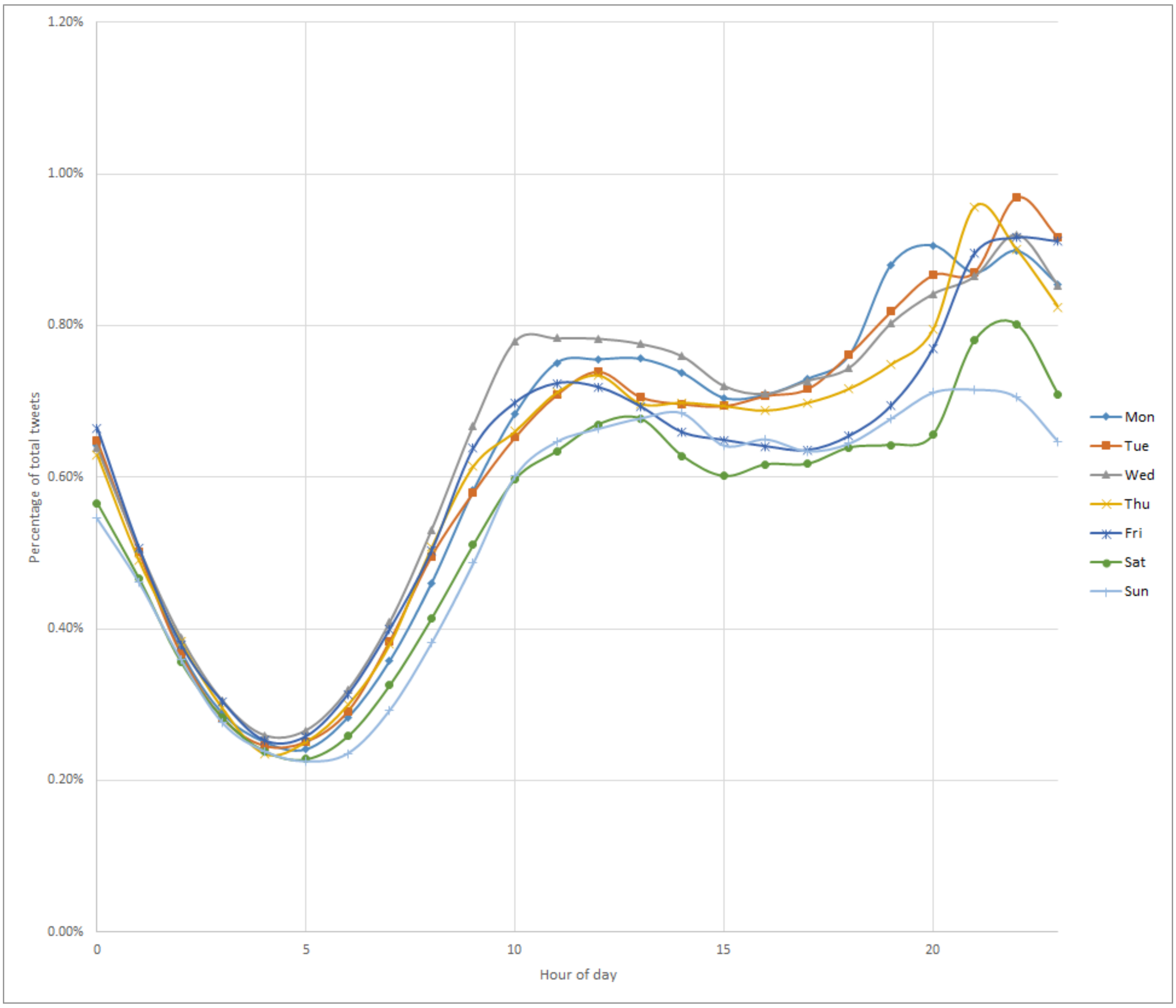


Figure 7.5: Tweets by hour of day as percentage of total for Pred NR.

Friday again, perhaps surprisingly, shows lower percentages from approximately midday until 8 PM, thereafter it increases comparably with Monday through Thursday. For Pred R Saturday and Sunday also have the fewest tweets, with Friday the most between 9 PM and 10 PM. For Pred NR most days show a decrease from 10 or 11 AM until around three or 4 PM then a steady increase until 8 or 9 PM, followed by a decrease during the early morning hours until 5 AM. In contrast with this Pred R most days show a steady increase from 5 AM until 4 PM, then a drop at 5 PM, followed by a sharp increase until 9 or 10 PM, followed by a similar decrease until 5 AM.

Table 7.5 tabulates the maxima and minima of percentage tweets throughout the week by dataset. This shows that the minimum percentage of tweets occurs on Sunday except for in Inp R when it occurs on Tuesday, the maximum percentage of tweets is more

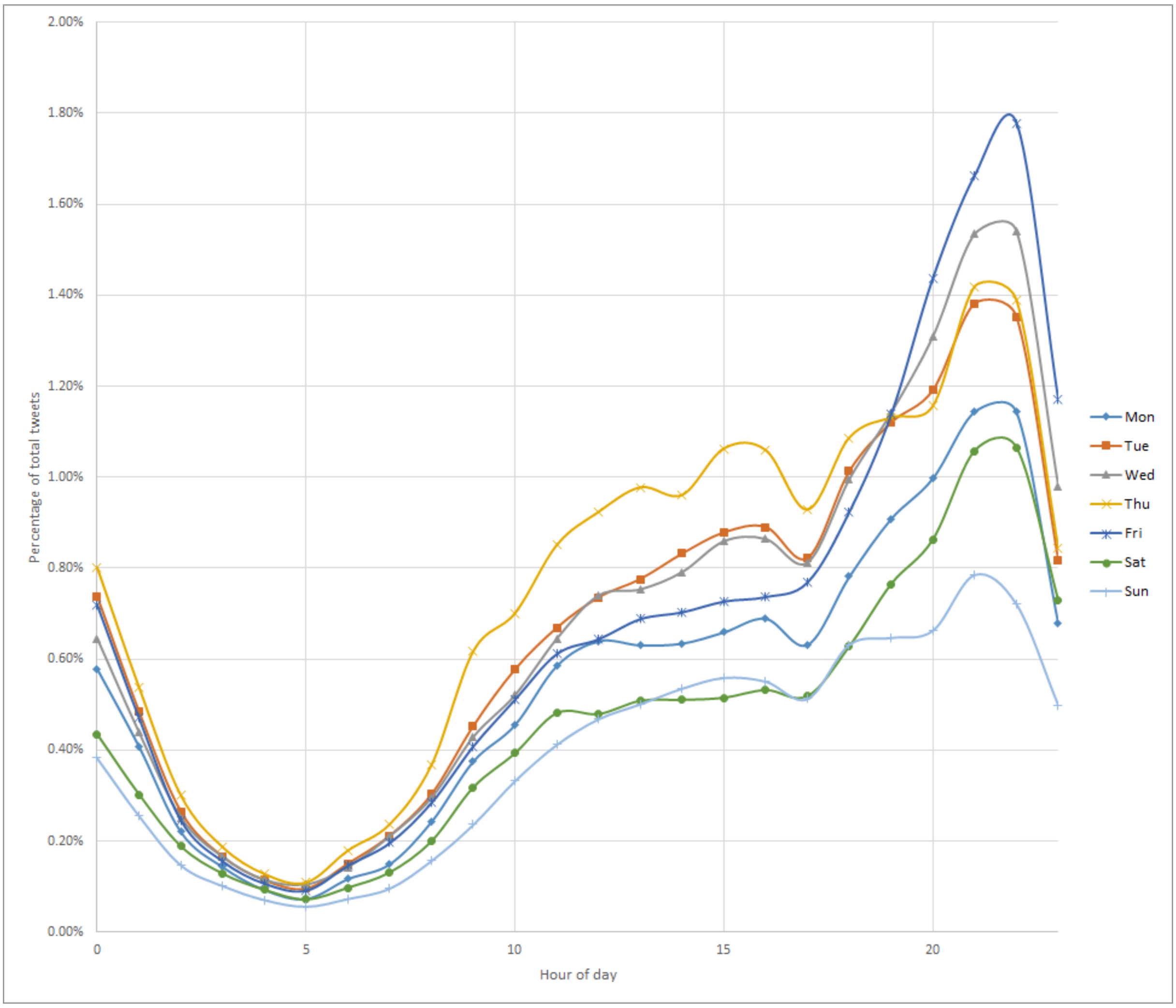


Figure 7.6: Tweets by hour of day as percentage of total for Pred R.

varied occurring on Friday, Saturday, Wednesday and Thursday for Inp NR, Inp R, Pred NR, and Pred R respectively.

Table 7.5: Maxima and minima of percentage tweets throughout the week by dataset.

	Min %	**Min Day**	**Max %**	**Max Day**
Data				
Inp NR	13.71	Sunday	14.94	Friday
Inp R	18.07	Tuesday	12.54	Saturday
Pred NR	12.8	Sunday	15.35	Wednesday
Pred R	9.4	Sunday	17.96	Thursday

Table 7.6 tabulates the the least and most active hours by percentage of tweets throughout the week by dataset. It can be seen that for Inp NR the quietest hour is 3 or 4 AM throughout the week, for Inp R the quietest hour is more variable being anywhere from 3 to 7 AM, for Pred NR it is always 4 or 5 AM and for Pred R it is always 5

AM. For Inp NR the busiest hour is 9 PM for Monday, Tuesday and Thursday, 8 PM for Wednesday, 4 PM for Saturday and 11 AM on Sunday. For Inp R there is also variability Monday, Tuesday, Wednesday, Saturday and Sunday are busiest during the 'normal peak' hours of 8 PM through 10 PM, but Thursday is busiest at 3 PM and Friday busiest at 6 PM. For Pred NR and Pred R the busiest hour is always 9 or 10 PM.

Table 7.6: Maxima and minima of percentage tweets by hour throughout the week by dataset.

	Monday		Tuesday		Wednesday		Thursday		Friday		Saturday		Sunday	
Data	**Min**	**Max**	**Min**	**Max**	**Min**	**Max**	**Min**	**Max**	**Min**	**Max**	**Min**	**Max**	**Min**	**Max**
Inp NR	4	21	3	21	4	20	4	21	4	21	3	16	4	11
Inp R	3	21	4	22	6	21	4	15	6	18	5	22	7	20
Pred NR	5	20	4	22	4	22	4	21	4	22	5	22	5	21
Pred R	5	21	5	21	5	22	5	21	5	22	5	22	5	21

Figure 7.7 shows the percentage of tweets by day of the week for each of the four datasets. It illustrates the reduction in tweeting over the weekend, and also the peaks of racist tweeting in Inp R on Tuesday and in Pred R on Thursday and Friday.

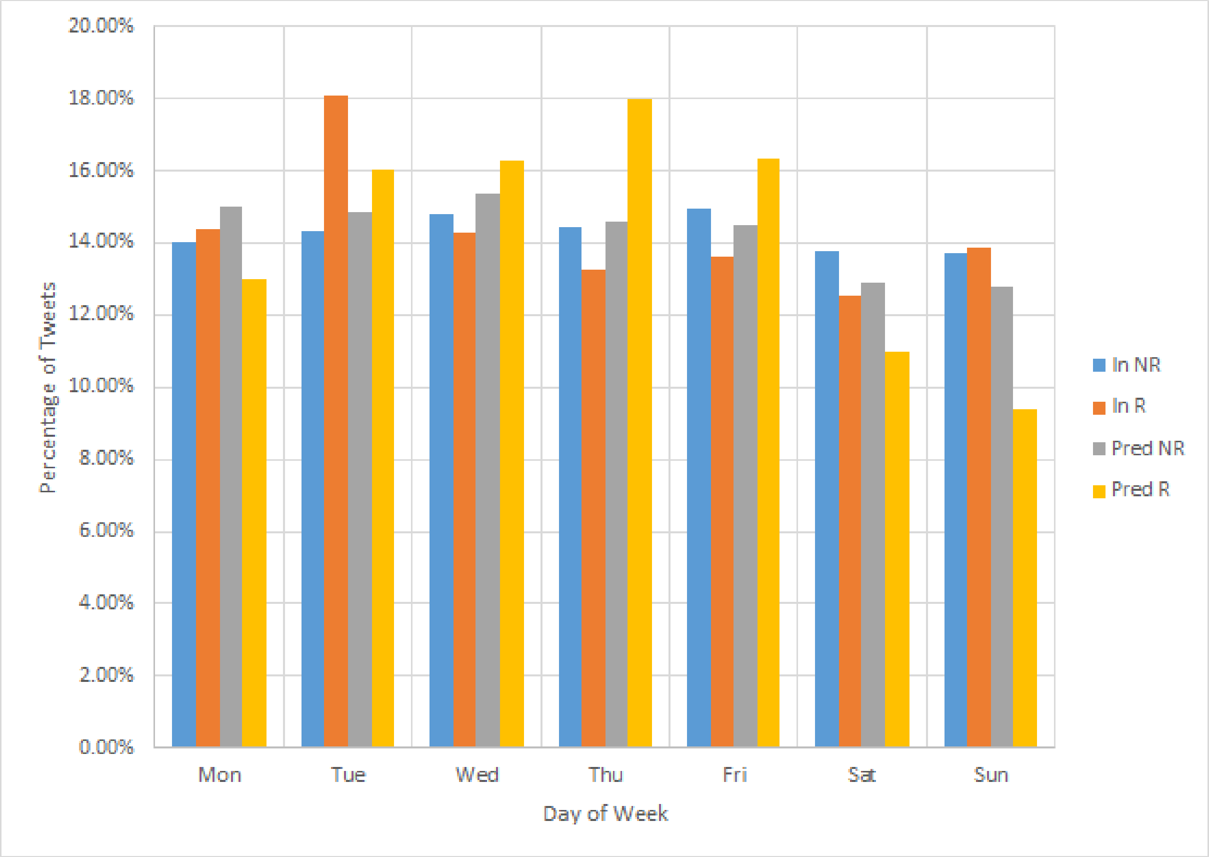


Figure 7.7: Tweets as percentage of total by day for the four datasets.

There is little in the literature to compare these temporal results with. Where temporal features are examined such as in Dadvar et al. (2013), they tend to be used as proxies for geographical information, the time zone of a tweet being used as an indication of the geographical location of the tweeter. This lack of attention in the literature and the results of this research indicate that temporal factors of tweets, and perhaps other social media, might be a fruitful area to explore further.

In the next section the results of the qualitative analysis are presented.

7.3 Qualitative results

7.3.1 NLTK Results

NLTK was used to produce the data in Table 7.7. This table shows word counts from the popular[6] racist tweets data. It gives the top 25 of each of the following: the most common words in the dataset, the most common words in the dataset excluding stopwords, the most common hashtags in the dataset, the most common bigrams in the dataset, and the most common trigrams in the dataset.

It can be seen from the table that the words 'nigger' and 'nigga' are the two most prevalent words, each being even more prevalent than the words 'a' and 'the'. These data show the expletive ridden nature of the tweets, with four swear words in the top 12 most common words, or top seven if stopwords are excluded. Indeed the vocabulary is extremely negative, of the top 25 words excluding stopwords, 14 are one of: ethnic slurs, swear words, negative adjectives and the words 'hate' and 'bitch'.

The top 25 hashtags show that very few hashtags were shared within this dataset. There were only four hashtags with counts greater than one: '#nigger', '#', '#hahn' and '#istandwithhatespeech'. The #istandwithhatespeech hashtag is interesting, since it arose in protest to the crackdown on hate speech and the agreement of a code of conduct by Microsoft, YouTube, Twitter and Facebook with the European Commission, to remove hate speech 24 hours after it was reported on their platforms (BBC, 2016). Use of the hashtag can be seen in the nonracist data, for example in the following tweet:

> Because censoring any type of speech, including hate speech is a slippery slope I'm not willing to go down #IStandWithHateSpeech.

This hashtag can also be seen in the racist data, examples being:

[6] These were the tweets in Pred R that had been retweeted more than once. Any of this sample that were determined to be non-racist by the author of this thesis were excluded.

Table 7.7: Top 25 most common, excluding stopwords, hashtags, bigrams and trigrams from the popular racist tweets.

Mst Cmn		**Ex Stop**		**Hashtags**		**Bigrams**		**Trigrams**	
nigga	2129	nigga	2129	#nigger	5	ass nigga	947	nigger nigger nigger	271
nigger	1835	nigger	1835	#	4	stupid nigger	451	bitch ass nigga	217
a	1471	ass	1174	#hahn	2	nigger nigger	347	alex stupid nigger	202
ass	1174	fuck	1097	#istandwithhatespeech	2	bitch ass	237	dumb ass nigga	52
fuck	1097	stupid	691	#lifelessonslearnedin5words	1	alex stupid	210	stupid ass nigga	43
you	842	fucking	589	#smh	1	fucking nigger	203	everyone type chat	41
stupid	691	bitch	348	#lookonthebrightsidein5words	1	fuck nigga	195	chat alex stupid	40
fucking	589	alex	268	#helpvex	1	nigga fuck	123	nigga shut fuck	35
the	587	shut	256	#commitacrimein5words	1	shut fuck	97	ass nigga keep	34
is	575	u	143	#bones	1	nigga shut	80	fake ass nigga	34
i	513	dumb	121	#yalldont	1	dumb ass	72	type chat alex	34
bitch	348	shit	115	#tweetlyricsthathavetobeshouted	1	fuck nigger	68	you'sbitch ass	31
up	339	get	107	#alllivesmatters	1	type chat	67	nigga even chase	29
that	312	type	105	#ps4share	1	stupid ass	62	fuck nigga fuck	29
this	291	chat	101	#nbafinals	1	everyone type	53	chase nigga even	29
my	279	im	98	#ifdawins	1	fuck fuck	49	even chase liquor	28
alex	268	cant	96	#bitchuguessedit	1	chat alex	42	stupid fucking nigger	28
shut	256	like	92	#gamergate	1	nigga fucking	42	fuck fuck nigga	27
me	252	yo	83	#lhhatl	1	ugly ass	41	ugly ass nigga	27
with	251	ugly	73	#police	1	fake ass	39	fuck ass nigga	25
to	226	fuckin	72	#uniteblue	1	stupid fucking	37	would chase nigga	24
in	212	chase	64	#whatihatein5words	1	give fuck	37	fuck bitch ass	24
and	184	faggot	63	#seoplus2016	1	nigga keep	36	everybody type chat	23
don't	179	hate	61	#zakirnaiks	1	fuck niggas	35	nigga give fuck	23
be	144	everyone	61	#future	1	can't fuck	33	nigga keep wildin	23

> As a nigger #IStandWithHateSpeech because I'm not a whiny nigger,

and

> #IStandWithHateSpeech nigger faggot tranny spic kike,

both of which may or may not have been sent with racist intent, but are certainly likely to cause offence.

'Alex' is the only name in the top 25 words, which is a reference to the YouTuber, Keemstar.[7]

All of the bigrams contain insults, slurs or swear words, or combinations of these, except for three of them: 'type chat', 'everyone type' and 'chat Alex', which are all also related to Keemstar. There is a similar situation for trigrams with the addition of one trigram which belongs to neither of these groups, 'even chase liquor' . This trigram is a reference to the lyrics of the song, Slayed, by Kodak Black, which was released in 2016, the time of the collection of the data, which contained the lyric, 'Why the fuck would I chase a bitch when I don't even chase the liquor.' There are multiple variations of this in the dataset, the most prevalent changing the word 'liquor' to 'nigga'.

These word and n-gram frequencies highlight the seasonal nature of the data. While much of the vocabulary such as the racial epithets and swear words will always have high frequencies in vocabularies of these kinds of tweets, other phrases frequencies will only be elevated for short periods, such as the mentions of Keemstar, whose notoriety was at its peak during summer 2016. The frequencies of the racial slurs was similar to that seen by Bartlett et al. (2014), in particular, the most frequently occurring being 'nigga' and the second most frequently 'nigger', with these two being the only slurs that were in the 20 most frequently occurring words in the sample.

[7] Keemstar, is alleged to have used the phrases 'Alex is a stupid nigger' and 'Everyone type in chat Alex is a stupid nigger'. Google trends searches peaked for these phrases during summer 2016, the time period during which the data for this research was collected (Spider-Byte, 2016).

7.3.2 Word Cloud Results

NVivo produce word clouds for different groupings of the Pred R data: the first level is exact matches only, the next level of grouping for NVivo is for words with the same stem, then synonyms, then specialisations and finally generalisations. The word clouds for each level of grouping are given in Figures 7.8, 7.9, 7.10, 7.11 and 7.12, respectively.

Figures 7.8 shows the prominence of the words ‘nigger’ and ‘nigga’, which appear as the two biggest words in the cloud, then the next level of prominence includes two words: ‘ass’ and ‘fuck’. Then two more words are prominent: ‘fucking’ and ‘stupid’.

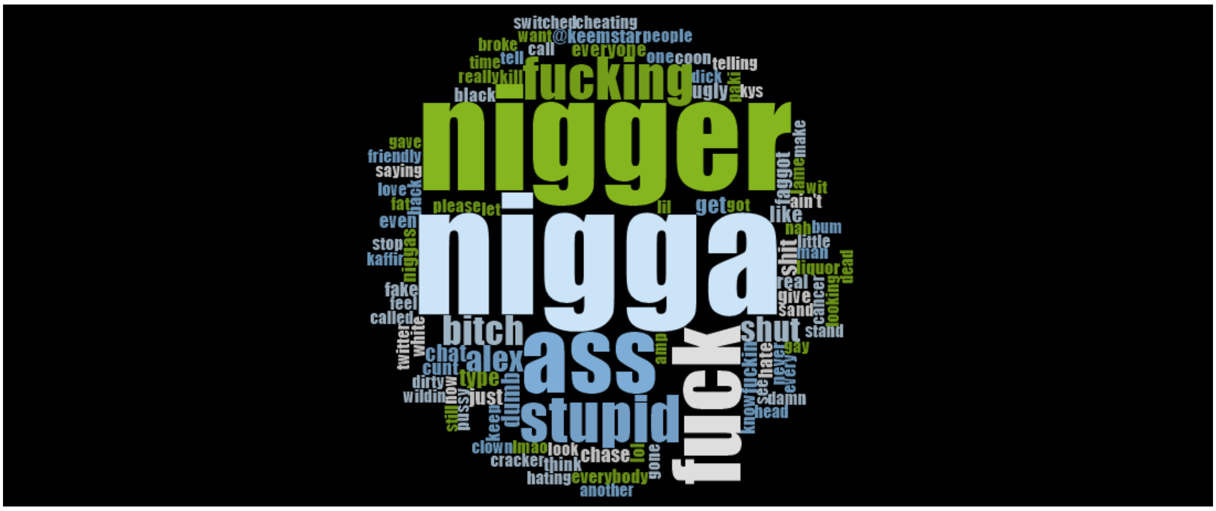


Figure 7.8: Word cloud of the popular racist tweets, exact matches only.

Figure 7.9 shows how the reduction of the set of words by stemming alters the word cloud. Now ‘niggers’ and ‘niggas’, are the most prominent, although less prominent than they were, ‘fucks’ becomes the third most prominent, followed by ‘ass’ and then ‘stupid’.

Figure 7.10 shows how the reduction of the set of words by grouping the words by synonyms alters the word cloud. In this diagram ‘nigga’ is by far the most prominent word, followed by ‘ass’ , then ‘stupid’, and ‘fuck’.

Figure 7.11 shows how the reduction of the set of words by grouping the words by specialisations alters the word cloud. Now the word ‘black’ has equal prominence to

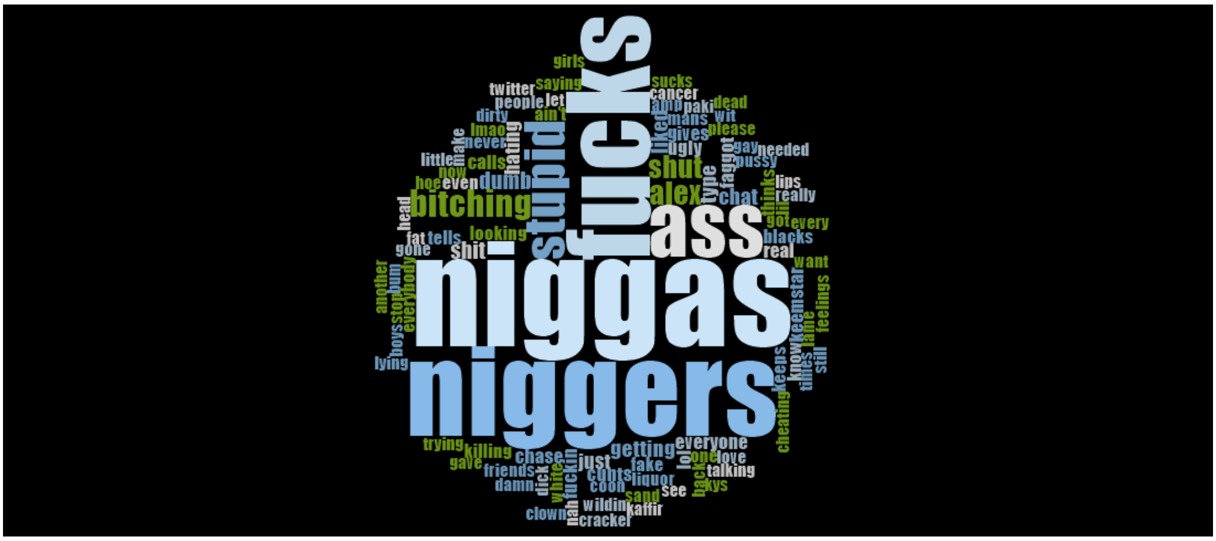


Figure 7.9: Word cloud of the popular racist tweets, words with the same stem.

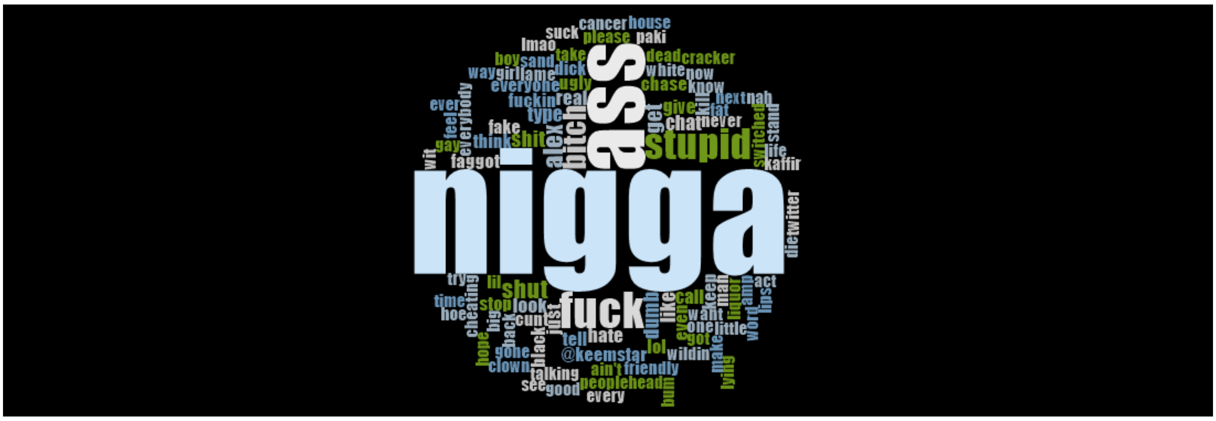


Figure 7.10: Word cloud of the popular racist tweets, synonyms.

‘nigga’, followed by ‘ass’, ‘sex’ and ‘relate’.

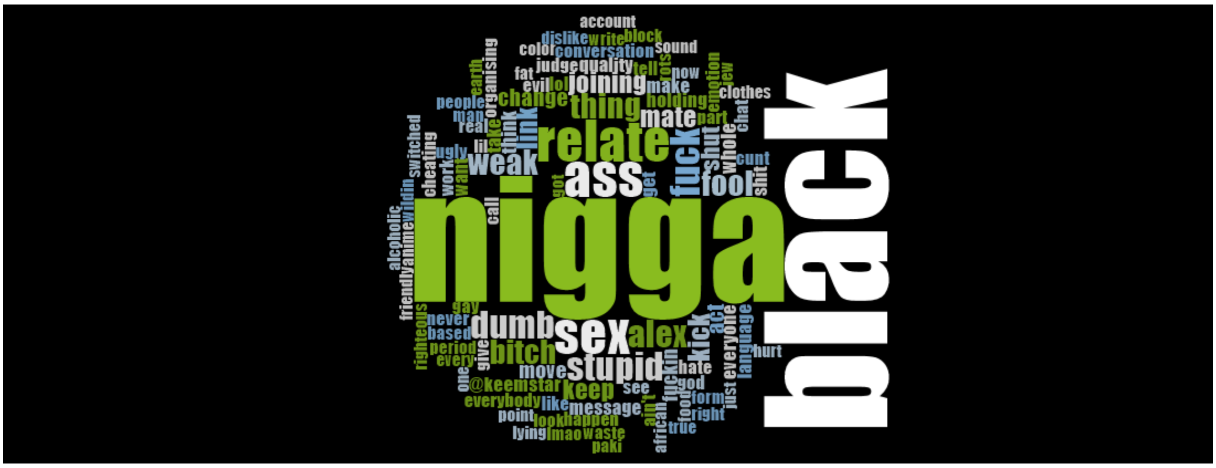


Figure 7.11: Word cloud of the popular racist tweets, specialisations.

Figure 7.12 shows how the reduction of the set of words by grouping the words by generalisations alters the word cloud. It is fairly similar to the specialisations word cloud. All these word clouds provide a good illustration of the prevalence of messages related to black ethnic slurs in the tweets, and the preponderance of swearing and other hateful words. The synonym word cloud in particular is quite striking, showing the dominance of the synonym ‘nigga’.

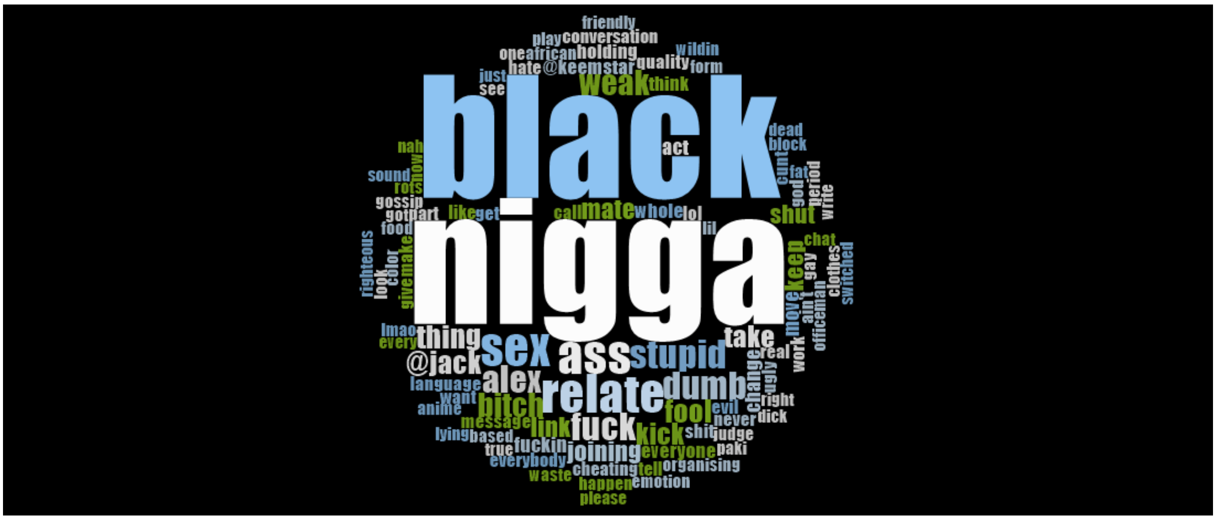


Figure 7.12: Word cloud of the popular racist tweets, generalisations.

7.3.3 Word Tree Results

NVivo produced the word trees given in Figures 7.13, 7.14, 7.15, 7.16, 7.17 and 7.18. For the more prevalent words, word trees become unwieldy, so word trees for ‘nigga’ and ‘nigger’ are not reproduced in this thesis. Instead Figures 7.13, 7.14, 7.15, 7.16, 7.17 and 7.18 give word trees for ‘gay’ at the level of stemming, ‘gay’ at the level of generalisations, ‘kike’ at the level of stemming, ‘kike’ at the level of generalisations, ‘kys’ at the level of generalisations, and ‘kill’ at the level of generalisations, respectively.

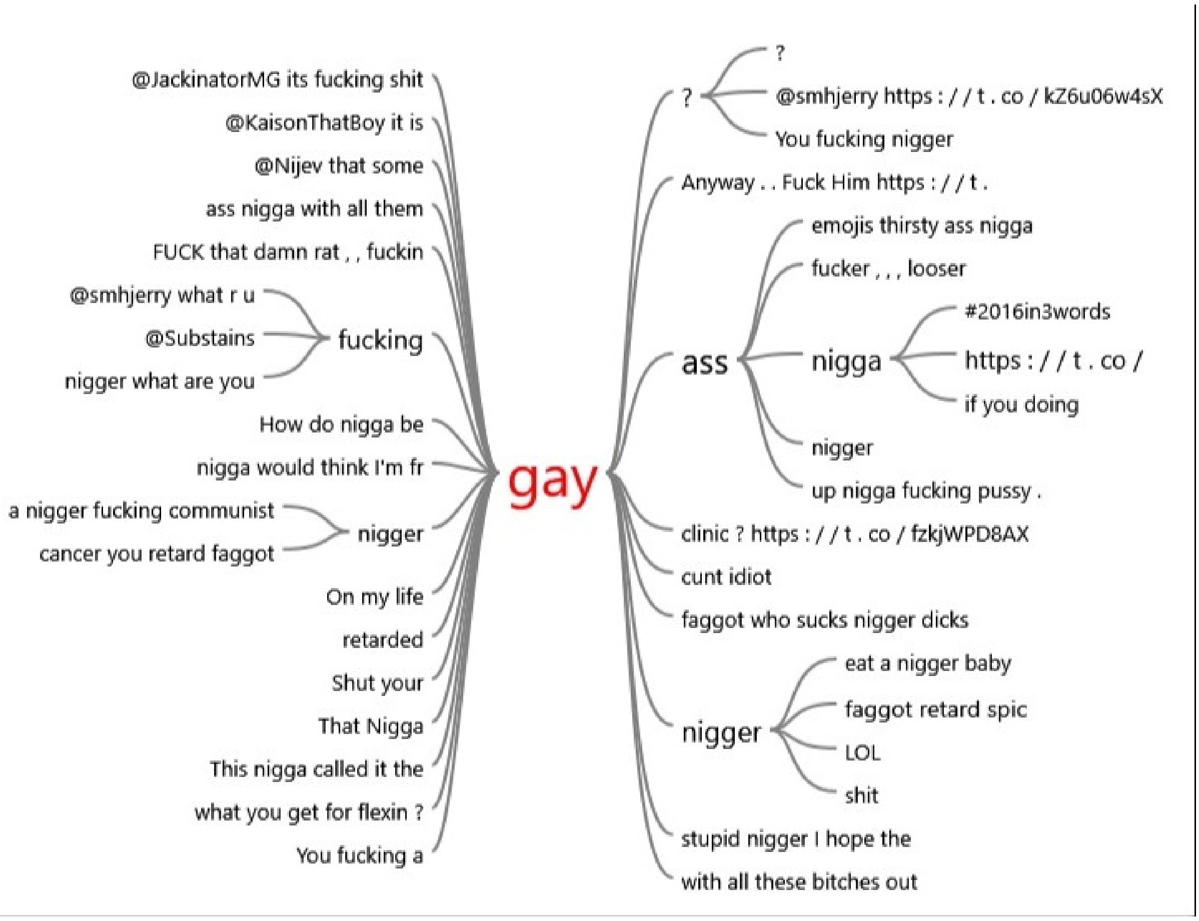


Figure 7.13: Word tree for ‘gay’ at the level of stemming.

Figure 7.13 displays the word tree for ‘gay’ at the level of stemming, which is identical to the word tree for ‘gay’[8] at the level of exact matches. The branches display the relentlessly negative use of the word. Interestingly its use as an adjective preceding the word ‘nigga’ is always with the addition of the word ‘ass’, whereas its use as an adjective preceding the word ‘nigger’ is more varied. When the word ‘gay’ is presented in a word tree at the level of generalisations (Figure 7.14), NVivo changes ‘gay’ to ‘faggot’, and there are far more complex branches both before and after the word, again all are relentlessly negative.

[8] When the word ’gay’ is used in the tweets it is always used in a pejorative sense.

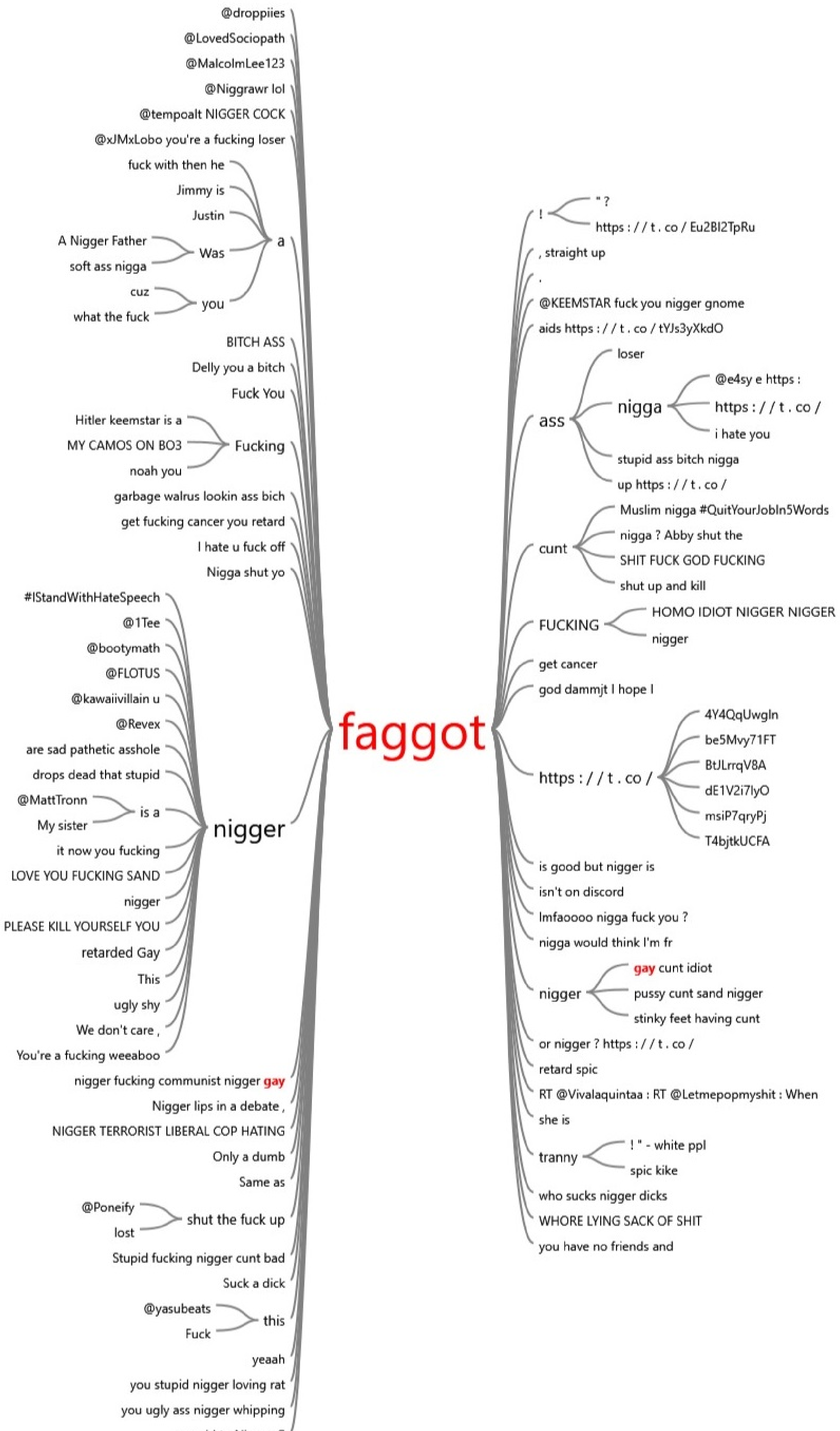
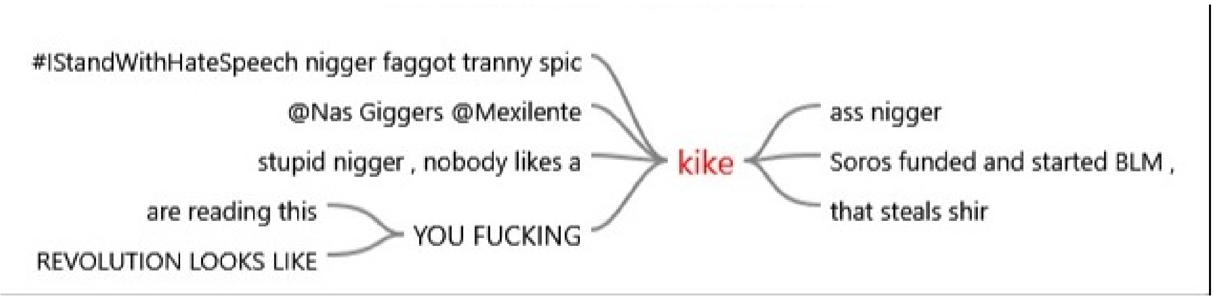

faggot
@droppiies
@LovedSociopath
@MalcolmLee123
@Niggrawr lol
@tempoalt NIGGER COCK
@xJMxLobo you're a fucking loser
fuck with then he
Jimmy is
Justin
A Nigger Father
soft ass nigga
Was
cuz
what the fuck
you
a
BITCH ASS
Delly you a bitch
Fuck You
Hitler keemstar is a
MY CAMOS ON BO3
noah you
Fucking
garbage walrus lookin ass bich
get fucking cancer you retard
I hate u fuck off
Nigga shut yo
#IStandWithHateSpeech
@1Tee
@bootymath
@FLOTUS
@kawaiivillain u
@Revex
are sad pathetic asshole
drops dead that stupid
@MattTronn
My sister
is a
it now you fucking
LOVE YOU FUCKING SAND
nigger
PLEASE KILL YOURSELF YOU
retarded Gay
This
ugly shy
We don't care ,
You're a fucking weeaboo
nigger
nigger fucking communist nigger gay
Nigger lips in a debate ,
NIGGER TERRORIST LIBERAL COP HATING
Only a dumb
Same as
@Poneify
lost
shut the fuck up
Stupid fucking nigger cunt bad
Suck a dick
@yasubeats
Fuck
this
yeaah
you stupid nigger loving rat
you ugly ass nigger whipping
!
" ?
https : / / t . co / Eu2BI2TpRu
, straight up
.
@KEEMSTAR fuck you nigger gnome
aids https : / / t . co / tYJs3yXkdO
ass
loser
nigga
@e4sy e https :
https : / / t . co /
i hate you
stupid ass bitch nigga
up https : / / t . co /
cunt
Muslim nigga #QuitYourJobIn5Words
nigga ? Abby shut the
SHIT FUCK GOD FUCKING
shut up and kill
FUCKING
HOMO IDIOT NIGGER NIGGER
nigger
get cancer
god dammjt I hope I
https : / / t . co /
4Y4QqUwgIn
be5Mvy71FT
BtJLrrqV8A
dE1V2i7lyO
msiP7qryPj
T4bjtkUCFA
is good but nigger is
isn't on discord
lmfaoooo nigga fuck you ?
nigga would think I'm fr
nigger
gay cunt idiot
pussy cunt sand nigger
stinky feet having cunt
or nigger ? https : / / t . co /
retard spic
RT @Vivalaquintaa : RT @Letmepopmyshit : When
she is
tranny
! " - white ppl
spic kike
who sucks nigger dicks
WHORE LYING SACK OF SHIT
you have no friends and

Figure 7.15 is a word tree for the word ‘kike’ at the level of stemming, which is a compact tree with few branches especially after the word. Unsurprisingly, all uses of this word are negative.



#IStandWithHateSpeech nigger faggot tranny spic
@Nas Giggers @Mexilente
stupid nigger , nobody likes a
are reading this
REVOLUTION LOOKS LIKE
YOU FUCKING
kike
ass nigger
Soros funded and started BLM ,
that steals shir

Figure 7.15: Word tree for ‘kike’ at the level of stemming.

For the word tree for ‘kike’ at the level of generalisations (Figure 7.16), NVivo changes ‘kike’ to ‘yid’, but now is missing the references to ‘kike’. This new word tree is the same as the word tree for the word ‘yid’ up to the level of stemming. This tree shows that all the instances in these data, of this word, relate to Tottenham Hotspur players, Eric Dier and Harry Kane and ex-Tottenham player, Jermaine Defoe. As with all the other words in these word trees, ’yid’ is used in negative messages, there are no positive uses, in the manner of which fans of the team Tottenham Hotspur, often used the word to identify themselves.

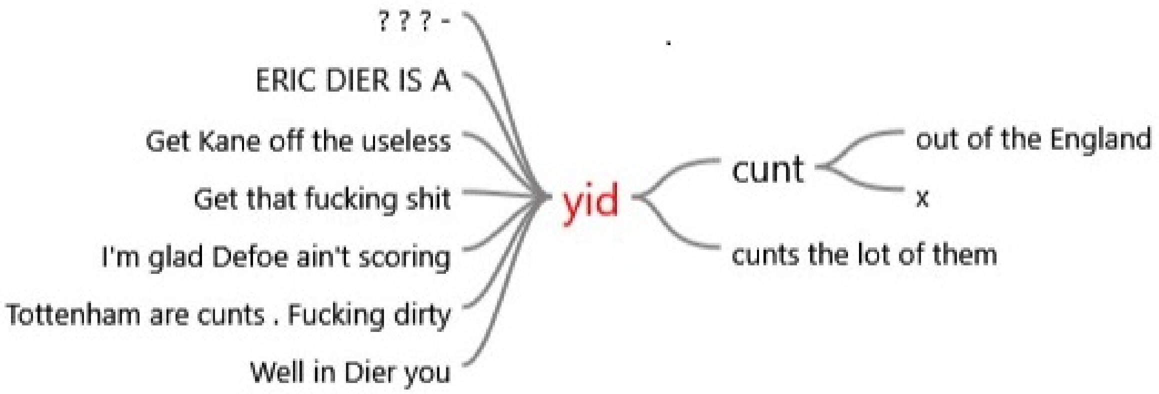


Figure 7.16: Word tree for ‘kike’ at the level of generalisations.

Figure 7.17, shows the word tree for ‘kys’ (‘kys’ is an acronym meaning ‘kill yourself’) at the level of generalisations. It shows that ‘kys’ is used with the word ‘nigger’ but not the word 'nigga', in this dataset. When 'kys' is used before the word 'nigger' it always has at least one other word in between, for example 'kys stupid nigger'.

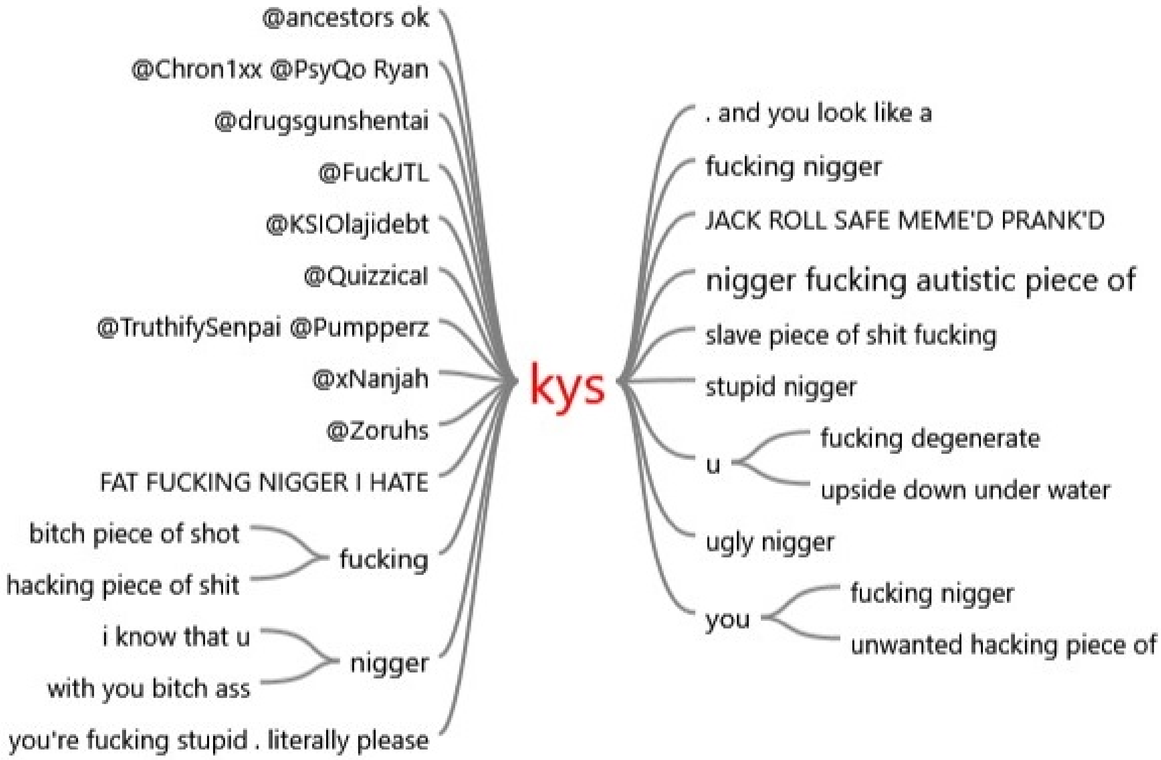


Figure 7.17: Word tree for 'kys' at the level of generalisations.

Figure 7.18, shows the word tree for 'kill' at the level of generalisations. The word tree for 'kill' is far more complex than the one for 'kys', although the most common word preceding 'kill' is 'fucking' and the most common word succeeding it is 'yourself'.

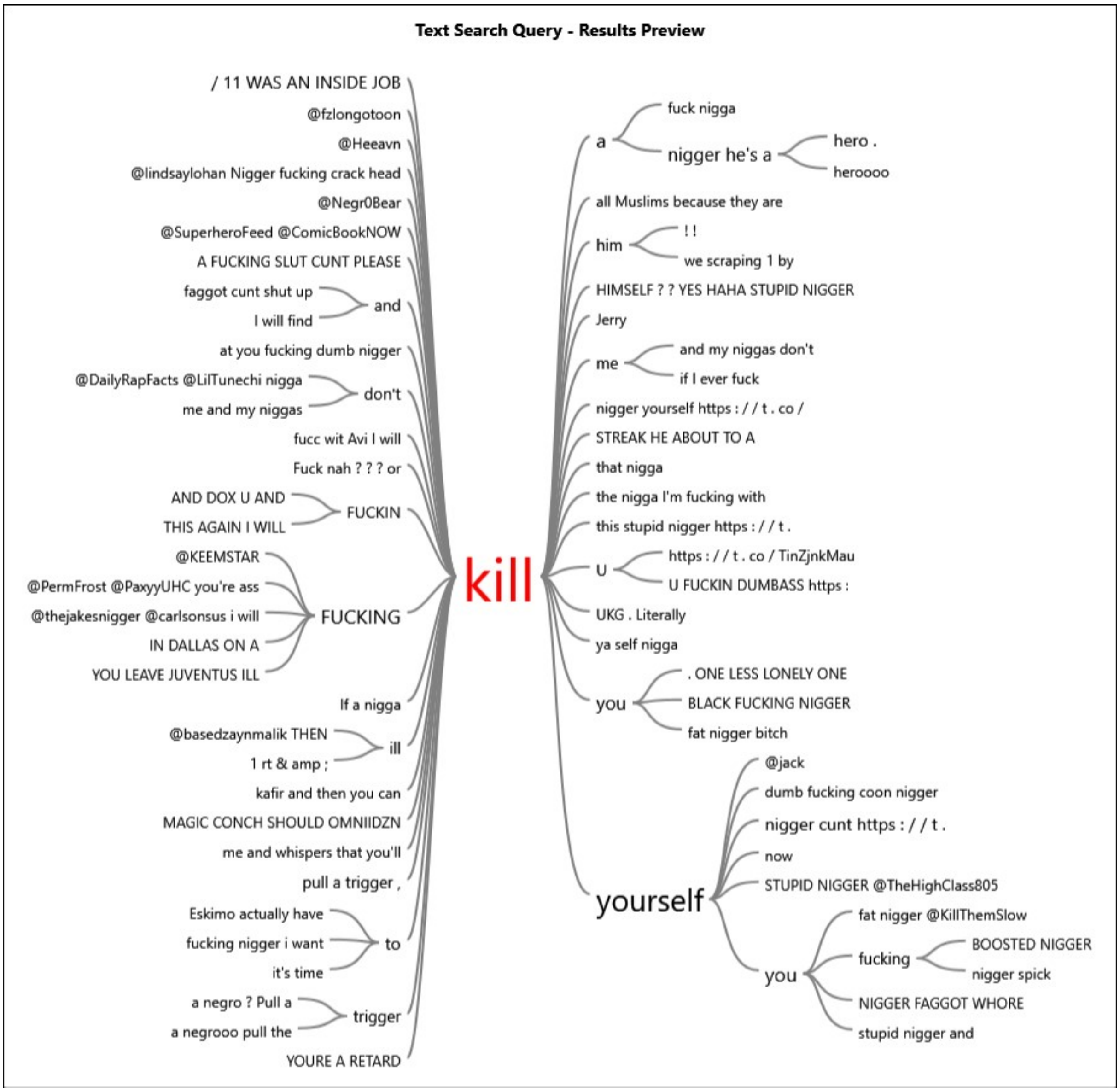


Figure 7.18: Word tree for 'kill' at the level of generalisations.

In the following section a qualitative analysis and discussion of the tweet data is given, along with a discussion of the grounded theory analysis and results.

7.3.4 Discursive Analysis Results

As well as the previous analyses, prior to the thematic analysis of grounded theory, some tweet examples were analysed discursively, to give an idea of the kinds of data making up the dataset. There are some examples of anti-white tweets with references to 'crackers' and 'whitey', but most tweets denigrate other ethnicities, the vast majority of which are anti-black.

In the Pred R dataset, the tweets that have large numbers of retweets are mostly these kind with the word 'nigga', which may or may not be used with racist intent. While such tweets may be of interest, for example considering why people create such tweets using such language, they are not a primary concern when identifying racist tweeters. Other measures, that are more suited to identify potentially problematic racist tweeters were discussed in Section 5.14.

7.3.5 Analysis of the Popular Retweets Results

When a tweet is retweeted, the retweeter might have a different understanding of its use, as might the wider audience of the tweets.[9] The 20 most popular retweeted tweets from Pred R are given in Table 7.8

Table 7.8 shows the top 20 retweeted tweets. For all of these whether they are racist or not, is, at best, debatable. These were predicted as racist because the machine learning predictor does a poor job of handling the word, 'nigga', flagging a number of false positives as racist. This is unsurprising since it is very difficult for humans to determine the intent behind the use of the word. However as already noted, even uses of this word with benign intent, are potentially problematic, and so their flagging by the software may not necessarily be problematic.

Some interesting characteristics of all the discrsive sample tweets will now be discussed.

In the dataset there are a large number of tweets of the form 'X is a stupid nigger', where X is a name, often in this dataset X equals 'Alex'. See Section 7.3.1 for an

[9] This audience might be quite different to the audience of the original tweet, since, for example, the tweet might be retweeted several hours after being first sent, and so a different time zone and therefore geographical region would likely be the audience of the retweet.

Table 7.8: Top 20 most retweeted tweets in Pred R.

Text	Count
Why black people don't swim , nigga almost drowned and tried to play it off ??? https://t.co/kihaADVhBb	20,465
Welcome to AMERICA, where ABC gives the KKK an interview, but we can't voice #BlackLivesMatter https://t.co/mxHwGpPBbl	13,919
You don't even want Muslims in the country. Shut the fuck up nigga https://t.co/NIrd3CxYul	5,096
When A nigga Don't Give A fuck About You?? https://t.co/VfztxI1Ggb	4,018
offensive gf: thinks she's black, calls you nigga, tells you to "shove your nigger cock in her nigger lips" https://t.co/3TtZan8g9i	3,534
When you know somebody a fuck nigga and they try to say what's up to you?????? https://t.co/U4t9PkBtFb	2,523
nigga I can't stop fucking laughing, the hell is this https://t.co/ZsfFPrTJIm	2,032
Asap Rocky nigger https://t.co/pvjDmoa3dB	1,714
When you get promoted from Field hand to House nigger https://t.co/GXmCe2lGmg	1,475
Every nigga that listens to Lil Uzi lip syncs like this https://t.co/c0RmL1NJn6	1,448
I'm a man,a son,a dad,a bro trolled n abused by lady presstitute &amp; her paki frnds abused my nation whom to contact ? https://t.co/oFt7Vxxqej	1,341
When your girl try to give a nigga you don't fuck with a hug https://t.co/0iacwQNUPf	1,009
Popson, African Twitter coon obsessed with Black culture and Black women–yet still hates Black people. https://t.co/h8blbpKEMp	944
when ya girl start singing "you was just another nigga on the hit list" https://t.co/Au0MFfKOxY	820
Bitch, if you don't get yo saddle back ass off my nigga Mater https://t.co/9Z0iyg55Dh	817
Man fuck this nigga ?? https://t.co/7ZMjZmhnwh	810
everybody type in the chat alex is a stupid nigger	795
get away from my nigga mater with that trash ass flag https://t.co/hfUauSrJn0	787
alex is a stupid nigger	758
never forget about the girl that was there for you when you was just a broke nigga telling her your dreams.	724

explanation of this.

There are a large number of vitriolic and aggressive tweets of the form ‘[do something] nigger’, where ‘do something’is a negative action. These range from:

@immediateIy go away nigger,

to

@ancestors ok kys nigger,

where ‘kys’ stands for kill yourself. Many of these are orders to ‘shut up’, and this concept of telling people that their messages are unwanted or worthless, is seen throughout the dataset.

The phrase ‘a fucking nigger’ occurs quite frequently in the dataset, but ‘a fucking nigga’ does not occur at all.

As mentioned in section 2.1, Mondal et al. (2017) when identifying hate speech on Twitter and Whisper, rather than using keywords hateful keywords used the sentence template: I <intensity><user intent><hate target>. This pattern was seen in a relatively small number of tweets, around 4% of those predicted racist tweets that were retweeted. Also a large number of these ended with ‘ass nigga’ suggesting perhaps the intent of the sender was nonracist.

These data, the word cloud results from Section 7.3.2 and the word tree results from Section 7.3.3 provides some interesting insights into the structural forms of the racist tweet data. Certainly the structure used by Mondal et al. (ibid.) seems far too simplistic. Using their sentence template: I <intensity><user intent><hate target> would exclude many of the forms of hate seen in Section 7.3.3. Many of the tweets were of the form <hate target>< negative action> or the reverse, for example ’kys nigger’. In terms of collecting data it would also be more fruitful just searching for the two words: ’nigga’ and ’nigger’, although of course this would limit the racist tweets, excluding other racial epithets. Also in terms of identification, synonyms for homosexuality would be fruitful indicators of potentially racist tweets, or more generally just abusive tweets, since such synonyms are seen in a large number of abusive messages, for example the phrase: ’nigger faggot’ is seen with high-frequency in the tweets.

The second most popular retweeted tweet in the sample (with 2542 retweets), contains the following text:

> When A Nigga Don’t Give A figgk About You?? `https://t.co/VfztxI1Ggb`

the URL directs to an ‘account suspended’ message on Twitter. Without the accompanying visual imagery it is difficult to determine intent of this tweet, it seems unlikely that its use of nigga is not intended to be racist.

There are a number of references within the tweets to ‘one less lonely nigger’. This is a reference to Justin Bieber who recorded a video version of his hit song ‘One Less Lonely Girl’ in which he sang the lyric ‘nigger’ instead of ‘girl’.

The majority of the predicted racist tweets contain either ‘nigger’ or ‘nigga’, as would be expected since they were two of the largest original data files collected.

The stereotyping of black people’s bodies is seen in the following tweets:

> @sayetaryor @WhiteWizaard @mc_morn @TalibKweli I’ll slap the ash of ya chapped Nigger lips in a debate, faggot.

This is similar rhetoric to that seen in Brown’s (2009) analysis of white supremacist discourse, where whites are regarded as a physically superior race, and blacks are seen as subhuman, close to non-human animals.

There are a number of uses of animalistic metaphors also seen in in Brown's (2009) analysis. For example, the use of the word 'chimp' in the following tweet:

> @KojoLDN dumb nigger you started this conversation with me! another low IQ chimp.

Black people's ancestors are likened to animals such as in this tweet, when the death of a gorilla is dismissed:

> who gives a figgk about a stupid gorilla nobody gives a shit when a nigger dies so why do people care when their ancestors do

There are pejorative references to 'nigger lovers', that is people who see blacks as equal to whites (Helms, 1984). This can be seen in the following tweet:

> @JPaul_KY @Coximus2 @ironghazi @HarrisNye I bet you're a nigger lover

There are other instances of racism aimed at different ethnicities for example:

> @Evade_Svaarj @cLaw_Jitta @Gxlaxy___ Shut it you fat paki cigg, get back in the tesco van and drive back to your paki country prick

Use of slurs against the Roma is common in Europe, and McGarry (2017) argues that Romaphobia is often seen as the last 'acceptable' form of racism. In the tweet data, there are references to 'gypos' a derogatory term in the UK, used against the Roma (Twomey, 2017), for example the following tweet

> @ShameADriver i give credit to the disabled Gypo who leapt out of his well laden van in Boston today `https://t.co/Duy4r2f09i`,

points to an image of a fully laden van parked outside a shop. The reply to this tweet is:

> that lot will be dumped in a country lane by the end of the day.

Minorities are often seen as ‘dirty’ such as in this example, related to Tottenham Hotspur Football Club:

> You know what. Tottenham are ciggs. fucking dirty yid ciggs the lot of them

which may be a reference to ‘dirty play’ i.e. unsportsman-like behaviour. There are however, many instances of where dirty is used to mean ‘unclean’ such as the following:

> @czrlit @Tagzhh @xKalizs @hussi96 IM NOT HAHAHA YOU fucking DIRTY UGLY SOMALIAN cunt HAHAHAHA STUPID fucking paki cunt

There are references to politicians, such as Barack Obama seen in the following tweet:

> 7 years of hell all caused by that nigger you worship. I’m done with it. #Obama figgk him.

Also, of course, there are references to Donald Trump. For example, the third most popular retweeted tweet in the sample (with 2,428 retweets) is:

> You don’t even want Muslims in the country. Shut the figg up nigga https://t.co/NIrd3CxYul

the URL directs to the @realDonaldTrump Twitter account.

The sixth most popular retweeted tweet in the sample (with 1,341 retweets) is:

> I'm a man,a son,a dad,a bro trolled n abused by lady presstitute &amp; her paki frnds abused my nation whom to contact ? `https://t.co/oFt7Vxxqej`

the URL directs to a message on Maneka Gandhi's India's, Union Minister for Women and Child Development, Twitter account, which reads:

> Are you a woman who is trolled or abused? Inform me at gandhim@nic.in.

There are references to other politicians with racist slurs, for example:

> #UniteBlue Get Behind @VoteFedalei he is the Dem running against greaseball Trey Gowdy .RT..RT...RT..VOTE or DONATE `https://t.co/ohuvxgUD4j`

the URL points to Chris Fedelai's Twitter page who was running to represent South Carolina's 4th Congressional District.[10] 'Greaseball' is a slur usually aimed at people of Italian American ethnicity, which Trey Gowdy is not, so possibly this was used to connote political sleaziness.

There are a number of instances of wishing cancer on people, such as:

> @xZekkay GET FKN CANCER U LIL Nigger

The use of capital letters in this tweet is commonly seen in the data. Capitalising text is a common rhetorical device used for emphasis, akin to shouting in the real world.

There are a number of sport-related attacks using racist and other stereotyping, for example the following tweet uses the word 'pikey'[11] the tweeter appears to be a Chelsea Football Club supporter and the tweet is aimed at West Ham United fans Twitter page:

[10] He lost.

[11] 'Pikey' is a term used in a derogatory manner often signifying a slur on Roma people

> ? pikey ciggs ? https://t.co/5sr0rwyrUm.

Paul Pogba, is a victim of racist abuse from both Juventus fans not wanting him to leave their club:

> @paulpogba YOU fucking Nigger IF YOU LEAVE JUVENTUS ILL fucking KILL YOU BLACK fucking Nigger

and fans of other clubs:

> Pogba nigger fuck off.

There are a lot of uses of the word ‘yid’ a derogatory term used against Jews and also Tottenham Hotspur, who have traditionally had a significant number of Jewish followers, but also used by them as a signifier of belonging to a minority group, similar to the way nigga is used by African Americans. For example the following tweet refers to, Harry Kane, Tottenham’s leading goalscorer, during a game where he played for England:

> Get Kane off the useless yid cunt.

Jews are also slurred with use of the word ‘kike’. For example this attack on the Jewish American financier George Soros:

> fuck YOU, @georgesoros, THIS IS WHAT A REVOLUTION LOOKS LIKE YOU fucking KIKE.

While of course there is a potential worldwide audience for any tweet, many of tweets such as the sports ones already mentioned, are very specific to a particular country, or at least a specific sporting team. Tweets that insult West Ham United fans are unlikely to travel very far beyond the UK, albeit many sports teams have fans worldwide. [12]

Aside from the sporting tribalism, there are other examples of regional identities and cultures seen in the data. For example in the following tweet, there are mentions of ‘unbelieving kaffir bastards’:

> Shoutout to all the unbelieving kaffir bastards not participating in Ramadan, hope it’s LITTY for y’all ???

’Kaffir’ is an Arabic term meaning ‘infidel’, but is also used in South Africa, as a derogatory term for a black person.

The most retweeted tweet in the sample (retweeted 3,534 times) is the following:

> offensive gf: thinks she’s black, calls you Nigger, tells you to "shove your nigger cock in her nigger lips" `https://t.co/3TtZan8g9i`

the link points to a picture of Kylie Jenner, an American reality TV personality. This tweet appears to be some kind of joke, presumably without racist intent, although use of the word nigger even in humour is problematic. It also objectifies women, thus it is an example of intersectionality: racism intersecting with misogyny.

There are instances of abuse aimed at a person’s body shape, intersecting with racist abuse, such as the following:

> @NaS_Nevaeh WHY BITCH Nigger BLOCK YOU fucking WHALE ASS Nigger BITCH

There are also other examples of intersectionality, for example

[12] Although this is quite unlikely for West Ham.

@Cluhrk get fucking cancer you retard faggot nigger gay cigg idiot

uses the words ‘gay’ and ‘faggot’ and ‘retard’ as insults, thus intersecting homophobic and disablist discourse with racism.

This analysis gives an initial insight into the nature of the tweets. This is explored further in the following section where the results of the grounded theory analysis are given.

7.3.6 Grounded Theory Results

Grounded theory aims to reach theoretical saturation i.e. a point at which no new concepts/categories are generated via theoretical sampling. For this research concepts were found to repeat quite quickly even over a few tweets. Since the tweets were selected based on their containing a racist keyword, it is perhaps unsurprising that theoretical saturation was reached relatively quickly. While there are discussions of conversations in the initial analysis of the tweet, most of the analysis is performed using each tweet as a stand-alone piece of data. This was done because of the complexity of obtaining and analysing conversations as a whole. The implications of this are discussed in Section 8.5.

To illustrate how concepts emerged from the data, three examples of the grounded theory coding are given in Table 7.9.

The set of focus codes generated by the data are the following:

Abuse
Request to harm themselves
Misogyny
Homophobia
Anti-Muslim
Anti-Semitic

Table 7.9: Three examples of grounded theory coding.

The Raw Data	**Initial Coding**	**Focused Coding**	**Theoretical Coding**
Shut the fuck up nigger	anti-black	Hatred of a group	Defence of in-group
	Abuse		
	Silencing the views of others	Hatred of opinions of others	
	Use of expletives	Linguistic	
@Cluhrk get fucking cancer you retard faggot nigger gay cigg idiot	anti-black	Hatred of an individual	Defence of in-group
	Homophobia	Sexuality	
	Abuse		
	Wish illness on target	Hatred of an individual	
	Use of expletives	Linguistic	
	Target of abuse is stupid	Disability	
I'm a man,a son,a dad,a bro trolled n abused by lady presstitute &amp; her paki frnds abused my nation whom to contact ?	anti-Asian	Hatred of a group	Defence of in-group
	Misogyny	Hatred of a group	
	Sarcasm	Linguistic	

Anti-black

Anti-white

Anti-Asian

Anti-non-Muslim

Anti-Roma

Abuse of celebrities

Sport

Politics

Humans as animals

Use of expletives

Bodily form

I'm a victim

Sex

Whites who befriend blacks

Wish illness on target

Go back to your own country

Silencing the views of others

Anger

Use of capital letters

Repetition of words

Use of URL

Use of emoticon

Use of hashtag

Use of sarcasm

Target of abuse is dirty

Target of abuse is stupid

Target of abuse is lazy

Target of abuse is dishonest

Target of abuse is ugly

Target of abuse is fake

In further analysis of these concepts, it was found that the rhetoric expressed in the tweets was so narrow that it could be fairly satisfactory summed up by a single theoretical concept, the defence of the in-group. This is consistent with the literature on hate groups and hate speech, where in-group protection is a common theme.

There is not much comment in the automated racist detection literature on the forms that the hate takes, this is in part due to the fact that the researchers focused on, for example tweet specifically aimed against Blacks (Kwok and Wang, 2013). As expected there are the major (in terms of number of population) ethnicities represented in terms of hate: black, white and Asian. Perhaps less surprisingly are the anti-Roma tweets, since this is a relatively small dataset. Again, perhaps unsurprisingly, there were a large number of misogynist and homophobic tweets, but the homophobia was so ubiquitous, that the intersectionality of it and racism will be fruitful to explore further in terms of the automated detection of such tweets. There are also a large number of tweets where celebrities were the target of abuse, this to be expected since platforms such as Twitter have high levels of discussion about celebrities. Sporting and political figures were also prevalent as targets of the abuse, in particular footballers. There were also a surprisingly large number of references to humans as animals, again a theme like homophobia, that requires more intersectional analysis. Indeed, while homophobia has a somewhat obvious intersectional relationship with racism, references to humans as animals perhaps is less obvious and so on interesting artefact of the data. There were also a large number of references to bodily form and also sexual function. Again this is an interesting area of intersectionality with racism. The tweets were very aggressive, and exhibited large amounts of anger, with swearing being almost universal in the racist tweets, and there were large amounts of capital letters, emoticons, hashtags and sarcasm. The majority of the racist tweets were insults calling the target: dirty, stupid, lazy, dishonest, ugly and fake. There were surprisingly few calls for the target to 'go back to their own country' but these were visible. A large number of the abusive tweets wished illness, usually cancer on the target. The themes seen in tweets are reflective of the larger racist literature, although there are some dissimilarities, with lower than expected calls for people of different ethnicities to return to their 'homelands', and the very significant amounts of homophobia perhaps more than would be expected.

The single theme result is somewhat disappointing, but perhaps to be expected, as there is a lot of repetition in the tweets. Despite this, these data are useful for an important part of this thesis: the identification of racist accounts. The set of focused coding codes were used as input into a machine learning algorithm to identify racist tweeters, as discussed in Sections 5.14 and 7.4.

7.4 Accounts Results

. From the Pred R dataset there are 462,145 accounts that sent at least one original tweet or retweet. The top 20 accounts for the measures discussed in Section 5.14 are given in the following tables: Table 7.10, Table 7.11, Table 7.12 and Table 7.13 which show the top 20 'influential racists' from Pred R, in order of *oCount*, *rDistinctCount*, *rCount*, and *retweetRatio*, respectively.

Table 7.10: Top 20 'Influential Racists' from Pred R, in order of *oCount*.

Account	oCount	rofothers_cnt	rCount	rDistinctCount	retweetRatio
Account1	256	0	6	3	2
Account2	185	1	33	22	1.50
Account3	115	0	0	0	0
Account4	107	0	1	1	1
Account5	98	0	0	0	0
Account6	89	16	12	11	1.09
Account7	88	31	1	1	1
Account8	88	31	16	14	1.14
Account9	87	0	0	0	0
Account10	83	15	28	17	1.65
Account11	82	15	32	22	1.45
Account12	81	0	4	4	1
Account13	79	23	13	12	1.08
Account14	78	18	18	14	1.29
Account15	77	0	0	0	0
Account16	76	0	1	1	1
Account17	75	0	0	0	0
Account18	75	0	1	1	1
Account19	75	34	54	44	1.23
Account20	75	56	199	41	4.85

There are 65 distinct accounts in these tables. The accounts in the tables are colour-coded green for Table 7.10, blue for Table 7.11, yellow for Table 7.12 and orange for Table

Table 7.11: Top 20 'Influential Racists' from Pred R, in order of *rCount*.

Account	oCount	rofothers_cnt	rCount	rDistinctCount	retweetRatio
Account21	1	0	17 868	29	616
Account22	1	14	13 919	1	13 919
Account23	1	0	8037	1	8037
Account24	10	1	7203	14	515
Account25	7	0	5096	1	5096
Account26	4	3	4022	4	1006
Account27	2	0	3539	13	272
Account28	1	0	3517	24	147
Account29	1	4	2523	1	2523
Account30	34	16	2480	38	65
Account31	3	5	2480	38	65
Account32	1	0	2182	15	145
Account33	1	0	1986	7	284
Account34	2	0	1917	8	240
Account35	10	14	1898	5	380
Account36	2	0	1756	1	1756
Account37	1	1	1729	5	346
Account38	1	0	1717	4	429
Account39	2	9	1700	2	850
Account40	1	0	1579	19	83

Table 7.12: Top 20 'Influential Racists' from Pred R, in order of *rDistinctCount*.

Account	oCount	rofothers_cnt	rCount	rDistinctCount	retweetRatio
Account53	39	20	105	49	2.14
Account19	75	34	54	44	1.23
Account54	1	0	54	44	1.23
Account20	75	56	199	41	4.85
Account30	34	16	2480	38	65.26
Account31	3	5	2480	38	65.26
Account55	23	5	275	37	7.43
Account56	3	4	35	31	1.13
Account57	12	2	687	30	22.90
Account58	72	4	40	30	1.33
Account21	1	0	17 868	29	616.14
Account59	30	9	672	25	26.88
Account60	34	27	463	25	18.52
Account28	1	0	3517	24	146.54
Account61	15	5	726	24	30.25
Account62	31	19	360	24	15
Account63	61	51	133	24	5.54
Account64	0	1	133	24	5.54
Account65	20	36	351	22	15.95
Account11	82	15	32	22	1.45

Table 7.13: Top 20 'Influential Racists' from Pred R, in order of *retweetRatio*.

Account	oCount	rofothers_cnt	rCount	rDistinctCount	retweetRatio
Account22	1	14	13 919	1	13 919
Account23	1	0	8037	1	8037
Account25	7	0	5096	1	5096
Account41	1	4	2523	1	2523
Account42	2	0	1756	1	1756
Account43	1	1	1100	1	1100
Account26	4	3	4022	4	1006
Account39	2	9	1700	2	850
Account44	1	8	820	1	820
Account45	2	0	1549	2	775
Account46	0	1	1424	2	712
Account47	1	0	700	1	700
Account21	1	0	17 868	29	616
Account24	10	1	7203	14	515
Account48	5	1	945	2	473
Account38	1	0	1717	4	429
Account49	6	8	834	2	417
Account50	0	2	820	2	410
Account51	0	12	798	2	399
Account52	1	2	398	1	398

7.13. The accounts are anonymised, given the name 'AccountN', where N is from 1 to 65. If an account is in more than one table it is given the colour of the first table it appears in, this being arbitrary, and just an indicator to show some accounts occur in more than one table.

In Table 7.10, the Accounts are ranked by *oCount*, the most prolific creator of tweets is Account1, an account that another Account claims is their bot account, used to offend people. This account created 256 original tweets, and did not retweet any other tweets. Indeed, of the top five original tweeting accounts there was only a single retweet, this suggests that these accounts are all *bot accounts*, that is accounts that automatically tweet, or are created as accounts that appear to disseminate automatically. Only three of Account1's tweets were retweeted, on average twice each. It does not make the top 20 of each of the other measures. Account16 is a nonracist account, but appears here due to it creating a number of tweets containing the phrase, 'Kaffir lime leaf'. The only Accounts from this table which are seen in any of the other top 20 tables, are Account11, Account19 and Account20, which all appear in Table 7.12. Account11, Account19 and Account20's tweets are largely made up of racist comments, and at the time of writing (March 2018) these accounts had been suspended by Twitter. Account2 and Account5 also have a high amount of racist content, and have been suspended. Account3 is a prolific account with over 24,000 tweets, many of which appear to be designed to offend, a large number of which are racist in nature. This account has not been suspended. Account10 does not appear to be explicitly racist, but has a lot of abusive content with the word 'nigga' in it.

There is some overlap between Tables 7.11, 7.12 and 7.13, as might be expected since a higher *rCount* means *retweetRatio* will be higher, and those with a large number of retweets are likely to have a large number of distinct retweets.

While Twitter does seem to be having some success of suspending the more prolific racist accounts, it has missed some accounts such as Account3, an account that may be of interest to law enforcement. Of course these are only 'top 20' data, so the picture might be quite different for the millions of other accounts. At first sight it seems that Twitter's efforts in policing itself are working well. However analysing accounts is very time-consuming, since each account's tweeting history has to be examined, so it is hard to say whether Twitter's efforts are working well for a large number of accounts.

These metrics were used in a machine learning process that identifies racist tweeting accounts, as discussed in the next section.

7.4.1 Accounts Machine Learning Results

. As discussed in Section 5.14.1, for each account the following variables: *oCount*, *rCount*, *rDistinctCount*, *followers_count*, and *Vocabulary*, where *oCount*, *rCount* and *rDistinctCount* were used as input features in an SVM model. The model was trained on the data split 70:30 training to test data and run with tenfold cross validation. The metrics of the model were an accuracy of 0.87, precision=0.68, recall=0.66, and F-score=0.67 for both $\beta = 1$ and $\beta = 0.5$. The D2 dataset also had a retweet dataset extracted from it, with the same variables as mentioned above. The model was then run on this D2 dataset. A random sample of 200 accounts was drawn from the predictions and their metrics were: accuracy of 0.80, precision=0.63, and recall=0.60, and F-score=0.61 for $\beta = 1$ and 0.62 for $\beta = 0.5$.

As discussed in Section 5.14.1, a similar procedure was performed, with the results from the grounded theory analysis. The set of focused coding codes were used as categorical input data. The model was again trained on the data split 70:30 training to test data and run with tenfold cross validation. The metrics of the model were an accuracy of 0.82, precision=0.69, recall=0.70, and F-score=0.70 for $\beta = 1$ and 0.69 for $\beta = 0.5$. This model was then run on the same random sample of 200 accounts from D2 in order to evaluate the grounded theory machine learning predictions. Each of the 200 accounts was coded in the same way as the earlier grounded theory analysis. Their metrics were: accuracy of 0.84, precision=0.60, recall=0.57, and F-score=0.61 for $\beta = 1$ and 0.59 for $\beta = 0.5$. As can be seen the accuracy was worse for the grounded theory analysis compared to the retweet data, but the other metrics were better.

Finally both sets of features were combined, the retweeted data and the grounded theory data both used as features. The metrics of the model were an accuracy of 0.88, precision=0.73, recall=0.71 and F-score=0.72 for both $\beta = 1$ and $\beta = 0.5$. Again the model was tested against the D2 dataset and the metrics were: accuracy of 0.88, precision=0.64, recall=0.66 and F-score=0.65 for both $\beta = 1$ and $\beta = 0.5$. So it can be seen that both sets of features performed well, but the combined features were the best performing input, suggesting that both retweeting data and grounded theory data show promise in the prediction of racist tweeting accounts.

7.5 Chapter Summary

This chapter presents results for the machine learning procedures and qualitative analysis. For the machine learning results first the effects of varying oversampling fraction of negatives are given. This is followed by a discussion of text, user, geographical and temporal features and their efficacy as input to the machine learning algorithms. Features are further considered in both discussing whether text should be treated as Ngrams or BOW and whether hour, hour+day or neither should be used as additional features. Then seven different algorithms are compared by analysing their metrics.

The temporal aspects of the tweets are further discussed in relation to the different datasets including both the input and predicted data. Then the qualitative results are discussed, first summary data from an NLTK analysis is given, followed by analysis of word clouds and word trees of the data. Then a brief discursive analysis is given followed by a grounded theory analysis of the data.

Finally accounts are discussed and results of machine learning processes using metrics from the accounts, from the grounded theory analysis and a combination of these are given.

The datasets used were: the hand coded nonracist tweets, designated as *Inp NR*; the hand coded racist tweets, designated as *Inp R*; the predicted nonracist tweets, designated as *Pred NR*, and the predicted racist tweets, designated as *Pred R*.

The effects of varying oversampling was investigated by examining the output for different *fraction of negatives*. Fraction of negatives is the proportion of negatives in relation to the proportion of positives being one. The fraction of negatives of 0.0385 (equivalent to a ratio of positives to negatives of 1 to 0.4) was chosen as the metrics for this value were optimal.

Each of the features discussed in Section 6.4.3 were added to N5 and run with a SVM model with tenfold cross validation. The model was also run with no additional features and this was denoted by BASELINE in the table. Only HOUR_TWEET was an improvement on the BASELINE. DAY_TWEET was the next best performing additional feature, and these two features were used in further analysis. The rest performed poorly and so were not used further. An SVM model was run a total of 12 times. There were four different word features: BOW, Bi, Tri and N5 corresponding to bag of words features, bigram features, trigram features and Ngrams from 1 to 5 combined. For each of the different word features, the model was run three times with the following additional features: hour and day as additional features, hour as an additional feature and neither additional feature. The best combination overall was N5 with either hour or neither as an additional feature but since hour and day had positive effects on some algorithm/feature combinations this was included as well, and when algorithms were compared they were run with N5, N5 plus hour and N5 plus hour and day as their input feature sets.

The algorithms NB, LR, SVM, RF, DT, GBT and ANN were tested with three different feature sets as input: text, text+hour and text+hour+day, giving 7*3 = 21 different combinations of algorithms, features and preprocessing. SVM performed best in terms of accuracy, precision and recall and so was used in the prediction stage. NB also had reasonable accuracy, precision and recall, but not as good as SVM. LR had good accuracy but poor precision and recall, and the other models: RF, DT, GBT and ANN all had reasonable accuracy but poor precision and recall. The addition of our as a feature was generally beneficial, although not always, and the addition of hour plus day was for the most part not beneficial. The best performing model that of SVM with hour as an additional feature, was run on the D1 dataset minus the hand annotated sample. This generated a predicted values dataset which was analysed as two separate datasets: one containing predicted racist tweets (Pred R) and the other containing predicted non-racist tweets (Pred NR).

In terms of the hour of day of the tweets, both the Inp NR and Inp R data show much more variability than Pred NR and Pred R, especially Inp R. This is to be expected because of their much smaller sample size. Inp R in particular shows a lot of variability both between hours and between days. The number of tweets generally decreased from around midnight to 4 or 5 in the morning, and then gradually increased to around 10 AM. The period from 10 AM to approximately 8 AM is more variable for both Inp NR and Inp R; there is a lot of variability and any trend is unclear. For Pred NR there is a slight downward trend until 4 PM, and for Pred R there is an upward trend until 4 PM and then a slight reduction to 5 PM. For each there is then an increase to a peak of 9 or 10 PM followed by reduction towards the early-morning minimum.

With respect to hourly percentages of tweets *by day*, the data from both Inp NR and

Inp R is inconclusive, except that surprisingly for Inp R Friday is the day with the lowest percentage of tweets, and its late-night peak is also less pronounced than the other days except Sunday and Thursday. For Pred NR Saturday and Sunday have the fewest tweets for much of the day and their late-night peaks are less pronounced than the other days. The other days of the week show reasonably similar patterns throughout the day, although Friday again, perhaps surprisingly, shows lower percentages from approximately midday until 8 PM, thereafter it increases comparably with Monday through Thursday. For Pred R Saturday and Sunday also have the fewest tweets, with Friday the most between 9 PM and 10 PM. For Pred NR most days show a decrease from 10 or 11 AM until around three or 4 PM then a steady increase until 8 or 9 PM, followed by a decrease during the early morning hours until 5 AM. In contrast with this Pred R most days show a steady increase from 5 AM until 4 PM, then a drop at 5 PM, followed by a sharp increase until 9 or 10 PM, followed by a similar decrease until 5 AM. The minimum percentage of tweets occurs on Sunday except for in Inp R when it occurs on Tuesday, the maximum percentage of tweets is more varied occurring on Friday, Saturday, Wednesday and Thursday for Inp NR, Inp R, Pred NR, and Pred R respectively.

For Inp NR the quietest hour is 3 or 4 AM throughout the week, for Inp R the quietest hour is more variable being anywhere from 3 to 7 AM, for Pred NR it is always 4 or 5 AM and for Pred R it is always 5 AM. For Inp NR the busiest hour is 9 PM for Monday, Tuesday and Thursday, 8 PM for Wednesday, 4 PM for Saturday and 11 AM on Sunday. For Inp R there is also variability Monday, Tuesday, Wednesday, Saturday and Sunday are busiest during the 'normal peak' hours of 8 PM through 10 PM, but Thursday is busiest at 3 PM and Friday busiest at 6 PM. For Pred NR and Pred R the busiest hour is always 9 or 10 PM.

For the qualitative analysis the tweets in Pred R that had been retweeted more than once were analysed. The words 'nigger' and 'nigga' were the two most prevalent words, the data showed the expletive ridden nature of the tweets, with four swear words in the top 12 most common words, or top seven if stopwords are excluded. Indeed the vocabulary is extremely negative, of the top 25 words excluding stopwords, 14 are one of: ethnic slurs, swear words, negative adjectives and the words 'hate' and 'bitch'. There were very few hashtags in the data.

'Alex' is the only name in the top 25 words, which is a reference to the YouTuber, Keemstar. All of the bigrams contained insults, slurs or swear words, or combinations of these, except for three of them: 'type chat', 'everyone type' and 'chat Alex', which are all also related to Keemstar. There is a similar situation for trigrams.

The word clouds illustrated the prevalence of messages related to black ethnic slurs in the tweets, and the preponderance of swearing and other hateful words. The word tree for 'gay' shows the relentlessly negative use of the word. Interestingly its use as an adjective preceding the word 'nigga' is always with the addition of the word 'ass', whereas its use as an adjective preceding the word 'nigger' is more varied. The word tree for 'kys' (lq kill yourself') showed shows that 'kys' is used with the word 'nigger' but not the word 'nigga', in this dataset. When 'kys' is used before the word 'nigger' it always has at least one other word in between, for example 'kys stupid nigger'. The word tree for 'kill' is far more complex than the one for 'kys', although the most common word preceding 'kill' is 'fucking' and the most common word succeeding it is 'yourself'. The tweet examples that were analysed discursively showed some examples of anti-white tweets, but most tweets denigrated other ethnicities, the vast majority of which were anti-black.

For all of for all of the top 20 retweeted tweets predicted by machine learning whether they are racist or not, is, at best, debatable. These were predicted as racist because the machine learning predictor does a poor job of handling the word, 'nigga'.

In the dataset there are a large number of tweets of the form 'X is a stupid nigger', where X is a name, often in this dataset X equals 'Alex'. There are a large number of vitriolic and aggressive tweets of the form '[do something] nigger', where 'do something'is a negative action. Many of these are orders to 'shut up', and this concept of telling people that their messages are unwanted or worthless, is seen throughout the dataset. The phrase 'a fucking nigger' occurs quite frequently in the dataset, but 'a fucking nigga' does not occur at all.

Mondal et al. (2017) when identifying hate speech on Twitter and Whisper, rather than using keywords hateful keywords used the sentence template: I <intensity><user intent><hate target>. This pattern was seen in a relatively small number of tweets, around 4% of those predicted racist tweets that were retweeted. Also large number of these ended with 'ass nigga' suggesting perhaps the intent of the sender was nonracist.

The stereotyping of black people's bodies is a common theme in the tweets, as is likening of black people to animals.

There are instances of racism aimed at different ethnicities. There are references to 'gypos' a derogatory term in the UK, used against the Roma. There are a number of attacks on Jews and Muslims.

Minorities are often seen as 'dirty'.

There are attacks on politicians and sporting celebrities.

There are a number of instances of wishing cancer on people.

There are a number of examples of intersectionality: racism intersecting with attacks on women, bodily form, homosexuality and disability.

The use of capital letters is commonly seen in the data.

From the grounded theory analysis, a set of focus codes was generated by the data and from these, it was found that the rhetoric expressed in the tweets was so narrow that it could be fairly satisfactory summed up by a single theoretical concept, the defence of the in-group.

From the Pred R dataset there were 462,145 accounts that sent at least one original tweet or retweet. An examination of the top 20 'influential racists' from Pred R, in order of *oCount*, *rDistinctCount*, *rCount*, and *retweetRatio*, showed that the most prolific accounts were likely bot accounts, since they did not retweet. The top 20 in *oCount* showed little overlap with the other three measures. Some of the accounts had been suspended by Twitter, others not.

For each account the following variables: *oCount*, *rCount*, *rDistinctCount*, *followers_count*, and *Vocabulary*, where *oCount*, *rCount* and *rDistinctCount* were used as input features in an SVM model. The model was trained on the data split 70:30 training to test data and run with tenfold cross validation. The metrics of the model were an accuracy of 0.87, precision=0.68, recall=0.66, and F-score=0.67 for both $\beta = 1$ and $\beta = 0.5$.

The D2 dataset also had a retweet dataset extracted from it, with the same variables as mentioned above. The model was then run on this D2 dataset. A random sample of 200 accounts was drawn from the predictions and their metrics were: accuracy of 0.80, precision=0.63, and recall=0.60, and F-score=0.61 for $\beta = 1$ and 0.62 for $\beta = 0.5$.

A similar procedure was performed, with the results from the grounded theory analysis. The set of focused coding codes were used as categorical input data. The model was again trained on the data split 70:30 training to test data and run with tenfold cross validation. The metrics of the model were an accuracy of 0.82, precision=0.69, recall=0.70, and F-score=0.70 for $\beta = 1$ and 0.69 for $\beta = 0.5$.

This model was then run on the same random sample of 200 accounts from D2 in order to evaluate the grounded theory machine learning predictions. Each of the 200 accounts was coded in the same way as the earlier grounded theory analysis. Their metrics were: accuracy of 0.84, precision=0.60, recall=0.57, and F-score=0.61 for $\beta = 1$ and 0.59 for $\beta = 0.5$. As can be seen the accuracy was worse for the grounded theory analysis compared to the retweet data, but the other metrics were better.

Finally both sets of features were combined, the retweeted data and the grounded theory data both used as features. The metrics of the model were an accuracy of 0.88, precision=0.73, recall=0.71 and F-score=0.72 for both $\beta = 1$ and $\beta = 0.5$. Again the model was tested against the D2 dataset and the metrics were: accuracy of 0.88, precision=0.64, recall=0.66 and F-score=0.65 for both $\beta = 1$ and $\beta = 0.5$. So it can be seen that both sets of features performed well, but the combined features were the best performing input, suggesting that both retweeting data and grounded theory data show promise in the prediction of racist tweeting accounts.

There now follows the final chapter of the thesis, which discusses these results in light of the theoretical framework and research questions.

Chapter 8

Discussion

This thesis has explored the application of novel machine learning and big data techniques to the automated identification of racist tweets and tweeters, along with a qualitative analysis of racist tweet data.

It was found that due to the scale of the Twitter data a Hadoop/Spark cluster was necessary to perform machine learning routines. Using this to identify racist tweets, a systematic comparison of seven different algorithms, and a large number of textual, user-derived and geographical features was performed. New features: time of day and day of week were also evaluated. It was found that the combination of support vector machines with hour of day as additional feature was optimal for accuracy (0.93) and AUPRC (0.86).

Additionally it was discovered that a novel machine learning system using metrics from the racist tweets, concepts from the grounded theory and a combination of the two as feature inputs performed well in identifying racist accounts. All three sets of features gave accuracy of at least 0.82.

Analysis of the time of day of the tweets showed a more rapid decline in tweeting after midnight in the racist tweets than the nonracist tweets, and the peak days of tweeting were Thursday and Friday respectively. This is contrary to what might be expected under a hypothesis that low self-control, which might occur after alcohol consumption, leads to antisocial behaviour.

It emerged from the analysis of the tweets containing ethnic slurs, that in many instances the ambiguity of the tweets meant that they were difficult to classify, for both humans and machines, as to whether the tweeter's intentions were racist or not, the word 'nigga' being particularly problematic.

Grounded theory analysis of the tweets showed extremely narrow rhetoric that could be summarised in a single theoretical concept: the defence of the in-group.

This chapter presents a discussion of these findings in the context of the theoretical framework presented in Chapter 4 and the research questions stated in Chapter 1. This chapter is structured according to the research questions, followed by discussion of the original contribution of the thesis, limitations of the research and opportunities for future work.

8.1 Automated Racist Tweet Identification

The first broad question addressed by the research (with its concomitant sub questions) was:

- Q1: Is it possible to have an efficient, accurate and reliable automated racist tweet identifier?
  - Q1-1: how can reliability, efficiency and accuracy be measured?
  - Q1-2: what are the current approaches to automated racist tweet identification, and can they be improved upon?
  - Q1-3: what data do the current approaches to automated racist tweet identification use as input, and are there any other possibilities?

Q1 can be answered in the affirmative: there are a number of examples in the literature of reliable, efficient and accurate automated racist tweet identifiers, which are nearly exclusively examples of machine learning systems. The aim of this research was to see if these existing machine learning methods could be improved upon, and to systematically review the current methods. The model that performed best overall was SVM, which had the maximum value for all of the metrics: curacy, AUPRC, AUROC and F-score for $\beta = 0.5$ and 1. The second best model overall was NB; it had the second best overall AUPRC, AUROC, and F-scores for both $\beta = 1$ and $\beta = 0.5$, although LR outperformed it in accuracy.

Efficiency of the machine learning identifiers is barely addressed in the literature. The closest any researchers come to discussing efficiency is stating the number of pieces of data collected and analysed. These range in size from a few thousand to hundreds of millions. The research is often framed as ‘big data’ research with a concomitant suggestion that this requires specialised processing. When dealing with big data-scale data, questions of efficiency are very pertinent but hard to answer. This is because handling such levels of data efficiently requires extensive computing power and this is usually performed by utilising a Hadoop/Spark system. Such systems are relatively easily scalable and so efficiency is more a matter of how many machines are included in a cluster. Unlike the rest of the literature, this thesis specifically discusses big data and its handling via Hadoop and Spark, and the Spark machine learning algorithms created were reasonably fast (runs typically took between five and 15 minutes), and so the system can be thought of as efficient, for the sample data used. Since the system was designed as a Hadoop system, larger amounts of data such as those available from the Twitter firehose API can be accommodated fairly easily.

Nearly all[1] of the literature used the metrics: accuracy, AUROC, AUPRC and F-score to measure accuracy. These, however do not evaluate a classifier with respect to how well it performs with unseen data. Cross validation is an evaluation method that overcomes this issue and the reliability of such systems is normally addressed by the use of K-fold cross validation. Most of the literature uses either fivefold or tenfold cross validation and this research used tenfold. K-fold cross validation in effect tests algorithm on K different

[1] The rest use a subset of these metrics.

samples and thus aids in reliability.

Many of the existing solutions compared the use of BOW and Ngrams as features. Results varied but generally Ngrams (especially 1-5grams) were optimal, and indeed the results of this research supported this. It was found that with an SVM classifier, the best combination overall was N5 with either hour or neither as an additional feature. For these two feature sets all of their metrics were better than any of the other feature sets. However the use of N5 was not always beneficial. For example, if hour plus day was used as an additional feature, accuracy and AUPRC were slightly higher for BOW than N5.

With regards to other features there were again mixed results. It was found that none of the features from the literature, that data were available for, were beneficial. Two new features: hour of day and day of week were compared with a baseline of no additional features. Hour of day showed improvements in AUPRC and F-score for $\beta = 0.5$, compared with the baseline. In contrast day of the week performed worse than the baseline for all measures.

While some existing approaches use utc_offset (usually as an indicator of geographical location) as a feature, this research is the first to use hour and day of the tweet as features. It is perhaps surprising that geographical, rather than temporal data, is more often used in analyses, since the geographical information on a tweet is prone to inaccuracy. Additionally the temporal information on a tweet is far more prevalent; every tweet has at least timestamp information, something which cannot be said to be true for geographical information. As discussed in Chapter 7 there is little in the literature to compare these temporal results with. Where temporal features are examined such as in Dadvar et al. (2013), they tend to be used as proxies for geographical information, the time zone of a tweet being used as an indication of the geographical location of the tweeter. This lack of attention in the literature and the results of this research indicate that temporal factors of tweets, and perhaps other social media, might be a fruitful area to explore further.

These features were used as input to algorithms, the most common of which, in the literature, were NB, SVM and LR. Many researchers just used one algorithm, others compared NB and SVM. Two groups of researchers compared five different algorithms: Badjatiya et al. (2017) who compared LR, RF, SVM's, GBDTs and DNNs, although they did not use NB and they did not use any non-textual features, and Davidson et al. (2017) who used LR, NB, DT, RF and SVM but no ANN. Algorithm results were mixed, some researchers found SVM to be the optimum algorithm, whereas others found NB worked better. Those that used ANN, generally found that these performed well, although they were not necessarily compared with SVM. For this research SVM worked optimally, much better than ANN. The wide variety and results for the different algorithms calls into question research that only uses one algorithm, especially if this was not SVM.

8.2 Automated Identification of Influential Racist Twitter Accounts

The second broad question addressed by the research (with its concomitant sub questions) was:

- Q2: If such an identifier can be created, how can this further the ability to understand and identify influential racist Twitter accounts?
  - Q2-1: how should influential be defined?
  - Q2-2: what are the current approaches to identifying influential accounts, and can these be applied to racist Twitter accounts?

Although there is an extensive literature on racist *tweet* identification there is no such literature on the identification of racist *tweeters*. So it was necessary to create an automated system to detect them. This was accomplished by analysing the racist tweets identified by the initial machine learning process, and using aspects of these in further machine learning processes to identify racist accounts. The first step in this process was to answer Q2-1 and determine how influential should be defined. All the standard measures of influence on Twitter are used to identify prolific accounts and/or accounts that are highly connected by some measure within the Twitter network. Neither of these types of measures were suitable for this research since the accounts of interest were ones that were not that prolific (the maximum number of original tweets sent by the automatically identify racist accounts was in the hundreds), neither do they necessarily have that many connections.

Instead other measures were successfully used to identify such accounts: the total number of original tweets from an account, *oCount*, the number of distinct tweets that were retweeted from an account, *rDistinctCount*, and the total number of retweets of an account's tweets, *rCount*.

These were run with a SVM model which was trained on the data split 70:30 training to test data and run with tenfold cross validation. The metrics of the model were an accuracy of 0.87, precision=0.68, recall=0.66, and F-score=0.67 for both $\beta = 1$ and $\beta = 0.5$.

The D2 dataset also had a retweet dataset extracted from it, with the same variables as mentioned above. The model was then run on this D2 dataset. A random sample of 200 accounts was drawn from the predictions and their metrics were: accuracy of 0.80, precision=0.63, and recall=0.60, and F-score=0.61 for $\beta = 1$ and 0.62 for $\beta = 0.5$.

A similar procedure was performed, with the results from the grounded theory analysis. The set of focused coding codes were used as categorical input data. The model was again trained on the data split 70:30 training to test data and run with tenfold cross validation. The metrics of the model were an accuracy of 0.82, precision=0.69, recall=0.70, and F-score=0.70 for $\beta = 1$ and 0.69 for $\beta = 0.5$.

This model was then run on the same random sample of 200 accounts from D2 in order to evaluate the grounded theory machine learning predictions. Each of the 200 accounts was coded in the same way as the earlier grounded theory analysis. Their metrics were: accuracy of 0.84, precision=0.60, recall=0.57, and F-score=0.61 for $\beta = 1$ and 0.59 for $\beta = 0.5$. As can be seen the accuracy was worse for the grounded theory analysis compared to the retweet data, but the other metrics were better.

Finally both sets of features were combined, the retweeted data and the grounded theory data both used as features. The metrics of the model were an accuracy of 0.88, precision=0.73, recall=0.71 and F-score=0.72 for both $\beta = 1$ and $\beta = 0.5$. Again the model was tested against the D2 dataset and the metrics were: accuracy of 0.88, precision=0.64, recall=0.66 and F-score=0.65 for both $\beta = 1$ and $\beta = 0.5$.

So it can be seen that both sets of features performed well, but the combined features were the best performing input. These are interesting results: both retweeting data and grounded theory data show promise in the prediction of racist tweeting accounts.

8.3 Qualitative Analysis, Criminology, Psychology and Racist Tweets

The third broad question addressed by the research (with its concomitant sub questions) was:

- Q3: Can qualitative analysis, criminology and psychology be used to further the understanding of racist tweets?
  - Q3-1: what themes emerge from a qualitative analysis of racist tweets?
  - Q3-2: can criminological and psychological theories, such as Routine Activity Theory (RAT), be applied to racist accounts and tweeters?

Of course, the first question can be answered in the affirmative, qualitative analysis and criminology are both rich theoretical tools that can be used to analyse textual data. The short nature of tweets makes interpretation difficult compared with, say, narrative analysis of interviews, but, on the other hand, the vast amount of data available means that the qualitative analysis can be performed on a rich seam of data. Similarly criminological or psychological analysis of the tweets is hampered by the brevity of the texts, but is aided by the numerous examples available.

The language used in these tweets is similar to other research that has shown people being disinhibited in their language on the net. If there is a disinhibition effect, then it might be hypothesised that people tweeting late at night might be more disinhibited due to alcohol consumption and it would therefore be expected that there would be a correlation between amount of racism and time of day, particularly in the hour or two after midnight, traditionally the time when people make their way home after visiting a pub. However the data does not support this. If Pred R and Pred NR are compared both follow similar patterns, with tweets decreasing from midnight to around 4 or 5 in the morning, then the tweet activity increases to a maximum around 9 or 10 at night. Between midnight and 2 the racist tweets show a more steep decline to the nonracist tweets, contrary to what would be expected under the hypothesis. For Inp NR peak day for tweeting is Friday. If alcohol is a factor, then this is what would be expected of racist tweeting, since Friday is the busiest pub day. However for Pred R, which is a much larger dataset, the peak day is Thursday, contrary to what the hypothesis suggests. This is interesting since it suggests that racist tweets are dissimilar to for example, alcohol related violence, which correlates with the time period of an hour or two after midnight, related to public house and nightclub closings (Bromley and Nelson, 2002). While, of course, racist tweeting and alcohol-related violence are very different behaviours, if the main tenet of control theory holds, that is a lack of self-control makes deviant behaviour more likely, it would be expected that both behaviours would be more prevalent in the hours after midnight and on Friday and Saturday, which is not the case with the results of this research.

Qualitative analyses of racist talk have been mostly confined to white power music and far right organisations and political parties (Simi and Futrell, 2015; King, 2014; Martinez Jr and Selepak, 2014; Ben-David and Matamoros-Fernandez, 2016). Similar themes that emerged from these analyses include the ideas of the protection of the homeland, and calls for various kinds of ‘others’ to be expelled. This kind of rhetoric was rare in the samples analysed for this research, but the tweet:

Get out my country you paki ciggs! `https://t.co/g7NOXGmnhu` is an example of this type of aggressive defence of the 'homeland'. However this is not the norm, and there were surprisingly few uses of the words: 'country' or 'nation', even in the racist tweets. However the grounded theory analysis showed that virtually all of the racist tweets, could be seen to be part of a single theoretical grouping, the defence of the in-group. Tweets such as,

> @ancestors ok kys nigger,

were far more common in the dataset. Tweet such as this show an aggressive attack on a person via their ethnicity and so can be theorised to be a defence of the in group, in this case presumably whites.

The former tweet is also an example of a tweet that is unlikely to have been sent 'in the heat of the moment'. As far as can be seen, it is not a spur of the moment reaction, and required, at least a small amount of research, to include the URL. In contrast there were a number of angry tweets sent during sporting events, such as an England international football game during Euro 2016, that show perhaps more expressive and less descriptive slurs, such as the following:

> THAT WAS A FUCKING PENALTY WHAT IS THIS SHIT YOU YANK CUNTS.

This reference to 'yanks' is rare in the sampled data, and it refers to an American referee officiating an England game. There are no other examples of Americans being referred to as 'yanks', and the use here appears to be very much 'heat of the moment', suggesting that its creator is unlikely to slur Americans except in situations such as this. This suggests that, in terms of Crandall and Eshleman's (2003) justification-suppression model, for the latter tweet justification factors are momentary but very powerful, and it is possible that suppression factors were ignored. For the former tweet, while justification factors may also be strong here, it seems that suppression factors are lower, and the balance between justification and suppression is more carefully calculated.

Similarly from a control theory perspective, the second tweet is indicative of low self-control, whereas the first is not. There is considerable evidence in the data of correlates of low self-control, such as swearing and the use of capital letters for emphasis. There were, for example, in the predicted racist retweets dataset, four swear words in the top 12 most common words, or top seven if stopwords were excluded. Thus, for these types of tweets, lack of self-control may be an important determining factor.

Whether the tweets emanate from a lack of self-control, is related to whether rational or irrational processes were present in the decision to send a racist tweet. A lack of self-control may be indicative of an irrational choice being made and vice versa. Crandall and Eshleman (2003) argued that when theorising about prejudice, rationality should be avoided because it is impossible to determine whether an actor is acting rationally or not, and that the processes involved in prejudice, are the same whether the prejudice is performed rationally or irrationally. It is indeed, of course, impossible to determine whether tweets were sent rationally or irrationally. However the vocabulary and form of the tweets suggests the angry and spontaneous nature of many of them. If this can be linked with irrationality then the basic tenets of RAT do not apply well to racist tweeting in many instances. However, Yar (2005) argues, expressive and other non-instrumental crime that might be considered to be irrational, can still be said to have such a rational element since emotional responses may be rational reactions to certain situations, and so RAT may still apply to racist tweeting.

RAT contends that crime is not a random event, instead it occurs in a disproportionate amount in areas that are criminogenic, because they either have a high number of motivated offenders, have a high number of suitable targets or lack capable guardians, or some combination of these. If the racist tweets observed are analysed in light of this, the vast number of racist tweets and tweeting accounts supports the contention of RAT that ‘criminal’ events on Twitter are occurring because of the crime triangle: one or more of there being a high number of motivated offenders, suitable targets or lack of capable guardians. In the 79 days that it took to collect the dataset, D1, there were 462,145 accounts that sent at least one racist original tweet or retweet (these are from the Pred R dataset, and so are predicted, rather than actual racist tweets). This is a large number of potentially racist accounts, from only 1% of Twitter's output, the streaming API, and so there are indeed a high number of motivated offenders on Twitter.

The question of suitable targets is more complex. While it might be thought that there are a multitude of suitable targets since, in theory, any Twitter user can be exposed to any tweet, and twitter users number in the hundreds of millions, the number of users that actually view a racist tweet, is likely to be considerably fewer. However there are mechanisms in Twitter, such as keyword searching, that mean it is likely that many users will inadvertently receive racist tweets. Therefore the suitable target dimension of the crime triangle is also satisfied. Even if very few of the racist accounts are followed by other users, there are still many suitable targets on Twitter compared with the off-line world.

RAT contends that suitable targets follow the model of target attractiveness corresponding to the acronym VIVA (value, inertia,[2] visibility, and accessibility), which are rationally assessed by offenders as to whether a target is suitable for them, and VIVA's applicability to racist tweets can be discussed in light of the results. Most of the racist tweets were attacks on others, and so their value might be in providing a release for their anger and frustrations. They might give a sense of security, that attacking the outgroup can provide. The tweeter might be trying to impress or repel other users. There was, however, little evidence of a 'call to arms' or content that is intended more directly to mobilise others.

The visibility of messages on Twitter is unlimited, but it is possible to place restrictions on what is viewed in a Twitter stream, for example by changing the country setting of an account to Germany, where there is a proscription on neo-Nazi rhetoric (MacGuill, 2017). However the streaming API used to create the sample for this research does not limit its content.

The final dimension of VIVA, Accessibility is the same as visibility on Twitter. If a message can be seen it has already been accessed.

[2] Inertia has little applicability in an online environment, and so will not be discussed.

So the acronym VIVA has some applicability to racist tweeting, but of more practical import is whether the automated systems created can be used for capable guardianship.

8.4 Capable Guardianship and SADRTA

The first two dimensions of the crime triangle are, of course, important, but with respect to this research the most important dimension is that of the capable guardian.

Whether Twitter's efforts in identifying racist tweets is a capable guardian or not is debatable. Capable guardianship requires visibility, and so users must be aware that Twitter is 'policing' them. References to Jack not being happy in relation to a racist tweet, show there is some awareness amongst users that Twitter is monitoring its content. Whether the visibility of Twitter's racism detection processes is sufficient for capable guardianship needs to be tested further.

Capable guardianship also needs to be effective, if it becomes known that attempts at guardianship are cursory or ineffective in other ways, then they no longer have the power to dissuade potential offenders. From the data some racist accounts had been temporarily blocked by Twitter, whereas others had not. Whether Twitter is meeting the requirement of a capable guardian is debatable.

In 2017 it was argued that social media lacks true capable guardianship (Navarro et al., 2017), But there has been a shift towards the introduction of guardianship, capable or not, with the introduction of, for example, Twitter's 'black box' self policing. This has occurred in response to both negative publicity regarding extremism on social media platforms discouraging use of the platforms and also legislative calls from bodies such as the EU to clamp down on extremist speech. Whether the system SADRTA is an effective capable guardian is an important question and can be addressed by unpacking the dimensions of what a capable guardian is in light of RAT. There is little research on the guardianship dimension of RAT (Moule Jr and Powers, 2019) and what there is usually discusses other forms of cybercrime rather than hate speech. Bossler and Holt

(2009) discuss three forms of guardianship related to both the real world and online in the form malware-related cybercrime: physical guardianship, social guardianship and personal guardianship. Physical guardianship relates to a physical deterrent which might be locks in the real world, or anti-virus software online. Social guardianship refers to influence of others on behaviour, such as peer pressure which could be exerted in real world situation, or online. Personal guardianship concerns the awareness of individuals as to the risks of their behaviour in relation to either real-world online environments.

In relation to hate speech and Twitter physical guardianship can be performed by Twitter's efforts at automatically removing racist tweets and accounts and SADRTA. Guardianship works best when there is an awareness of the guardian's presence, as has been shown there is, or was, some awareness of Twitter's efforts at policing their platform, since this was newsworthy in 2017 when there was publicity regarding the EU's concerns over the amount of hate speech on social media. However there is little ongoing publicity and even if there is it is possible that a large number of users of Twitter would be unaware of it. If SADRTA is used, it might also gain some publicity, but this is unlikely to be sufficient for it to be sufficiently effective in creating an ongoing awareness in the majority of Twitter users.Instead perhaps what is needed is a visual reminder that an automated system is running, or likely to be used to analyse Twitter conversations ex post facto. This might be in the form of a visual identifier, that draws the eye on the Twitter platform, perhaps a red flashing symbol of some kind. This visual reminder of the presence of guardianship on Twitter would hopefully deter some users from posting extremist content. So while SADRTA works retroactively in identifying racist tweets and accounts, visual or other signalling of its presence can cause a proactive guardianship, perhaps somewhat akin to a police car driving by in the real world.

The social element of guardianship is perhaps less relevant to SADRTA, although it might come into play, for example in forum discussions when questions are asked relating to possible consequences of extremist posting. If there is knowledge of SADRTA in such an online community, the more a user interacts with such a community, the more likely they are to be aware of SADRTA and perhaps they would change their behaviour accordingly. While it would be nice to assume that this means they would post less, there is always the possibility that they would instead just tried to 'game the system' and change the content of the messages to avoid being flagged by SADRTA. Consequently there is a need for continuous improvement in SADRTA's recognition of what is and what is not racist speech. Fortunately machine learning is ideally suited to perform such ongoing improvements in prediction. SADRTA would need to be modified to include new data as racist tweets and tweeters are encountered and labelled, samples of these would need to be checked by humans and flagged to provide new input data to SADRTA's machine learning algorithms. Any new flagged data should improve SADRTA's racist detection models. The ability of machine learning systems to 'learn' from new data is what makes them better at handling such problems as opposed to say, decision rule based systems, which would need to be recoded with different rules based on new data situations that they encounter.

Personal guardianship is related to an individual's awareness of the risks involved in their behaviour, so is the result of a complex interaction of different influences and characteristics of the individual. Internal moral codes would contribute to personal guardianship, as well as external influences such as those gained in social situations, meaning that social guardianship influences personal guardianship. It is also influenced by physical guardianship since a person's knowledge of risk would be related to their perception of any deterrent measures related to their behaviour. How much of an influence systems such as SADRTA have on personal guardianship is impossible to know without the system being in place, and users discussing their behaviours in relation to it. As already mentioned there is (or there was in 2017) some knowledge of Twitter's desire to reduce racist speech on its platform, it remains to be seen whether this knowledge is still at the forefront of the minds of most users of Twitter.

SADRTA is potentially a capable guardian for Twitter. It is not currently visible to Twitter users, and so does not meet the traditional definition of capable guardian, although this could be changed with publicity if it was being used by law enforcement. It does however meet the capability requirement, it performed well at identifying both racist tweets and tweeters, identifying certain racist accounts that Twitter did not. Unlike Twitter's secretive efforts, its details are publicly available, and it will be made available for public use. This transparency will aid in its role as capable guardians, since any publicity for this research means Twitter users are more likely to be aware of the scrutiny placed on their tweets.

While Twitter may or may not be capably guarding the target, the other controllers: the handler of the offender and the manager of the place of RAT 3 (Eck and Madensen, 2015), are largely absent in relation to racist tweets. The offender may have family or friends that attempt to limit their access to the internet and hence Twitter, or they may have a court order restricting their online activities, however the ease of access to the internet and Twitter mean any restrictions are likely to be easily circumvented. If the 'place' of the crime triangle, is the internet rather than Twitter, then there are managers, in some sense at least, of the place. Access to the internet usually requires a username and password combination, as do most websites. Whether these can be thought to be management is an interesting question.There is some research in the off-line world on the effect of place managers but research on online place managers is minimal, and what there is examines cyber trespass (Remrey, 2016). From a cyber trespass viewpoint is clear that physical control such as user password access do have a deterrent effect and so can be thought of as effective place managers. However much like there are ways of avoiding place managers in the real world there are also ways of bypassing cyber place managers, such as using stolen login credentials. Of course users would be logged in to Twitter, so this type of online place management is not relevant to SADRTA. Account and login credentials are not even necessarily useful to law enforcement in investigations, since it is trivial to create an anonymous account to access services such as Twitter.

RAT 4 introduced the idea of super controllers which are networks that the other controllers are part of, which provide incentives for controllers to reduce crime (Eck and Madensen, 2015). The warnings from the European Commission of retaliation if internet entities such as Twitter do not reduce hate speech on their platforms, could be used as evidence of a super controller network in action. However Sampson et al.'s (2010) notion of super controllers is something of a catchall definition, with 10 different kinds of super controllers that can be formal, diffuse or personal. It is difficult to think of an individual or organisation that is not controlled by super controllers, whether the individual or organisation is a handler, manager or capable guardian or none of these things. Despite this it is of interest to discuss how SADRTA might be situated with respect to super controllers. If SADRTA is used is not working in isolation but instead as part of a network of software, managers, and other actors that are interested in its ability to reduce the incidence of racist speech on Twitter. It would seem prudent that the use and development of any such software was not done in isolation, but instead linked to efforts on other social media platforms, such as Facebook. A racist speech super controller network might include combined efforts between such platforms and controlling the problem. Of course getting such large organisations to work together provides its own challenges, for example it may be deemed as not worthwhile since platforms such as Twitter and Facebook may be too technically different from one another for a joint effort to be worthwhile. However even if there is no explicit agreement, it is the norm these days for creators of software to share their code by placing it on repositories such as Git hub, that is making it *open source* and allowing anybody to adapt the code for their own purposes. So open sourcing in a way might be argued to be acting in the interests of a super controller. The political dimension, such as European Union requirements and directives, will also be part of such a super controlling network that SADRTA would be situated within.

RAT does not explain the disproportionate victimisation of blacks seen in the tweets, nor does it consider the motivations of offenders. Cyber-Routine Activities Theory (CRAT, Choi and Lee, 2017) integrates the concepts of LET and RAT to computer crime victimisation, arguing that differences in online lifestyles and digital guardianship are correlated with online victimisation. From this it might be hypothesised that back people have online lifestyles that mean they are more likely to be victimised by racist tweeters. This is highly unlikely, and suggests that CRAT does not explain racist tweeting well. In fact, since tweets are atomic pieces of discourse, the notion of who is being victimised is at issue. While tweets can be directed at specific users, many are not and these, it might be argued, are victimising an entire ethnicity. It could also be argued that the victimisation is more widespread, since racist tweets are often perceived as unpleasant whether they are directed at the ethnicity of the reader, or not.

Due to the ambiguity of its use the word ‘nigga’ was particular problematic for both human and machine labelling of tweets. There is generally a consensus that often ‘nigga’ is used as a positive identifier of the self, although whether this is acceptable is another matter, since it can be argued that using it in this way helps reinforce power differentials (Andrews, 2014). It is also used pejoratively as a variation on ‘nigger’.

An example to liilustrate the difficulty, of handling this word, is the most popular retweet in the poular racist tweets dataset was:

> Why black people don’t swim , nigga almost drowned and tried to play it off ??? https://t.co/kihaADVhBb,

which was retweeted 20,465 times, It is impossible to determine how each of the participants: creator, retweeter, target of the tweet, and the audience of the tweet react to its content. The word ‘nigga’ might have been used by the original sender of the tweet, to show affection or solidarity with its target, or it might be used in the same form as ‘nigger’, that is its use was intended to be racist.

This is a major challenge in identification of racist tweets, due to ubiquity of the word in the dataset. ‘Nigga’ Was the most common word in the popular racist tweets dataset, with 2,129 uses. The, usually, less ambiguous ‘nigger’ was second with 1,835 uses.

Machine learning is a good tool to handle this type of issue, since it can learn from an initial input of human-labelled instances. However machine learning models are only as good as the data that they use as input. Since the word ‘nigga’ presents problems for human interpretation, the quality of the input data cannot be guaranteed.

For this research the input data was hand annotated by a number of researchers. To determine the interrater reliability of the researchers, the same three datasets were given to each of the annotators. These datasets were 100 tweet samples containing the words: ‘nigga’, ‘nigger’ and ‘wop’ and were rated at the beginning, middle and end of the annotation process respectively. Krippendorff’s α was 0.60, 0.92, and 0.90 for ‘nigga’, ‘nigger’ and ‘wop’ annotations respectively. The low value of α for ‘nigga’ in relation to the others, may be due to practice effects, but it seems likely that it is at least in part due to the ambiguity of the word meaning humans struggle to agree on its usage.

One solution to this is to analyse more instances of tweets containing the word, and to see if particular sentence structures can be utilised in helping determine whether a tweet is racist or not. For example, discovering patterns in the tweet data, such as the use of the word ‘gay’ as an adjective preceding the word ‘nigga’ always occurred with the addition of the word ‘ass’, may aid in human annotation of the tweets by allowing groupings of particular forms of sentence to be treated in the same way, with respect to annotation.

Despite these challenges, SADRTA is likely to become more adept at handling the word ’nigga’ since machine learning systems, are likely to improve with the more data they encounter. As new labelled examples are entered into the system, the models the system produces are based on more information, the more data the system encounters more accurate its predictions are likely to be. While the labelling of such tweets is also challenging, the ability of SADRTA to learn from such data, is perhaps the only practical way to tackle this issue.The handling of this word highlights the importance of input data into a machine learning system. As discussed elsewhere, for machine learning to perform optimally, the selection of features is crucial. Although it is possible to use statistical analysis to aid in the choice of features, much of the machine learning literature, particularly in the exploratory phase such as the use of systems like SADRTA, uses an approach that subjectively identifies the likely most relevant features. With datasets such as those used in this research, with relatively small numbers of potential features, this is a sensible strategy, although much of the published research does not do this systematically. SADRTA, on the other hand, was developed by identifying as close as possible to the full set of potential features from the data. Unlike other researchers in this area, focus was also given to meta data and whether it provides improvements to the machine learning’s predictive capability.While the addition of meta data as features did not have a significant impact on the performance of the machine learning, there were some improvements when hour of day was used as a feature, suggesting that the submit data might be a fruitful area of research. As seen in Chapter 7 the addition of meta data features did not provide the benefits seen by Dadvar et al. (2012) and Dadvar et al. (2013) although their comparisons were with a different baseline. The selection the baseline may be the key factor here, since for this research the baseline included all the text of the tweets in the form of Ngrams, whereas Dadvar et al. (2012) and Dadvar et al. (2013) used only a subset of the text as baseline. Most other researchers used BOW or Ngrams as baseline, and it is unfortunate that the only systematic comparison of meta data features did not.

8.5 Limitations

As is the case with any research, the methodology, and analysis and interpretation of the results are subject to limitations.

Despite the use of K-fold cross validation in this research, it is possible that the models created may be overfitting the data, and possibly poorly model any new data they encounter. This is particularly a worry for the accounts' machine learning processes, since there were only 250 labelled accounts as input. This small sample may mean that the model is being overfit to a very specific small sample. That being said, there are instances in the literature of such small samples being used to determine racist tweets.

The focus of the research was on what was called 'personal racism', that is pieces of text that explicitly contain racist terms. The reason for this was pragmatic: the aim was to collect a large number of potentially racist tweets. However the downside of this was that the data would exclude any implicitly racist tweets that do not contain any racist words.

When handling the field *utc_offset*, any tweets without a value for this field were removed from the analysis. This may introduce bias into the data, since a proportion of those removed may have not set this value since their time zone matched the default one. Additionally, like all the Twitter data in this research, an assumption was made that the data was correct. Since there is no access to Twitter's servers other than via the API, there is no way to test the veracity of the data.

Most of the analysis in this research was performed using each tweet as a stand-alone piece of data. This was done because of the difficulty and complexity of analysing conversations as a whole.

Use of the machine learning systems such as SADRTA will always have its own limitations. Machine learning itself is a process that predicts future events from limited data. While machine learning is certainly flavour of the month at the moment, and there is publicity around its use in many fields and many different applications, it is not always successful or even applicable to every task for which it is used. While SADRTA, and other systems in the literature, have had some success in predicting racist tweets, it is unlikely that machine learning will ever be 100% accurate in such a domain, and even the most successful predictive algorithms currently, such as SADRTA, only have accuracy in the 80 to 90% range. Additionally as is discussed elsewhere, accuracy is not the be all and end all of prediction, both precision and recall need to be examined, and for the best predictors in this domain, precision and recall rarely rise above 75%. SADRTA has only provided incremental improvements on existing systems, and informal testing of the system during development, suggested that it was difficult to obtain significant improvements in evaluation metrics. Large increases in the amount of input data did not provide the expected increases in accuracy, precision and reliability.

As well as limitations in the accuracy precision and recall of SADRTA, there are also limits as to the applicability of conclusions drawn from its outputs. Some of the analysis presented herein is based on SADRTA's predicted racist tweets data. While sampling showed that SADRTA does a reasonable job of predicting such tweets, there will inevitably be false negatives and false positives within such a dataset. Hence analyses based on these data can only be tentative as they will be biased by both racist tweets ignored by SADRTA (false negatives) and nonracist tweets marked as racist by SADRTA (false positives).

There are also limitations on the use of the grounded theory data. Grounded theory is by its nature subjective, and there is always the possibility that other coders would draw different conclusions about the data, and generate different theoretical concepts from it. The use of the grounded theory data, and racist tweet metrics to predict racist accounts also has its limits. While these metrics worked well in the limited sample available, it would be necessary to compare other metrics and grounded theory results and other datasets to see if they gave similar results. That is would 'any old data' provide similar results, there being nothing special about the metrics chosen or the grounded theory data. There was some informal testing of this which suggested it was not the case, that is the metrics chosen and the grounded theory data did provide useful information and make the predictions more accurate, with better precision and recall, but this is only a very tentative conclusion.

8.6 Original Contribution

This research provides an extensive review of the current machine learning methods for identifying racist tweets. All of the algorithms utilised in the literature are compared, along with most of the features that have been used. There is no other work in the literature that compares all these algorithms and matches their efficacy with such a large set of features. The algorithms were tested against a hand annotated racist tweet corpus created for the research consisting of 84,000 labelled tweets. This corpus is one of the largest such corpuses created.

A new automated system, SADRTA, that automatically identifies both racist tweets and tweeters was created. This system comprises novel machine learning techniques and ran on a Spark/Hadoop cluster, such systems are very limited in the literature and this is a novel methodology that builds Hadoop cluster from scratch and runs Spark and Hortonworks Hadoop on it. It also uniquely uses grounded theory data as input to machine learning routines aimed at identifying racist tweet accounts.

While some existing approaches use utc_offset (usually as an indicator of geographical location) as a feature, this research is the first to use hour and day of the tweet as features.

None of the literature after identifying racist tweets makes the seemingly obvious next step to identifying racist accounts. This research utilises machine learning algorithms and features from both a numerical analysis and a grounded theory analysis of the popular racist retweets, to identify potentially racist accounts.

The literature is sparse in relation to methodological and theoretical linkage between big data and criminology. This thesis addresses this issue, in particular methodological considerations of utilising Hadoop and Spark are detailed. For this a new definition of big data is given, one which focuses on processing of data, rather than the volume, velocity and variety of most feature based definitions.

The thesis argues that the main difficulty facing both human and machine labelling of racist tweets, is the ambiguity of the use of the word 'nigga'.

This thesis provides first use of grounded theory in analysing racist tweets.

8.7 Future Work

There is considerable evidence of intersectionality in the data. The work on intersectionality and racist speech is in a nascent stage (Burnap and Williams, 2016) and it would be interesting to test the models further with respect to the interaction of different forms of abuse.

As discussed in Chapter 3 Rattansi (2007) questioned the dichotomous identification of racists versus non-racists. They argued that there needs to be a spectrum of racism and that this might mean that currently there is too much focus on acts of extreme racism. This research is indeed focused on what might be called 'acts of extreme racism', the data that has been analysed has a least one thing in common, that is it contains an ethnic slur. This means it is a very overt form of racism and so is likely to fit into the category of ‘extreme racism’. Despite this, even within these data there are levels of racism, or at least levels of unpleasantness. For example the tweet:

> @Cluhrk get fucking cancer you retard faggot nigger gay cigg idiot,

is likely to be seen as more extremely racist and unpleasant, than the tweet:

> @ShameADriver i give credit to the disabled Gypo who leapt out of his well laden van in Boston today `https://t.co/Duy4r2f09i`.

It would be interesting to see if a classifier could be trained to generate a probabilistic level of racism and/or offensiveness of the tweet. Such a classifier might be able to help with the difficulty of handling the word ‘nigga’ if enough examples of its use are annotated on a scale of racism and offensiveness.

In this research all the data used from Twitter is historical data, i.e. it occurred in the past.[3] The main reasons for this are that the data was collected before the analytical programs were written, and that a large amount of data was required to test whether the system could handle big data scale data. Of course, Twitter generates huge amounts of data every second, and ideally it is this current streaming data that would be analysed. However all the procedures illustrated herein, are equally applicable to streaming data as they are to historic data. Spark can easily handle data streamed from Twitter, and the ML techniques used could equally be applied to streaming data.

The data found in this research contain a large number of expletives and was very abusive in nature. This indicates low self-control on the part of the tweeter, and it would be interesting to explore the correlation between low self-control and these types of tweets further.

[3] Although technically it might be argued that as soon as *any* data is created it becomes historical data, the distinction is that the data collected was stored to disk, then programs were written to analyse it. The programs created as part of this research could be converted to analyse *streaming* data, that is data that is generated continuously. Once the data is ingested into the system, it will have the same format as the historical data, and the same processing would occur on it.

The time of day data suggests that racist abuse online differs from offline crime and may be fruitful to theoretically explore the link between time and racist tweeting further.

Criminals are like anybody else, in that they perform most of their activities within their activity and awareness spaces (Brantingham and Brantingham, 2015). It can be argued that such spaces apply to the online world as well, and racists may have an activity space which encompasses Twitter, Facebook and other sites. This research is focused on Twitter but it would be interesting to see how online racists moved between sites and how their activities are shaped by the different interactions available in the different forms of online social spaces.

The problems of handling the word 'nigga' require further annotation of sample tweets, an analysis of how to handle various forms of tweets encountered using the word.

The complexity of Twitter conversations (Murthy, 2012) and how to include them into the racist tweet identifier could be explored.

Finally, Twitter's role as a capable guardian, and the efficacy of its attempts at removing hate speech from its platform need to be further investigated. The data for this research was collected prior to the announcements around Twitter's identification of hateful accounts. It would be interesting to compare levels of racist tweeting in these data with data collected more recently, which, if Twitter's efforts are successful, should show a reduction in the level of racist tweeting.

Appendix A

Additional Code Listings

List of Code A.1: An example Spark program written in Scala that processes text data.

```
import org.apache.spark.sql.SparkSession
import org.apache.spark.ml.feature.{RegexTokenizer, Tokenizer}
import org.apache.spark.ml.feature.StopWordsRemover
//the following line is for Spark less than 2
// val hc = new org.apache.spark.sql.hive.HiveContext(sc)
val spark = SparkSession.builder.enableHiveSupport().getOrCreate()
spark.sql("SET hive.support.sql11.reserved.keywords=false;")

val dfHiveX = spark.sql("""SELECT input_file_name(), *, user.utc_offset, user.followers_count as fol, unix_timestamp(created_at, ’EEE MMM d HH:mm:ss Z yyyy’) as ut_date,
HOUR(from_unixtime(unix_timestamp(created_at, ’EEE MMM d HH:mm:ss Z yyyy’) - 3600 + COALESCE(cast(user.utc_offset as bigint),0))) as hour,
from_unixtime(unix_timestamp(created_at, ’EEE MMM d HH:mm:ss Z yyyy’) - 3600 +
    COALESCE(cast(user.utc_offset as bigint),0),’E’) as day,
MONTH(from_unixtime(unix_timestamp(created_at, ’EEE MMM d HH:mm:ss Z yyyy’) -
    3600 + COALESCE(cast(user.utc_offset as bigint),0))) as month FROM xnodup
    WHERE racist is NOT null """)

val fractions = Map(false -> 0.2, true -> 1.0)
val dfHive = dfHiveX.stat.sampleBy("racist", fractions, 36L)

spark.sql("set hive.execution.engine=tez;")
val dfHive1 = spark.sql("""SELECT input_file_name(), *, user.utc_offset,
    unix_timestamp(created_at, ’EEE MMM d HH:mm:ss Z yyyy’) as ut_date,
HOUR(from_unixtime(unix_timestamp(created_at, ’EEE MMM d HH:mm:ss Z yyyy’) -
    3600 + COALESCE(cast(user.utc_offset as bigint),0))) as hour,
from_unixtime(unix_timestamp(created_at, ’EEE MMM d HH:mm:ss Z yyyy’) - 3600 +
    COALESCE(cast(user.utc_offset as bigint),0),’E’) as day,
MONTH(from_unixtime(unix_timestamp(created_at, ’EEE MMM d HH:mm:ss Z yyyy’) -
    3600 + COALESCE(cast(user.utc_offset as bigint),0))) as month FROM ynodup
    WHERE racist is NOT null""")

//STOPWORDS
import org.apache.spark.sql.DataFrame
def RemoveStopwords (dfin : DataFrame) : DataFrame = {
    val stopwords = sc.textFile("/media/ed/Seagate/data/stopwords.txt")

    val tokenizer = new Tokenizer().setInputCol("text").setOutputCol("words")
    val wordsData = tokenizer.transform(dfin)

    // remove stop words
    val remover = new StopWordsRemover().setInputCol("words").setOutputCol("NoStop")
    val dfNoStop= remover.transform(wordsData)
    dfNoStop.select("racist", "NoStop").show(100, false)
    //~ dfNoStop.printSchema()
    dfNoStop
}
val Array(dfNoStop, dfNoStop1)=Array(dfHive, dfHive1).map(RemoveStopwords _)

import org.apache.spark.sql.functions.udf

val urls = "((https?|ftp|gopher|telnet|file|Unsure|http):((//)|(\\\\))+[\\w\\d:#@%/;$()~_?\\+-=\\\\\\.&]*)"

val digits = "\\b\\d+\\b"
val alphanumerics = "(\\b[0-9]+[a-zA-Z]+\\b)|(\\b+[a-zA-Z]+[0-9]+\\b)|(\\b[a-zA-Z]+[0-9]+[a-zA-Z]+\\b)"
val white_space = "\\s+"
val small_words ="\\b[a-zA-Z0-9]{1,2}\\b"
val emojis = ".*[^\u0000-\uFFFF].*"

def noPuncOrURLs(lines: Seq[String]) = lines flatMap { line =>
  "[a-zA-Z@]+".r findAllIn line.replaceAll("((https?|ftp|gopher|telnet|file|
   Unsure|http):((//)|(\\\\))+[\\w\\d:#@%/;$()~_?\\+-=\\\\\\.&]*)", "") map (_
   .toLowerCase)
}

val removeRegexUDF = udf(
    (input: Seq[String]) => noPuncOrURLs(input).filterNot(s => s.matches(urls)
   || s.matches(digits) || s.matches(alphanumerics) || s.matches(white_space)
   || s.matches(small_words)|| s.matches(emojis))
)

// Apply the UDF to change the source dataset
def RemoveURLs (dfin : DataFrame) : DataFrame = {
val dfout= dfin.withColumn("Notquite", removeRegexUDF('NoStop))
dfout
}
val Array(dfNoURL, dfNoURL1)=Array(dfNoStop, dfNoStop1).map(RemoveURLs)
// Install language models
val version = "3.6.0"
val model = s"stanford-corenlp-$version-models" // append "-english" to use the
    full English model
val jars = sc.asInstanceOf[{def addedJars: scala.collection.mutable.Map[String,
    Long]}].addedJars.keys // use sc.listJars in Spark 2.0
if (!jars.exists(jar => jar.contains(model))) {
  import scala.sys.process._
  s"wget http://repo1.maven.org/maven2/edu/stanford/nlp/stanford-corenlp/
   $version/$model.jar -O /tmp/$model.jar".!!
  sc.addJar(s"/tmp/$model.jar")
  sc.addJar(s"/home/ed/.ivy2/jars/$model.jar")
}

import org.apache.spark.sql.functions.{udf, lit}

val mkString = udf((a: Seq[String]) => a.mkString(" "))
val lemmaText:  (Column) => Column = (x) => {lemma(x) }

def MakeLemmas (dfin : DataFrame) : DataFrame = {
    val dfString = dfin.withColumn("NoURL_string", mkString($"Notquite"))
    val dfout = dfString.withColumn("NoURL", lemmaText(col("NoURL_string")))
    dfout
}

val Array(dfLemma, dfLemma1)=Array(dfNoURL, dfNoURL1).map(MakeLemmas)

import com.databricks.spark.corenlp.functions._
import org.apache.spark.sql.types._
import org.apache.spark.sql.Column
import org.apache.spark.ml.feature.NGram

val ngramCreator = udf((xs: Seq[String], n: Int) =>
  (1 to n).map(i => xs.sliding(i).filter(_.size == i).map(_.mkString(" "))).
   flatten)

```

```
spark.udf.register("ngramCreator", https://www.reddit.com/r/television/comments/8ru56y/goliath_season_two_is_a_really_weird_mess/)

import org.apache.spark.ml.feature.SQLTransformer

val ngramer =  new SQLTransformer().setStatement(
  """SELECT *, ngramCreator(Notquite, 5) AS NoURL FROM __THIS__"""
)


val Array(ngramDataFrame, ngramDataFrame1) = Array(dfLemma, dfLemma1).ngramer
```

List of Code A.2: An example Spark program written in Scala that performs ML routines on the preprocessed textual data.

```
import org.apache.spark.ml.classification.LinearSVC
import org.apache.spark.ml.feature.{HashingTF, IDF, Tokenizer}
import org.apache.spark.mllib.util.MLUtils
import com.databricks.spark.corenlp.functions._
import org.apache.spark.sql.functions.udf
import com.databricks.spark.corenlp.functions._
import org.apache.spark.sql.{DataFrame, Dataset, Row, SparkSession}
import org.apache.spark.ml.Pipeline
import org.apache.spark.ml.feature.{VectorAssembler, IndexToString, StringIndexer, VectorIndexer}
import org.apache.spark.ml.evaluation.BinaryClassificationEvaluator
```

```

spark.sql("set hive.execution.engine=tez;")

val data = dfLemma.withColumn("racist", 'racist.cast("String"))
val labelIndexer = new StringIndexer().setInputCol("racist").setOutputCol("label").setHandleInvalid("skip").fit(data)
val hashingTF = new HashingTF().setInputCol("NoURL").setOutputCol("hash-tf").setNumFeatures(20000)
val idf = new IDF().setInputCol("hash-tf").setOutputCol("hash-tfidf")
val hourIndexer = new StringIndexer().setInputCol("hour").setOutputCol("hour-idx").setHandleInvalid("skip")
val dayIndexer = new StringIndexer().setInputCol("day").setOutputCol("day-idx").setHandleInvalid("skip")
val va = new VectorAssembler().setInputCols(Array("hour-idx",  "day-idx", "hash-tfidf")).setOutputCol("features")
val Array(training, test) = data.randomSplit(Array(0.8, 0.2))
val lsvc = new LinearSVC().setMaxIter(5).setRegParam(0.1).setFeaturesCol("features")
val labelConverter = new IndexToString().setInputCol("prediction").setOutputCol("predictedLabel").setLabels(labelIndexer.labels)
val pipeline = new Pipeline().setStages(Array(labelIndexer,  hourIndexer, dayIndexer,  hashingTF, idf, va,    lsvc, labelConverter))

import org.apache.spark.ml.linalg.DenseVector
import org.apache.spark.ml.tuning.{ParamGridBuilder, CrossValidator}
val nFolds: Int = 5
val paramGrid = new ParamGridBuilder().addGrid(idf.minDocFreq, Array(1, 10)).build()
val evaluator = new BinaryClassificationEvaluator().setLabelCol("label").setRawPredictionCol("rawPrediction")
val cv = new CrossValidator().setEstimator(pipeline).setEvaluator(evaluator).setEstimatorParamMaps(paramGrid).setNumFolds(nFolds)
val cvModel = cv.fit(training)

// Save model
cvModel.write.overwrite().save("target/tmp/lsvcModel")

val bestModel = cvModel.bestModel.asInstanceOf[org.apache.spark.ml.PipelineModel]

val predictionAndLabel = bestModel.transform(test).select(’Prediction, ’racist, ’label, ’NoURL)

// Print the average metrics per ParamGrid entry
val avgMetricsParamGrid = cvModel.avgMetrics
println( "avgMetricsParamGrid " + avgMetricsParamGrid)

// Combine with paramGrid to see how they affect the overall metrics
val combined = paramGrid.zip(avgMetricsParamGrid)

// Explain params for each stage
val bestHashingTFNumFeatures = bestModel.stages(3).asInstanceOf[org.apache.
```

```
    spark.ml.feature.HashingTF].explainParams
val bestIDFMinDocFrequency = bestModel.stages(4).asInstanceOf[org.apache.spark.
    ml.feature.IDFModel].explainParams
val bestDecisionTreeDepth = bestModel.stages(7).asInstanceOf[org.apache.spark.
    ml.classification.LinearSVCModel].explainParams
println( "bestModel " + bestModel)
println( "bestHashingTFNumFeatures " + bestHashingTFNumFeatures)
println( "bestIDFMinDocFrequency " + bestIDFMinDocFrequency)

// retrieving the best model’s params
val bestEstimatorParamMap = cvModel.getEstimatorParamMaps.zip(
    avgMetricsParamGrid).maxBy(_._2)._1
println(s"Best params:\n$bestEstimatorParamMap")

//metrics
Timelog.timer("metrics")
import org.apache.spark.mllib.evaluation.BinaryClassificationMetrics

def formatArrayColumn(arrayColumn: Column): Column = {
  concat(lit("["), concat_ws(", ", arrayColumn), lit("]")).as(s"format(${
    arrayColumn.expr})")
}

val result = predictionAndLabel.withColumn("NoURL", formatArrayColumn($"NoURL")
    )
val scoreAndLabels = result.select("label", "prediction").rdd.map(row => (row.
   getAs[Double]("prediction"), row.getAs[Double]("label")))
val metrics = new BinaryClassificationMetrics(scoreAndLabels)
val accuracy = 1.0 * scoreAndLabels.filter(x => x._1 == x._2).count() / test.
   count()

// Precision-Recall Curve
val PRC = metrics.pr

// F-measure
val f1Score = metrics.fMeasureByThreshold.collect()
f1Score.foreach { case (t, f) =>
  println(s"Threshold: $t, F-score: $f, Beta = 1")
}

val beta = 0.5
val fScore = metrics.fMeasureByThreshold(beta).collect()
fScore.foreach { case (t, f) =>
  println(s"Threshold: $t, F-score: $f, Beta = 0.5")
}

// AUPRC
val auPRC = metrics.areaUnderPR
println("Area under precision-recall curve = " + auPRC)

// AUROC
val auROC = metrics.areaUnderROC
```

```
println("Area under ROC = " + auROC)

// best and worst features
import org.apache.spark.ml.classification.LinearSVCModel
val LSVCModel= bestModel.stages(6).asInstanceOf[org.apache.spark.ml.classification.LinearSVCModel]

//print latex line
println(f"SVM & BOW+Hour+Day & $accuracy%.3f &  $auPRC%.3f &  $auROC%.3f &  ${f1Score(0)._2}%.3f & ${fScore(0)._2}%.3f \\\\ ")

//now transform the big data
val data1 = dfLemma1.withColumn("racist", 'racist.cast("String"))
val dfPredictions = bestModel.transform(data1.withColumnRenamed("hash-tfidf", "features"))
```

List of Code A.3: An example tweet showing JSON format.

```
{
   "contributors":None,
   "truncated":False,
   "text":u"nigger hahahahahahha. \U0001f602",
   "is_quote_status":False,
   "in_reply_to_status_id":None,
   "id":666180981775175680,
   "favorite_count":0,
   "_api":<tweepy.api.API object at 0x1a27690>,
   "author":User(follow_request_sent=None,
   profile_use_background_image=True,
   _json=   {
      "follow_request_sent":None,
      "profile_use_background_image":True,
      "default_profile_image":False,
      "id":2175770720,
      "verified":False,
      "profile_image_url_https":"https://pbs.twimg.com/profile_images
    /665778074412171264/0xRrGzL5_normal.jpg",
      "profile_sidebar_fill_color":"7AC3EE",
      "profile_text_color":"3D1957",
      "followers_count":1378,
      "profile_sidebar_border_color":"FFFFFF",
      "id_str":"2175770720",
      "profile_background_color":"000000",
      "listed_count":3,
      "profile_background_image_url_https":"https://pbs.twimg.com/
    profile_background_images/645528085501095936/yQIQJIXY.jpg",
      "utc_offset":-28800,
      "statuses_count":54182,
      "description":"dont kill my high because your low.",
      "friends_count":1769,
      "location":"squad6815",
      "profile_link_color":"000000",
      "profile_image_url":"http://pbs.twimg.com/profile_images/665778074412171264/OxRrGzL5_normal.jpg",
      "following":None,
      "geo_enabled":True,
      "profile_banner_url":"https://pbs.twimg.com/profile_banners/2175770720/1447568838",
      "profile_background_image_url":"http://pbs.twimg.com/profile_background_images/645528085501095936/yQIQJIXY.jpg",
      "name":"ekaayy",
      "lang":"en",
      "profile_background_tile":False,
      "favourites_count":27748,
      "screen_name":"jerickaaac",
      "notifications":None,
      "url":None,
      "created_at":"Tue Nov 05 10:03:20 +0000 2013",
      "contributors_enabled":False,
      "time_zone":"Pacific Time (US & Canada)",
      "protected":False,
      "default_profile":False,
      "is_translator":False
   },
   id=2175770720,
   _api=<tweepy.api.API object at 0x1a27690>,
   verified=False,
   profile_image_url_https="https://pbs.twimg.com/profile_images
```

```
 /6657780744l2171264/OxRrGzL5_normal.jpg",
profile_sidebar_fill_color="7AC3EE",
is_translator=False,
geo_enabled=True,
profile_text_color="3D1957",
followers_count=1378,
protected=False,
location="squad6815",
default_profile_image=False,
id_str="2175770720",
utc_offset=-28800,
statuses_count=54182,
description="dont kill my high because your low.",
friends_count=1769,
profile_link_color="000000",
profile_image_url="http://pbs.twimg.com/profile_images/665778074412171264/
 OxRrGzL5_normal.jpg",
notifications=None,
profile_background_image_url_https="https://pbs.twimg.com/
 profile_background_images/645528085501095936/yQIQJIXY.jpg",
profile_background_color="000000",
profile_banner_url="https://pbs.twimg.com/profile_banners
 /2175770720/1447568838",
profile_background_image_url="http://pbs.twimg.com/profile_background_images
 /645528085501095936/yQIQJIXY.jpg",
screen_name="jerickaaac",
lang="en",
profile_background_tile=False,
favourites_count=27748,
name="ekaayy",
url=None,
created_at=datetime.datetime(2013,
11,
5,
10,
3,
20   ),
contributors_enabled=False,
time_zone="Pacific Time (US & Canada)",
profile_sidebar_border_color="FFFFFF",
default_profile=False,
following=False,
listed_count=3),
"_json":{
   "contributors":None,
   "truncated":False,
   "text":u"nigger hahahahahahha. \U0001f602",
   "is_quote_status":False,
   "in_reply_to_status_id":None,
   "id":666180981775175680,
   "favorite_count":0,
   "source":"<a href="      http://twitter.com/download/iphone" rel="
```

```
nofollow">Twitter for iPhone</a>",
   "retweeted":False,
   "coordinates":None,
   "timestamp_ms":"1447664897756",
   "entities":{
      "user_mentions":[

      ],
      "symbols":[

      ],
      "hashtags":[

      ],
      "urls":[

      ]
   },
   "in_reply_to_screen_name":None,
   "id_str":"666180981775175680",
   "retweet_count":0,
   "in_reply_to_user_id":None,
   "favorited":False,
   "user":{
      "follow_request_sent":None,
      "profile_use_background_image":True,
      "default_profile_image":False,
      "id":2175770720,
      "verified":False,
      "profile_image_url_https":"https://pbs.twimg.com/profile_images
/665778074412171264/OxRrGzL5_normal.jpg",
      "profile_sidebar_fill_color":"7AC3EE",
      "profile_text_color":"3D1957",
      "followers_count":1378,
      "profile_sidebar_border_color":"FFFFFF",
      "id_str":"2175770720",
      "profile_background_color":"000000",
      "listed_count":3,
      "profile_background_image_url_https":"https://pbs.twimg.com/
profile_background_images/645528085501095936/yQIQJIXY.jpg",
      "utc_offset":-28800,
      "statuses_count":54182,
      "description":"dont kill my high because your low.",
      "friends_count":1769,
      "location":"squad6815",
      "profile_link_color":"000000",
      "profile_image_url":"http://pbs.twimg.com/profile_images
/665778074412171264/OxRrGzL5_normal.jpg",
      "following":None,
      "geo_enabled":True,
      "profile_banner_url":"https://pbs.twimg.com/profile_banners
/2175770720/1447568838",
         "profile_background_image_url":"http://pbs.twimg.com/profile_background_images/645528085501095936/yQIQJIXY.jpg",
         "name":"ekaayy",
         "lang":"en",
         "profile_background_tile":False,
         "favourites_count":27748,
         "screen_name":"jerickaaac",
         "notifications":None,
         "url":None,
         "created_at":"Tue Nov 05 10:03:20 +0000 2013",
         "contributors_enabled":False,
         "time_zone":"Pacific Time (US & Canada)",
         "protected":False,
         "default_profile":False,
         "is_translator":False
      },
      "geo":None,
      "in_reply_to_user_id_str":None,
      "lang":"nl",
      "created_at":"Mon Nov 16 09:08:17 +0000 2015",
      "filter_level":"low",
      "in_reply_to_status_id_str":None,
      "place":None
   },
   "coordinates":None,
   "timestamp_ms":"1447664897756",
"entities":{
   "user_mentions":[

   ],
   "symbols":[

   ],
   "hashtags":[

   ],
   "urls":[

   ]
},
"in_reply_to_screen_name":None,
"in_reply_to_user_id":None,
"retweet_count":0,
"id_str":"666180981775175680",
"favorited":False,
"source_url":"http://twitter.com/download/iphone",
"user":User(follow_request_sent=None,
profile_use_background_image=True,
_json=   {
   "follow_request_sent":None,
   "profile_use_background_image":True,
   "default_profile_image":False,
    "id":2175770720,
    "verified":False,
    "profile_image_url_https":"https://pbs.twimg.com/profile_images/665778074412171264/0xRrGzL5_normal.jpg",
    "profile_sidebar_fill_color":"7AC3EE",
    "profile_text_color":"3D1957",
    "followers_count":1378,
    "profile_sidebar_border_color":"FFFFFF",
    "id_str":"2175770720",
    "profile_background_color":"000000",
    "listed_count":3,
    "profile_background_image_url_https":"https://pbs.twimg.com/profile_background_images/645528085501095936/yQIQJIXY.jpg",
    "utc_offset":-28800,
    "statuses_count":54182,
    "description":"dont kill my high because your low.",
    "friends_count":1769,
    "location":"squad6815",
    "profile_link_color":"000000",
    "profile_image_url":"http://pbs.twimg.com/profile_images/665778074412171264/0xRrGzL5_normal.jpg",
    "following":None,
    "geo_enabled":True,
    "profile_banner_url":"https://pbs.twimg.com/profile_banners/2175770720/1447568838",
    "profile_background_image_url":"http://pbs.twimg.com/
```

```
 profile_background_images/645528085501095936/yQIQJIXY.jpg",
   "name":"ekaayy",
   "lang":"en",
   "profile_background_tile":False,
   "favourites_count":27748,
   "screen_name":"jerickaaac",
   "notifications":None,
   "url":None,
   "created_at":"Tue Nov 05 10:03:20 +0000 2013",
   "contributors_enabled":False,
   "time_zone":"Pacific Time (US & Canada)",
   "protected":False,
   "default_profile":False,
   "is_translator":False
},
id=2175770720,
_api=<tweepy.api.API object at 0x1a27690>,
verified=False,
profile_image_url_https="https://pbs.twimg.com/profile_images
 /665778074412171264/0xRrGzL5_normal.jpg",
profile_sidebar_fill_color="7AC3EE",
is_translator=False,
geo_enabled=True,
profile_text_color="3D1957",
followers_count=1378,
protected=False,
location="squad6815",
default_profile_image=False,
id_str="2175770720",
utc_offset=-28800,
statuses_count=54182,
description="dont kill my high because your low.",
friends_count=1769,
profile_link_color="000000",
profile_image_url="http://pbs.twimg.com/profile_images/665778074412171264/
 0xRrGzL5_normal.jpg",
notifications=None,
profile_background_image_url_https="https://pbs.twimg.com/
 profile_background_images/645528085501095936/yQIQJIXY.jpg",
profile_background_color="000000",
profile_banner_url="https://pbs.twimg.com/profile_banners
 /2175770720/1447568838",
profile_background_image_url="http://pbs.twimg.com/profile_background_images
 /645528085501095936/yQIQJIXY.jpg",
screen_name="jerickaaac",
lang="en",
profile_background_tile=False,
favourites_count=27748,
name="ekaayy",
url=None,
created_at=datetime.datetime(2013,
11,
    5,
    10,
    3,
    20   ),
    contributors_enabled=False,
    time_zone="Pacific Time (US & Canada)",
    profile_sidebar_border_color="FFFFFF",
    default_profile=False,
    following=False,
    listed_count=3),
    "geo":None,
    "in_reply_to_user_id_str":None,
    "lang":"nl",
    "created_at":datetime.datetime(2015,
    11,
    16,
    9,
    8,
    17   ),
    "filter_level":"low",
    "in_reply_to_status_id_str":None,
    "place":None,
    "source":"Twitter for iPhone",
    "retweeted":False
}
```

List of Code A.4: Python program, data.py, used to capture tweets.

```

#!/usr/bin/env python

import sys,os
import tweepy
import json

if len(sys.argv) < 3:
    print 'Usage: ' + sys.argv[0] + " <'a', 'b' or 'c'>" + " <'1', '2' or '3'>"
    sys.exit(1)

#pass security information to variables
if sys.argv[2] == '1':
# the following security values have been removed to protect user's accounts
    consumer_key = ''
    consumer_secret = ''
    access_key = ''
    access_secret = ''
elif sys.argv[2] == '2':
    consumer_key = ''
    consumer_secret = ''
    access_key = ''
    access_secret = ''
elif sys.argv[2] == '3':
    consumer_key = ''
    consumer_secret = ''
    access_key = ''
    access_secret = ''

datadir='data/' + sys.argv[1] + '/'
if not os.path.exists(datadir):
    os.makedirs(datadir)

#use variables to access twitter
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_key, access_secret)
api = tweepy.API(auth)

outfile = open(datadir + 'MULTI.txt', 'a+')
if sys.argv[1] == 'a':
    racistwords = ["abbie", "abe", "abie", "abc", "abcd", "abid", "abeed", "abo", "abbo", "ali baba", "alligator bait" , "ann", "annamite", "ape", "apple", "arabush", "armo", "aunt jemima", "aunt jane", "aunt mary", "aunt sally", "banana", "beaner", "beaney", "beni-oui-oui", "bluegum", "boche", "bosche", "bosch", "boeotian", "boerehater", "boer-hater", "boer hater", "bog irish", "bogtrotter", "bog-trotter", "bohunk", "boong", "bong", "bung", "boonga", "boong", "bunga", "boonie", "bootlip", "bounty bar", "brownie", "buck", "buddhahead", "bule", "buffie", "burrhead", "burr-head", "burr head", "camel jockey", "canuck", "charlie", "chee-chee", "chi-chi", "cheesehead", "cheese-eating surrender monkey", "chernozhopy", "ching chong", "chinaman", "chink", "chonky", "chunky", "chunger", "christ killer", "cholo", "chug", "
    coconut", "coolie", "coon", "coonass", "coon-ass", "cracker", "crow", "
    curry-muncher", "cushi", "dago", "dego", "dal khor", "darky", "darkey", "
    darkie", "dink", "dogan", "dogun", "dothead", "dune coon", "eight ball", "
    eskimo", "eyetie", "flip", "fritz", "frog", "fuzzy-wuzzy", "gable", "gaijin
    ", "gin", "gin jockey", "golliwog", "gook", "gook-eye", "gooky", "gora", "
    goy", "goyim", "goyum", "greaseball", "greaser", "gringo", "groid", "gub",
    "gubba", "guizi", "guido", "guinea", "ginzo", "gweilo", "gwailo", "kwai lo"
    , "gyopo", "kushi", "kyopo ", "gyppo", "gippo", "gypo", "gyppie", "gyppy",
    "gipp", "hairyback", "hajji", "hadji", "haji", "half-breed", "haole", "heeb
    ", "hebe", "hillbilly", "honky", "honkey", "honkie", "hori", "hun", "hymie"
    , "ikey", "ike", "iky", "ikey-mo", "ikeymo", "indon", "injun", "jungle
    bunny", "jap", "japie", "yarpie", "jerry", "jigaboo", "jiggabo", "jigarooni
    ", "jijjiboo", "zigabo", "jig", "jigg", "jigga", "jigger", "jock", "jocky",
     "jockie", "jungle bunny"]

elif sys.argv[1] == 'b':
    racistwords = ["kaffir", "kaffer", "kaffir", "kafir", "kaffre", "kuffar", "
    caffer", "caffre", "kalar", "kano", "katsap", "kacap", "kacapas", "
    kharkhuwa", "khokhol", "kike", "kyke", "kimchi", "klansman", "kraut", "lebo
    ", "limey", "lubra", "lugan", "mabuno", "mahbuno", "macaca", "madrassi", "
    majus", "malakh-khor", "malaun", "mau-mau", "mick", "moon cricket", "
    mooncricket", "moskal", "mulignan", "mulignon", "mzungu", "niglet", "nig-
    nog", "nigger", "niger", "nig", "nigor", "nigra", "nigre", "nigar", "niggur
    ", "nigga", "niggah", "niggar", "nigguh", "niggress", "nigette", "nip", "
    nitchie", "neche", "neechee", "neejee", "nichi", "nichiwa", "nidge", "
    nitchee", "nitchy", "northern monkey", "nusayri", "oreo", "paddy", "paki",
    "palagi", "pancake face", "pancake", "papoose", "peckerwood", "pepper", "
    pepsi", "pickaninny", "piefke", "pikey", "piky", "piker", "pinoy", "plastic
     paddy", "pocho", "pocha", "polack", "pollock", "pom", "pohm", "pommy", "
    pommie", "pommie grant", "porch monkey", "prairie nigger", "quashie", "
    raghead", "rastus", "razakars", "redlegs", "redneck", "redskin", "roundeye"
    , "sambo", "sand nigger", "sassenach", "sawney", "scandihoovian", "seppo",
    "septic", "schvartse", "schwartze", "sheeny", "sheep shagger", "sheigetz",
    "shelta", "shiksa", "shine", "shkutzim", "shylock", "sideways vagina", "
    sideways pussy", "sideways cooter", "skinny", "slant", "slant eye", "slope"
    , "slopehead", "slopy", "slopey", "sloper", "slopi", "slopy", "sloppy", "
    smoked irish", "smoked irishman", "soosmar-khor", "sooty", "southern faerie
    ", "southern fairy", "spade", "spearchucker", "spic", "spick", "spik", "
    spig", "spigotty", "spook", "squarehead", "squaw", "squinty", "swamp guinea
    "]

elif sys.argv[1] == ’c’:
     racistwords = set(["tacohead", "taffy", "taff", "taig", "teague", "teg", "
    teig", "tar-baby", "teapot", "teuchter", "thicklips", "timber nigger", "
    tinker", "tynekere", "tinkere", "tynkare", "tynkere", "tynker", "tenker", "
    tinkar", "tyncar", "tinkard", "tynkard", "tincker", "towel head", "touch of
     the tar brush", "turco-albanian", "turk", "twinkie", "ukrop", "uncle tom",
     "wasp", "wetback", "wigger", "whigger", "wigga", "white nigger", "white
    trash", "whitey", "wog", "wop", "yank", "yankee", "yellow", "yid", "yuon",
    "zip", "zipperhead"])

#function to return empty string if string is none.
```

```
def xstr(s):
    return s or ’’

#create an object called ’customStreamListener’
class CustomStreamListener(tweepy.StreamListener):
    global writer

    def on_status(self, status):

        json.dump(status._json, outfile)
        outfile.write(’\n’)
        tweetText= status._json[’text’]
        print tweetText

        for word in tweetText.split():

            q=word.replace("#", "", 1).lower()
            if q in unicoderacistwords:
                print ’found:’, word
                filename=datadir+ q + ’.txt’
                wordfile=open(filename, ’a+’)
                json.dump(status._json, wordfile)
                wordfile.write(’\n’)

    def on_error(self, status_code):
        print >> sys.stderr, ’Encountered error with status code:’, status_code
        return True # Don’t kill the stream

    def on_timeout(self):
        print >> sys.stderr, ’Timeout...’
        return True # Don’t kill the stream

while True:
    sapi = tweepy.streaming.Stream(auth, CustomStreamListener())
    unicoderacistwords = [unicode(i, encoding=’UTF-8’) for i in racistwords]
    sapi.filter(languages=["en"], track=unicoderacistwords)
```

List of Code A.5: Python program, used to remove invalid JSON records from tweet files.

```

#!/usr/bin/env python
import json
from datetime import datetime
import time

start = time.time()
starttime = datetime.strptime(time.ctime(),"%a %b %d %H:%M:%S %Y")
starttext=’ Started at: ’ + str(starttime)

# fname = ’SMALLdata.txt’
# with open(fname, ’r’) as f:
import fileinput
import glob
import os
import sys, time
path= sys.argv[1]
os.chdir(path) # get the folder in the run configuration

# errors.txt one directory above the data, otherwise it uses it as input!!!!
errf=open("../errors.txt", ’w’)
errf.write(sys.argv[0]+ starttext)

results=open("../results.txt", ’w’)
results.write(sys.argv[0]+ starttext+ "\n"+ "\n")

for subdir, dirs, files in os.walk(path):
    for f in files:

        with open(f,"r") as infile:
            with open("../fixedfiles/" + f ,"wb") as output:
                for line in infile:
                    localtime = time.asctime( time.localtime(time.time()) )
                    try:
                        print ’TRY!!!’, infile, line[50:100]
                        tweet = json.loads(line)
                        output.write(line)

                    except ValueError as e:
                        print ’err’
                        errf.write( "\n" + "\n" + ’ERROR: file ’ + str(infile) + "\n"+ "\n" +’id: ’+  str(tweet[’id’])+ "\n"+ "\n"+ ’time: ’
                                    +localtime+’ date: ’+ tweet[’created_at’]+ "\n"+ "\n"+ line)
                        print(e)
                        continue


end = time.time()
duration=end-start
endtime = datetime.strptime(time.ctime(),"%a %b %d %H:%M:%S %Y")
endtext=’ Ended at ’ + str(endtime) + ’, completed in: ’ + str(duration) + ’ seconds’
sys.stdout.write(’\n’+ endtext)
errf.write(sys.argv[0]+ endtext)
results.write(sys.argv[0]+ endtext)
```

```
SELECT A.screen_name, A.id_str, A.oCount, A.rOfOthers_cnt, B.rCount, B.retweetRatio FROM

(select account.id_str AS id_str, account.screen_name AS screen_name,
       sum(case when retweeted_status  is null then 1 else 0 end) as oCount,
       sum(case when retweeted_status  is not null then 1 else 0 end) as rOfOthers_cnt
from predictions_d where prediction=1 and utc_offset is not null
group by account.id_str, account.screen_name) AS A


JOIN


(select retweeted_status.account.id_str AS id_str, retweeted_status.account.screen_name,
       sum(1) as rCount,
       sum(1) / count(distinct retweeted_status.id_str) as retweetRatio
from predictions_d where prediction=1 and utc_offset is not null and retweeted_status  is not null
group by retweeted_status.account.id_str, retweeted_status.account.screen_name) AS B
ON A.id_str = B.id_str
```

List of Code A.6: SQL that gives counts of orginal tweets, retweets of others, retweets and retweet ratio for an account. language

.

Appendix B

Record Counts and Metrics Table

Table B.1: Record counts of the 283 text files.

Filename	Count	Filename	Count	Filename	Count	Filename	Count
abbie.txt	75,807	abbo.txt	1,441	abcd.txt	88,930	abc.txt	1,541,224
abeed.txt	1,662	abe.txt	230,747	abid.txt	19,489	abie.txt	2,731
abo.txt	47,149	annamite.txt	118	ann.txt	991,463	ape.txt	345,857
apple.txt	11,864,652	armo.txt	2,822	banana.txt	1,580,088	beaner.txt	23,762
beaney.txt	298	bluegum.txt	158	boche.txt	614	boeotian.txt	189
bogtrotter.txt	1,109	bohunk.txt	132	bong.txt	337,086	boonga.txt	496
boong.txt	16,061	boonie.txt	7,154	bootlip.txt	48	bosche.txt	268
bosch.txt	150,821	brownie.txt	560,240	buck.txt	621,889	buddhahead.txt	37
buffie.txt	2,809	bule.txt	11,200	bunga.txt	15,360	bung.txt	14,912
burrhead.txt	824	caffer.txt	27	caffre.txt	29	canuck.txt	14,423
charlie.txt	7,803,810	chee-chee.txt	41	cheesehead.txt	35,060	chi-chi.txt	7,191
chinaman.txt	4,345	chink.txt	34,400	cholo.txt	23,706	chonky.txt	234
chug.txt	113,007	chunger.txt	22	chunky.txt	814,818	coconut.txt	1,228,445
coolie.txt	13,939	coon-ass.txt	53	coonass.txt	431	coon.txt	202,610
cracker.txt	304,673	crow.txt	341,117	curry-muncher.txt	11	cushi.txt	176
dago.txt	7,693	darkey.txt	531	darkie.txt	4,607	darky.txt	4,479
dego.txt	2,991	dink.txt	27,633	dogan.txt	8,038	dogun.txt	51
dothead.txt	82	eskimo.txt	69,807	eyetie.txt	49	flip.txt	2,972,449
fritz.txt	58,858	frog.txt	1,004,171	fuzzy-wuzzy.txt	86	gable.txt	28,121
gaijin.txt	6,348	gin.txt	489,042	ginzo.txt	312	gippo.txt	125
gipp.txt	1,942	golliwog.txt	1,945	gook-eye.txt	3	gook.txt	28,169
gooky.txt	782	gora.txt	16,567	goyim.txt	14,407	goy.txt	17,217
goyum.txt	52	greaseball.txt	2,234	greaser.txt	19,622	gringo.txt	29,761
groid.txt	325	gubba.txt	1,084	gub.txt	2,670	guido.txt	43,473
guinea.txt	211,482	guizi.txt	84	gwailo.txt	39	gweilo.txt	255
gyopo.txt	63	gypo.txt	2,584	gyppie.txt	4	gyppo.txt	213
gyppy.txt	35	hadji.txt	4,418	hairyback.txt	32	haji.txt	25,218
hajji.txt	13,953	half-breed.txt	1,781	haole.txt	5,601	hebe.txt	5,826
heeb.txt	709	hillbilly.txt	62,191	honkey.txt	9,108	honkie.txt	639
honky.txt	32,098	hori.txt	16,428	hun.txt	1,071,607	hymie.txt	605
ike.txt	176,068	ikey.txt	1,836	iky.txt	1,426	indon.txt	6,817
injun.txt	4,325	japie.txt	473	jap.txt	123,529	jerry.txt	1,217,221
jigaboo.txt	2,943	jiggabo.txt	42	jigga.txt	43,926	jigger.txt	6,635
jigg.txt	7,101	jig.txt	107,016	jijjiboo.txt	1	jockie.txt	132
jock.txt	187,115	jocky.txt	1,647	kacap.txt	9	kaffer.txt	444
kaffir.txt	10,866	kafir.txt	31,776	kalar.txt	425	kano.txt	245,217
katsap.txt	22	khokhol.txt	8	kike.txt	29,049	kimchi.txt	53,913
klansman.txt	19,284	kraut.txt	10,491	kuffar.txt	32,168	kushi.txt	8,573
kyke.txt	2,577	lebo.txt	15,898	limey.txt	5,294	lubra.txt	57
lugan.txt	62	macaca.txt	274	madrassi.txt	121	majus.txt	327
malaun.txt	62	mau-mau.txt	107	mick.txt	409,868	mooncricket.txt	100
moskal.txt	281	mulignan.txt	28	mulignon.txt	6	mzungu.txt	2,591
neche.txt	716	neechee.txt	23	neejee.txt	3	nichi.txt	1,273
nichiwa.txt	34	nidge.txt	900	nigar.txt	1,793	niger.txt	887,855
nigette.txt	27	niggah.txt	140,963	niggar.txt	3,862	nigga.txt	27,733,830
nigger.txt	340,785	niggress.txt	203	nigguh.txt	86,841	niggur.txt	2,171
niglet.txt	5,405	nig-nog.txt	91	nigor.txt	95	nigra.txt	2,214
nigre.txt	57	nig.txt	176,856	nip.txt	250,681	nitchie.txt	24
nitchy.txt	42	nusayri.txt	1,149	oreo.txt	985,132	paddy.txt	307,544
paki.txt	141,174	palagi.txt	9,662	pancake.txt	304,667	papoose.txt	20,904
peckerwood.txt	1,268	pepper.txt	924,920	pepsi.txt	413,081	pickaninny.txt	279
piefke.txt	7	piker.txt	1,734	pikey.txt	11,063	piky.txt	101
pinoy.txt	193,625	pocha.txt	1,924	pocho.txt	2,238	pohm.txt	20
polack.txt	2,837	pollock.txt	40,745	pommie.txt	1,250	pommy.txt	1,974
pom.txt	126,731	quashie.txt	399	raghead.txt	2,583	rastus.txt	145
razakars.txt	1,019	redlegs.txt	2,863	redneck.txt	184,775	redskin.txt	7,225
roundeye.txt	72	sambo.txt	13,532	sassenach.txt	3,150	sawney.txt	583
scandihoovian.txt	12	schwartze.txt	81	seppo.txt	374	septic.txt	58,390
sheeny.txt	16,589	shelta.txt	179	shiksa.txt	1,025	shine.txt	2,546,492
shylock.txt	2,877	skinny.txt	2,640,539	slant.txt	40,154	slopehead.txt	12
sloper.txt	860	slope.txt	136,616	slopey.txt	384	slopi.txt	17
sloppy.txt	378,197	slopy.txt	365	sooty.txt	7,933	spade.txt	226,593
spearchucker.txt	142	spick.txt	4,314	spic.txt	15,304	spig.txt	502
spik.txt	1,505	spook.txt	43,540	squarehead.txt	551	squaw.txt	14,295
squinty.txt	14,692	tacohead.txt	23	taff.txt	14,347	taffy.txt	103,844
taig.txt	8,616	tar-baby.txt	61	teague.txt	148,134	teapot.txt	98,550
teg.txt	15,007	teig.txt	366	tenker.txt	28	teuchter.txt	309
thicklips.txt	54	tincker.txt	3	tinker.txt	117,600	turk.txt	184,378
twinkie.txt	57,834	tynker.txt	1,844	ukrop.txt	2,604	wasp.txt	213,799
wetback.txt	16,017	whigger.txt	257	whitey.txt	84,534	wigga.txt	4,334
wigger.txt	5,569	wog.txt	8,209	wop.txt	184,768	yankee.txt	443,132
yank.txt	110,373	yarpie.txt	12	yellow.txt	5,316,829	yid.txt	5,784
yuon.txt	2,910	zipperhead.txt	327	zip.txt	919,457		

Table B.2: Record counts of the 169 text files with utc_offset aka DATAFILES.

Input Filename	Count	Input Filename	Count	Input Filename	Count
abbie.txt	36957	goy.txt	7031	nigra.txt	627
abbo.txt	195	goyim.txt	5485	nip.txt	111389
abc.txt	773852	greaseball.txt	675	nusayri.txt	20
abcd.txt	19019	greaser.txt	4926	oreo.txt	491599
abe.txt	114026	gringo.txt	14444	paddy.txt	138204
abeed.txt	417	gub.txt	792	paki.txt	48204
abid.txt	5995	guido.txt	20516	palagi.txt	3322
abie.txt	833	guinea.txt	102104	pancake.txt	158332
abo.txt	19522	gypo.txt	741	papoose.txt	12598
ann.txt	411567	hadji.txt	1608	pepper.txt	474916
ape.txt	177538	haji.txt	11224	pepsi.txt	195963
apple.txt	5403124	hajji.txt	873	piker.txt	475
armo.txt	889	half-breed.txt	466	pikey.txt	4245
banana.txt	759304	haole.txt	2027	pinoy.txt	51318
beaner.txt	10951	hebe.txt	2489	pocha.txt	364
bogtrotter.txt	46	hillbilly.txt	31115	pocho.txt	693
bong.txt	170746	honkey.txt	3106	polack.txt	310
boong.txt	8851	honky.txt	15974	pollock.txt	16469
boonie.txt	1879	hori.txt	7100	pom.txt	58253
bosch.txt	61636	hun.txt	436154	pommy.txt	518
brownie.txt	294893	ike.txt	73861	raghead.txt	606
buck.txt	318217	ikey.txt	449	razakars.txt	8
buffie.txt	273	indon.txt	2737	redlegs.txt	998
bule.txt	4269	injun.txt	954	redneck.txt	81009
bung.txt	7203	jap.txt	54980	redskin.txt	3376
bunga.txt	8151	jerry.txt	620535	sambo.txt	5981
canuck.txt	7961	jig.txt	58096	sassenach.txt	1072
charlie.txt	3285265	jigaboo.txt	1193	septic.txt	25316
cheesehead.txt	2984	jigg.txt	3386	sheeny.txt	7214
chi-chi.txt	2673	jigga.txt	24384	shine.txt	1144697
chinaman.txt	1702	jigger.txt	3504	shylock.txt	890
chink.txt	16318	jock.txt	64713	skinny.txt	1105832
cholo.txt	10902	jocky.txt	299	slant.txt	19081
chug.txt	60716	kaffir.txt	3156	slope.txt	71544
chunky.txt	170230	kafir.txt	11562	sloppy.txt	185581
coconut.txt	646299	kano.txt	93678	sooty.txt	3490
coolie.txt	6491	kike.txt	11462	spade.txt	92034
coon.txt	117264	kimchi.txt	28220	spic.txt	5276
cracker.txt	166150	klansman.txt	10967	spick.txt	1948
crow.txt	174041	kraut.txt	3631	spik.txt	184
dago.txt	4326	kuffar.txt	5784	spook.txt	23082
darkie.txt	1859	kushi.txt	1897	squaw.txt	4680
darky.txt	2061	kyke.txt	190	squinty.txt	7963
dego.txt	339	lebo.txt	6685	wetback.txt	4237
dink.txt	13921	limey.txt	2162	whitey.txt	44041
dogan.txt	3830	mick.txt	221396	wigga.txt	1654
eskimo.txt	33816	mzungu.txt	799	wigger.txt	2274
flip.txt	1317469	nig.txt	56353	wog.txt	3620
fritz.txt	28393	nigar.txt	152	wop.txt	107769
frog.txt	539818	niger.txt	292718	yank.txt	55660
gable.txt	14410	nigga.txt	15432550	yankee.txt	222196
gaijin.txt	3266	niggah.txt	64632	yellow.txt	2281063
gin.txt	262606	niggar.txt	922	yid.txt	2333
gipp.txt	502	nigger.txt	175018	yuon.txt	1001
golliwog.txt	595	nigguh.txt	40654	zip.txt	403861
gook.txt	12233	niggur.txt	563	TOTAL	41260026
gora.txt	6468	niglet.txt	2283		

Table B.3: Metrics for SVM N5+Hour for various values of fraction of negatives.

Frac	**Acc**	**AUPRC**	**AUROC**	**F-Score**	
				Beta=1	**Beta=0.5**
0.01	0.924	0.851	0.882	0.828	0.843
0.02	0.907	0.859	0.874	0.829	0.850
0.03	0.896	0.893	0.879	0.855	0.885
0.04	0.915	0.928	0.912	0.903	0.913
0.05	0.900	0.930	0.901	0.902	0.908
0.06	0.909	0.923	0.909	0.901	0.899
0.07	0.912	0.919	0.910	0.897	0.897
0.08	0.912	0.914	0.905	0.888	0.897
0.09	0.910	0.901	0.899	0.874	0.885
0.1	0.898	0.882	0.882	0.851	0.866
0.11	0.903	0.878	0.880	0.846	0.867
0.12	0.915	0.873	0.889	0.848	0.861
0.13	0.917	0.877	0.891	0.852	0.866
0.14	0.916	0.866	0.888	0.842	0.851
0.15	0.913	0.851	0.875	0.825	0.842
0.16	0.917	0.850	0.883	0.827	0.835
0.17	0.908	0.818	0.857	0.791	0.808
0.18	0.921	0.847	0.875	0.821	0.841
0.19	0.919	0.831	0.873	0.808	0.819
0.2	0.909	0.882	0.891	0.856	0.866
0.21	0.920	0.807	0.861	0.785	0.797
0.22	0.920	0.797	0.854	0.774	0.789
0.23	0.928	0.818	0.853	0.789	0.827
0.24	0.925	0.807	0.852	0.781	0.809
0.25	0.933	0.816	0.862	0.793	0.817
0.26	0.924	0.776	0.847	0.755	0.767
0.27	0.926	0.784	0.848	0.762	0.780
0.28	0.935	0.797	0.852	0.774	0.802
0.29	0.928	0.784	0.842	0.759	0.788
0.3	0.938	0.794	0.860	0.776	0.791
0.4	0.942	0.745	0.835	0.727	0.746
0.5	0.953	0.746	0.819	0.722	0.769
0.6	0.956	0.717	0.810	0.697	0.738
0.7	0.959	0.674	0.762	0.635	0.716
0.8	0.966	0.695	0.753	0.639	0.750
0.9	0.966	0.675	0.725	0.596	0.733
1.0	0.965	0.606	0.675	0.496	0.652

References

ADL (2014) *Best Practices for Responding to Cyberhate.* [Online; accessed 22-January-2018]. Available at `https://www.adl.org/best-practices-for-responding-to-cyberhate`.

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., and Passonneau, R. (2011) 'Sentiment analysis of twitter data', in: *Proceedings of the workshop on languages in social media.* Association for Computational Linguistics, pp. 30–38.

Agarwal, S. (2017) *Why big data projects fail and how to make 2017 different.* [Online; accessed 27-November-2017]. Available at `https://www.networkworld.com/article/3170137/cloud-computing/why-big-data-projects-fail-and-how-to-make-2017-different.html`.

Alapati, S. R. (2016) *Expert Hadoop Administration: Managing, Tuning, and Securing Spark, YARN, and HDFS.* Boston: Addison-Wesley Professional.

Alexa (2018) *twitter.com Traffic Statistics.* [Online; accessed 22-February-2018]. Available at `https://www.alexa.com/siteinfo/twitter.com`.

Alexander, B. and Levine, A. (2008) 'Web 2.0 storytelling: Emergence of a new genre', *EDUCAUSE review*, 43(6), pp. 40–56.

Alonso Alemany, L. and Carrascosa, R. (2011) ‘A system for adaptive information extraction from highly informal text’, in: *International Conference on Application of Natural Language to Information Systems.* Springer, pp. 145–152.

Anderson, L. and Lepore, E. (2013) ‘Slurring words’, *Noûs*, 47(1), pp. 25–48.

Andrews, K. (2014) ‘From the ‘Bad Nigger’ to the ‘Good Nigga’: an unintended legacy of the Black Power movement’, *Race & Class*, 55(3), pp. 22–37.

Apache.org (2018) *Extracting, transforming and selecting features.* [Online; accessed 1-March-2018]. Available at `https://spark.apache.org/docs/2.2.0/ml-features.html`.

Awan, I. and Zempi, I. (2017) ‘I Will Blow Your Face OFF - VIRTUAL and Physical World Anti-muslim Hate Crime’, *The British Journal of Criminology*, 57(2), pp. 362–380.

Badjatiya, P., Gupta, S., Gupta, M., and Varma, V. (2017) ‘Deep learning for hate speech detection in tweets’, in: *Proceedings of the 26th International Conference on World Wide Web Companion.* International World Wide Web Conferences Steering Committee, pp. 759–760.

Bamman, D., Dyer, C., and Smith, N. A. (2014) ‘Distributed Representations of Geographically Situated Language’, in: *Proceedings of the Annual Meeting of the Association for Computational Linguistics (Short Papers).* Vol. 828, p. 834.

Banton, M. (1998) *Racial theories.* Cambridge: Cambridge University Press.

Barendt, E. (2011) ‘Religious Hatred Laws: Protecting Groups or Belief?’, *Res Publica*, 17(1), pp. 41–53.

Barthes, R. (1978) *Image-music-text.* London: Fontana.

Bartlett, J., Reffin, J., Rumball, N., and Williamson, S. (2014) ‘Anti-social media’, *Demos*, pp. 1–51.

Bauman, Z. (1996) ‘From pilgrim to tourist–or a short history of identity’, *Questions of Cultural Identity*, 1, pp. 18–36.

BBC (2016) *Standing up for hate.* Available at `http://www.bbc.co.uk/news/blogs-trending-36427153`.

— (2018) *Met Police chief: Social media leads children to violence.* Available at `http://www.bbc.co.uk/news/uk-43603080`.

Beletsky, L., Wagner, K. D., Arredondo, J., Palinkas, L., Magis Rodríguez, C., Kalic, N., Ludwig-Barron, N., and Strathdee, S. A. (2016) ‘Implementing Mexico’s "Narcomenudeo" drug law reform: a mixed methods assessment of early experiences among people who inject drugs’, *Journal of Mixed Methods Research*, 10(4), pp. 384–401.

Ben-David, A. and Matamoros-Fernandez, A. (2016) ‘Hate speech and covert discrimination on social media: monitoring the Facebook pages of extreme-right political parties in Spain’, *International Journal of Communication*, 10, pp. 1167–1193.

Benesch, S. (2013) *Proposed Guidelines for Dangerous Speech.* [Online; accessed 27-November-2017]. Available at `https://dangerousspeech.org/guidelines/`.

Bermingham, A. and Smeaton, A. (2011) ‘On using Twitter to monitor political sentiment and predict election results’, in: *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pp. 2–10.

Beyer, J. L. (2012) *What does anonymity mean? Reddit, activism, and ’creepshots’.* [Online; accessed 27-November-2017]. Available at `http://www.jlbeyer.com/what-does-online-anonymity-mean-reddit-activism-and-creepshots/`.

Bhaskar, R. (2009) *Scientific realism and human emancipation.* Routledge.

Biernacki, P. and Waldorf, D. (1981) ‘Snowball sampling: Problems and techniques of chain referral sampling’, *Sociological Methods & Research*, 10(2), pp. 141–163.

Bird, S., Klein, E., and Loper, E. (2009) *Natural language processing with Python: analyzing text with the natural language toolkit.* Sebastopol, CA: O'Reilly Media, Inc.

Birkbeck, C. and LaFree, G. (1993) 'The situational analysis of crime and deviance', *Annual Review of Sociology*, 19(1), pp. 113–137.

Bleich, E. (2014) 'Freedom of expression versus racist hate speech: Explaining differences between high court regulations in the USA and Europe', *Journal of Ethnic and Migration Studies*, 40(2), pp. 283–300.

Bossler, A. M. and Holt, T. J. (2009) 'On-line activities, guardianship, and malware infection: An examination of routine activities theory.', *International Journal of Cyber Criminology*, 3(1),

Bourdieu, P. (1984) *Distinction: A social critique of the judgement of taste.* Harvard University Press.

Box, S. (2002) *Power, crime and mystification.* Routledge.

Boyd, D. and Crawford, K. (2012) 'Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon', *Information, Communication & Society*, 15(5), pp. 662–679.

Bradley, H. (2015) *Fractured identities: Changing patterns of inequality.* Chichester: Wiley.

Brantingham, P. L. and Brantingham, P. J. (1993) 'Nodes, paths and edges: Considerations on the complexity of crime and the physical environment', *Journal of Environmental Psychology*, 13(1), pp. 3–28.

— (2015) 'Understanding crime with computational topology', in: Andresen, M. and Farrell, G. (eds.) *The Criminal Act.* London: Palgrave, pp. 131–145.

Brantingham, P. J., Brantingham, P. L., et al. (1981) *Environmental criminology.* Beverly Hills, CA: Sage.

Brantingham, P. J., Brantingham, P. L., and Andresen, M. A. (2016) 'The geometry of crime and crime pattern theory', in: Wortley, R. and Townsley, M. (eds.) *Environmental Criminology and Crime Analysis.* Taylor & Francis, pp. 98–115.

Bromley, R. D. and Nelson, A. L. (2002) 'Alcohol-related crime and disorder across urban space and time: evidence from a British city', *Geoforum*, 33(2), pp. 239–254.

Brown, C. (2009) 'WWW.HATE.COM: White supremacist discourse on the internet and the construction of whiteness ideology', *The Howard Journal of Communications*, 20(2), pp. 189–208.

Brown, N., Newbury-Birch, D., McGovern, R., Phinn, E., and Kaner, E. (2010) 'Alcohol screening and brief intervention in a policing context: a mixed methods feasibility study', *Drug and Alcohol Review*, 29(6), pp. 647–654.

Bryant, R. (2014) 'Digital Crime', in: Bryant, R. and Bryant, S. (eds.) *Policing Digital Crime.* Farnham: Ashgate, pp. 1–42.

Bryant, R. and Day, E. (2014) 'Law and Digital Crime', in: Bryant, R. and Bryant, S. (eds.) *Policing Digital Crime.* Burlington, Vermont: Ashgate., pp. 83–110.

Bryant, R. (2008) 'The challenge of digital crime', in: Bryant, R. and Bryant, S. (eds.) *Investigating Digital Crime.* Citeseer, pp. 1–26.

Bryant, R. and Bryant, S. (2014) *Blackstone's Handbook for Policing Students 2015.* Oxford: Oxford University Press.

Burgess, R. L. and Akers, R. L. (1966) 'A differential association-reinforcement theory of criminal behavior', *Social Problems*, 14(2), pp. 128–147.

Burnap, P. and Williams, M. L. (2015) 'Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making', *Policy & Internet*, 7(2), pp. 223–242.

Burnap, P. and Williams, M. L. (2016) ‘Us and them: identifying cyber hate on Twitter across multiple protected characteristics’, *EPJ Data Science*, 5(1), p. 11.

Burnap, P., Williams, M. L., Sloan, L., Rana, O., Housley, W., Edwards, A., Knight, V., Procter, R., and Voss, A. (2014) ‘Tweeting the terror: modelling the social media reaction to the Woolwich terrorist attack’, *Social Network Analysis and Mining*, 4(1), p. 206.

Burnap, P., Rana, O. F., Avis, N., Williams, M., Housley, W., Edwards, A., Morgan, J., and Sloan, L. (2015) ‘Detecting tension in online communities with computational Twitter analysis’, *Technological Forecasting and Social Change*, 95, pp. 96–108.

Burnap, P. and Williams, M. L. (2014) *Hate speech, machine classification and statistical modelling of information flows on Twitter: Interpretation and communication for policy decision making.* Presented at: Internet, Policy & Politics, Oxford, UK, 26 September 2014.

Caruana, R. and Niculescu-Mizil, A. (2006) ‘An empirical comparison of supervised learning algorithms’, in: *Proceedings of the 23rd international conference on Machine learning.* ACM, pp. 161–168.

Cashmore, E. (1996) *Dictionary of race and ethnic relations.* Psychology Press.

Castells, M. (2011) *The rise of the network society.* Vol. 12. c: Wiley.

Cetina, K. K. (2009) ‘The synthetic situation: Interactionism for a global world’, *Symbolic Interaction*, 32(1), pp. 61–87.

Chan, J. and Bennett Moses, L. (2016) ‘Is big data challenging criminology?’, *Theoretical Criminology*, 20(1), pp. 21–39.

Charmaz, K. and Belgrave, L. L. (2007) ‘Grounded theory’, *The Blackwell Encyclopedia of Sociology*, Chattoo, S. and Atkin, K. (2012) ‘Race, ethnicity and social policy: theoretical concepts and the limitations of current approaches to welfare’, in: *Understanding Race and Ethnicity: Theory, History, Policy and Practice.* Bristol: Policy Press, pp. 17–38.

Chaudhry, I. (2015) ‘# Hashtagging hate: Using Twitter to track racism online’, *First Monday*, 20(2),

Chen, F. and Neill, D. B. (2014) ‘Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs’, in: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, pp. 1166–1175.

Chen, M. (2015) *TF-IDF, HashingTF and CountVectorizer.* Available at `https://mingchen0919.github.io/learning-apache-spark/tf-idf.html`.

Chen, Y., Zhou, Y., Zhu, S., and Xu, H. (2012) ‘Detecting offensive language in social media to protect adolescent online safety’, in: *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom).* IEEE, pp. 71–80.

Cheung, C. (2013) ‘Understanding factors associated with online piracy behaviour of adolescents’, *International Journal of Adolescence and Youth*, 18(2), pp. 122–132.

Choi, K.-S. (2008) ‘Computer crime victimization and integrated theory: An empirical assessment’, *International Journal of Cyber Criminology*, 2(1), p. 308.

Choi, K.-S. and Lee, J. R. (2017) ‘Theoretical analysis of cyber-interpersonal violence victimization and offending using cyber-routine activities theory’, *Computers in Human Behavior*, 73, pp. 394–402.

Clarke, R. V. (1983) ‘Situational crime prevention: Its theoretical basis and practical scope’, *Crime and Justice*, 4, pp. 225–256.

Clarke, R. V. (1980) ‘"Situational" Crime Prevention: Theory and Practice’, *The British Journal of Criminology*, 20(2), pp. 136–147.

Clarke, R. V. and Felson, M. (1993) *Routine activity and rational choice.* Piscataway, NJ: Transaction Publishers.

Cloudera (2014) *Apache Spark Resource Management and YARN App Models.* [Online; accessed 22-February-2018]. Available at `https://blog.cloudera.com/blog/2014/05/apache-spark-resource-management-and-yarn-app-models/`.

Cohen, L. E. and Felson, M. (1979) ‘Social Change and Crime Rate Trends: A Routine Activity Approach’, *American Sociological Review*, 44(4), pp. 588–608. issn: 00031224. Available at `http://www.jstor.org/stable/2094589`.

Cohen, L. E., Kluegel, J. R., and Land, K. C. (1981) ‘Social inequality and predatory criminal victimization: An exposition and test of a formal theory’, *American Sociological Review*, 46, pp. 505–524.

Cohn, D. Y. and Vaccaro, V. L. (2006) ‘A study of neutralisation theory’s application to global consumer ethics: P2P file-trading of musical intellectual property on the internet’, *International Journal of Internet Marketing and Advertising*, 3(1), pp. 68–88.

Cornish, D. B. and Clarke, R. V. (1986) *The Reasoning Criminal: Rational Choice Perspectives on Offending.* Piscataway, NJ: Transaction Publishers.

CPS (2017) *Racist and Religious Hate Crime - Prosecution Guidance.* [Online; accessed 27-November-2017]. Available at `http://www.cps.gov.uk/legal/p_to_r/racist_and_religious_crime/`.

— (2018) *Social Media: Guidelines on prosecuting cases involving communications sent via social media.* [Online; accessed 8-March-2018]. Available at `https://www.cps.`

`gov.uk/legal-guidance/social-media-guidelines-prosecuting-cases-involving-communications-sent-social-media`.

Crandall, C. S. and Eshleman, A. (2003) 'A justification-suppression model of the expression and experience of prejudice.', *Psychological Bulletin*, 129(3), p. 414.

Croom, A. M. (2015) 'Slurs, stereotypes, and in-equality: A critical review of 'How Epithets and Stereotypes are Racially Unequal'', *Language Sciences*, 52, pp. 139–154.

Dadvar, M., de Jong, F. M., Ordelman, R. J., and Trieschnigg, R. B. (2012) 'Improved cyberbullying detection using gender information', in: *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012).* Ghent University.

Dadvar, M., Trieschnigg, D., Ordelman, R., and Jong, F. de (2013) 'Improving cyberbullying detection with user context', in: *European Conference on Information Retrieval.* Springer, pp. 693–696.

Damji, J. (2016) *A Tale of Three Apache Spark APIs: RDDs, DataFrames, and Datasets.* [Online; accessed 27-November-2017]. Available at `https://databricks.com/blog/2016/07/14/a-tale-of-three-apache-spark-apis-rdds-dataframes-and-datasets.html`.

Danet, B. (1998) *Text as Mask. Gender, Play, and Performance on the Internet. I Jones, SG (ed.) CyberSociety 2.0. Revisiting Computer-Mediated Communication and Community.* Thousand Oaks: Sage Publications.

Databricks (2016) *Stanford CoreNLP wrapper for Apache Spark.* [Online; accessed 13-October-2016]. Available at `https://github.com/databricks/spark-corenlp`.

Davidson, T., Warmsley, D., Macy, M., and Weber, I. (2017) 'Automated hate speech detection and the problem of offensive language', in: *Proceedings of the Eleventh International Conference on Web and Social Media.* ICWSM.

Davis, J. and Goadrich, M. (2006) 'The relationship between Precision-Recall and ROC curves', in: *Proceedings of the 23rd international conference on Machine learning.* ACM, pp. 233–240.

Davis, N. J. (1972) 'Labeling theory in deviance research: A critique and reconsideration', *The Sociological Quarterly*, 13(4), pp. 447–474.

De Mauro, A., Greco, M., Grimaldi, M., Giannakopoulos, G., Sakas, D. P., and Kyriaki-Manessi, D. (2015) 'What is big data? A consensual definition and a review of key research topics', in: *AIP conference proceedings.* Vol. 1644. 1. AIP, pp. 97–104.

De Swert, K. (2012) 'Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha', *Center for Politics and Communication*, pp. 1–15.

Dean, J. and Ghemawat, S. (2008) 'MapReduce: simplified data processing on large clusters', *Communications of the ACM*, 51(1), pp. 107–113.

Delgado, R. (1982) 'Words that wound: A tort action for racial insults, epithets, and name-calling', *Harv. CR-CLL Rev.* 17, p. 133.

DeVan, A. (2016) *The 7 V's of Big Data.* [Online; accessed 30-November-2017]. Available at `https://www.impactradius.com/blog/7-vs-big-data/`.

Dinakar, K., Jones, B., Havasi, C., Lieberman, H., and Picard, R. (2012) 'Common sense reasoning for detection, prevention, and mitigation of cyberbullying', *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3), p. 18.

Djuric, N., Zhou, J., Morris, R., Grbovic, M., Radosavljevic, V., and Bhamidipati, N. (2015) 'Hate speech detection with comment embeddings', in: *Proceedings of the 24th international conference on world wide web.* ACM, pp. 29–30.

Donner, C. M., Marcum, C. D., Jennings, W. G., Higgins, G. E., and Banfield, J. (2014) 'Low self-control and cybercrime: Exploring the utility of the general theory of crime beyond digital piracy', *Computers in Human Behavior*, 34, pp. 165–172.

Doward, J. (2017) *Government's new online hate crime hub given just £200,000.* [Online; accessed 27-November-2017]. Available at `https://www.theguardian.com/society/2017/oct/14/government-criticised-for-low-funding-level-to-tackle-online-hate`.

Durkheim, E. (1933) 'The division of labor in Society',

dw.com (2018) *EU: Twitter, Facebook still in violation of the bloc's consumer law.* [Online; accessed 8-March-2018]. Available at `http://www.dw.com/en/eu-twitter-facebook-still-in-violation-of-the-blocs-consumer-law/a-42599157`.

Dwyer, A., Ball, M., Bond, C., Lee, M., and Crofts, T. (2017) *Reporting Victimisation to LGBTI (Lesbian, Gay, Bisexual, Transgender, Intersex) Police Liaison Services: A mixed methods study across two Australian states.* [Online; accessed 8-March-2018]. Available at `http://ecite.utas.edu.au/107082`.

Eck, J. (1994) 'Drug markets and drug places', *Unpublished PhD dissertation, University of Maryland, College Park, MD*,

Eck, J. E. (1995) *Examining routine activity theory: A review of two books.*

Eck, J. E. and Madensen, T. D. (2015) 'Meaningfully and artfully reinterpreting crime for useful science: an essay on the value of building with simple theory', in: Andresen, M. and Farrell, G. (eds.) *The Criminal Act.* London: Palgrave, pp. 5–18.

Edwards, L. (2012) 'Section 127 of the Communications Act 2003: Threat or Menace?', *SCL Journal*, 23(4), EJN Secretariat (2018) *Status of implementation of 2008/913/JHA*. [Online; accessed 06-March-2018]. Available at https://www.ejn-crimjust.europa.eu/ejn/EJN_Library_StatusOfImpByCat.aspx?CategoryId=60.

Elliott, I., Thomas, S. D., and Ogloff, J. R. (2011) 'Procedural justice in contacts with the police: Testing a relational model of authority in a mixed methods study.', *Psychology, Public Policy, and Law*, 17(4), p. 592.

Ellison, N. B. and Boyd, D. (2007) 'Social network sites: Definition, history, and scholarship', *Journal of computer-mediated Communication*, 13(1), pp. 210–230.

Engward, H. (2013) 'Understanding grounded theory', *Nursing Standard (through 2013)*, 28(7), p. 37.

European Commission (2018) *A Europe that protects: Commission reinforces EU response to illegal content online*. [Online; accessed 8-March-2018]. Available at http://europa.eu/rapid/press-release_IP-18-1169_en.htm.

Faith Matters (2017) *Tell MAMA 2016 Annual Report*. [Online; accessed 16-February-2018]. Available at https://tellmamauk.org/wp-content/uploads/2017/11/A-Constructed-Threat-Identity-Intolerance-and-the-Impact-of-Anti-Muslim-Hatred-Web.pdf.

— (2018) *About us*. [Online; accessed 16-February-2018]. Available at https://tellmamauk.org/about-us/.

Feilzer, M. Y. (2010) 'Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm', *Journal of mixed methods research*, 4(1), pp. 6–16.

Feldman, B. (2017) *Twitter Doesn't Want to Be Twitter.* [Online; accessed 8-March-2018]. Available at `http://nymag.com/selectall/2017/03/twitter-doesnt-want-to-be-twitter.html`.

Felson, M. (1986) 'Linking Criminal Choices, Routine Activities, Informal Control, and Outcomes', in: Cornish, D. B. and Clarke, R. V. (eds.) *The reasoning criminal: Rational choice perspectives on offending.* New York: Springer-Verlag, p. 119.

Felson, M. and Clarke, R. (1998) 'Opportunity Makes the Thief. Police Research Series Paper 98, Policing and Reducing Crime Unit', *Research, Development and Statistics Directorate. London: Home Office*,

Fenton, S. (1999) *Ethnicity: Racism, class and culture.* Lanham, MD: Rowman & Littlefield.

Finkelhor, D. and Asdigian, N. L. (1996) 'Risk factors for youth victimization: Beyond a lifestyles/routine activities theory approach', *Violence and victims*, 11(1), p. 3.

Fleiss, J. L. (1971) 'Measuring nominal scale agreement among many raters.', *Psychological bulletin*, 76(5), p. 378.

Fortuna, P. C. T. (2017) *Automatic detection of hate speech in text: an overview of the topic and dataset annotation with hierarchical classes.* [Online; accessed 27-November-2017]. Available at `https://repositorio-aberto.up.pt/bitstream/10216/106028/2/202853.pdf`.

Foundation, A. S. (2015) *HDFS Commands Guide.* [Online; accessed 22-January-2018]. Available at `https://hadoop.apache.org/docs/r2.7.1/hadoop-project-dist/hadoop-hdfs/HDFSCommands.html`.

Friedman, J. H. (2001) 'Greedy function approximation: a gradient boosting machine', *Annals of Statistics*, pp. 1189–1232.

Fuchs, C. (2017) *Social media: A critical introduction.* London: Sage.

Garland, J. and Chakraborti, N. (2012) ‘Divided by a common concept? Assessing the implications of different conceptualizations of hate crime in the European Union’, *European Journal of Criminology*, 9(1), pp. 38–51.

Garner, S. (2017) *Racisms: an introduction.* London: Sage.

Geer, D (2007) ‘The physics of digital law: searching for counterintuitive analogies’, *Cybercrime: Digital cops in a networked environment*, pp. 13–36.

Gelber, K. and McNamara, L. (2015) ‘The Effects of Civil Hate Speech Laws: Lessons from Australia’, *Law & Society Review*, 49(3), p. 631.

Ghemawat, S., Gobioff, H., and Leung, S.-T. (2003) ‘The Google file system’, in: *ACM SIGOPS operating systems review.* Vol. 37. 5. ACM, pp. 29–43.

Gillespie, A. A. (2010) ‘Racially offensive web postings’, *J. Crim. L.* 74, p. 205.

Gitari, N. D., Zuping, Z., Damien, H., and Long, J. (2015) ‘A lexicon-based approach for hate speech detection’, *International Journal of Multimedia and Ubiquitous Engineering*, 10(4), pp. 215–230.

Glaser, B. and Strauss, A. L. (1967) *The discovery of grounded theory: Strategies for qualitative research.* New York: Aldine.

Goffman, E. (1981) *Forms of talk.* Pennsylvania, PA: University of Pennsylvania Press.

Golash-Boza, T. (2016) ‘A critical and comprehensive sociological theory of race and racism’, *Sociology of Race and Ethnicity*, 2(2), pp. 129–141.

Google (2017) *Google Maps Geocoding API Usage Limits.* [Online; accessed 7-March-2018]. Available at `https://developers.google.com/maps/documentation/geocoding/usage-limits`.

Gordon, S. and Ford, R. (2006) 'On the definition and classification of cybercrime', *Journal in Computer Virology*, 2(1), pp. 13–20.

Goswami, S., Saha, U., and Bose, S. (2016) 'An Information Management System for Crime against Women in India from News: A Big Data Solution', *Data Mining and Knowledge Engineering*, 8(2), pp. 48–57.

Gottfredson, M. R. and Hirschi, T. (1990) *A general theory of crime.* Stanford, CA: Stanford University Press.

Goutte, C. and Gaussier, E. (2005) 'A probabilistic interpretation of precision, recall and F-score, with implication for evaluation.', in: *ECIR.* Vol. 5. Springer, pp. 345–359.

Gouws, S., Metzler, D., Cai, C., and Hovy, E. (2011a) 'Contextual bearing on linguistic variation in social media', in: *Proceedings of the Workshop on Languages in Social Media.* Association for Computational Linguistics, pp. 20–29.

Gouws, S., Hovy, D., and Metzler, D. (2011b) 'Unsupervised mining of lexical variants from noisy text', in: *Proceedings of the First workshop on Unsupervised Learning in NLP.* Association for Computational Linguistics, pp. 82–90.

Greevy, E. and Smeaton, A. F. (2004) 'Classifying racist texts using a support vector machine', in: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, pp. 468–469.

Guyon, I. and Elisseeff, A. (2003) 'An introduction to variable and feature selection', *Journal of machine learning research*, 3(Mar), pp. 1157–1182.

Habermas, J. (1989) 'The Structural Transformation of the Public Sphere', *Polity*, 7(8),

Habermas, J., Lennox, S., and Lennox, F. (1974) 'The public sphere: An encyclopedia article (1964)', *New German Critique*, (3), pp. 49–55.

Hall, S. and Du Gay, P. (2006) *Questions of cultural identity.* Crane Resource Centre.

Hardy, S.-J. and Chakraborti, N. (2017) *Hate Crime: Identifying and Dismantling Barriers to Justice.* Leicester: University of Leicester.

Harris, D. (2013) *The history of Hadoop: From 4 nodes to the future of data.* [Online; accessed 7-March-2018]. Available at `https://gigaom.com/2013/03/04/the-history-of-hadoop-from-4-nodes-to-the-future-of-data/`.

Hasanuzzaman, M., Dias, G., and Way, A. (2017) 'Demographic Word Embeddings for Racism Detection on Twitter', in: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* Vol. 1, pp. 926–936.

Hawdon, J., Oksanen, A., and Räsänen, P. (2017) 'Exposure to online hate in four nations: a cross-national consideration', *Deviant behavior*, 38(3), pp. 254–266.

Hechter, M. and Kanazawa, S. (1997) 'Sociological rational choice theory', *Annual review of sociology*, 23(1), pp. 191–214.

Heimerl, F., Lohmann, S., Lange, S., and Ertl, T. (2014) 'Word cloud explorer: Text analytics based on word clouds', in: *System Sciences (HICSS), 2014 47th Hawaii International Conference on.* IEEE, pp. 1833–1842.

Helms, J. E. (1984) 'Toward a theoretical explanation of the effects of race on counseling a Black and White Model', *The Counseling Psychologist*, 12(4), pp. 153–165.

Hindelang, M. J., Gottfredson, M. R., and Garofalo, J. (1978) *Victims of personal crime: An empirical foundation for a theory of personal victimization.* Cambridge, MA: Ballinger.

Hirschi, T., Gottfredson, M., et al. (1986) 'The distinction between crime and criminality', in: *Critique and explanation: Essays in honor of Gwynne Nettler.* Transaction Books New Brunswick, NJ, pp. 44–69.

Hoffman, B. (2015) *Motivation for learning and performance.* Cambridge, MA: Academic Press.

Hollis, M. E., Felson, M., and Welsh, B. C. (2013) 'The capable guardian in routine activities theory: A theoretical and conceptual reappraisal', *Crime Prevention and Community Safety*, 15(1), pp. 65–79.

Holmes, A. (2012) *Hadoop in practice.* Shelter Island, NY: Manning Publications Co.

Home Office (2016) *Action against hate: The UK Government's plan for tackling hate crime.* Available at `https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/543679/Action_Against_Hate_-_UK_Government_s_Plan_to_Tackle_Hate_Crime_2016.pdf`.

Horsman, G., Ginty, K., and Cranner, P. (2017) 'Identifying offenders on Twitter: A law enforcement practitioner guide', *Digital Investigation*, 57(6), pp. 448–454.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., and Mishra, S. (2015) 'Detection of cyberbullying incidents on the instagram social network', *Association for the Advancement of Artificial Intelligence*,

Hunton, P. (2012) 'Data attack of the cybercriminal: Investigating the digital currency of cybercrime', *Computer Law & Security Review*, 28(2), pp. 201–207.

Jain, A. and Bhatnagar, V. (2016) 'Crime data analysis using pig with hadoop', *Procedia computer science*, 78, pp. 571–578.

Japkowicz, N. and Shah, M. (2011) *Evaluating learning algorithms: a classification perspective.* Cambridge: Cambridge University Press.

Jeffery, C. R. (1993) 'Obstacles to the development of research in crime and delinquency', *Journal of research in crime and delinquency*, 30(4), pp. 491–497.

Jenkins, R. (1997) *Rethinking ethnicity: Arguments and explorations.* London: Sage.

Joachims, T. (1998) 'Text categorization with support vector machines: Learning with many relevant features', *Machine learning: ECML-98*, pp. 137–142.

Jones, T. and Newburn, T. (2002) 'The transformation of policing? Understanding current trends in policing systems', *The British journal of criminology*, 42(1), pp. 129–146.

Kantar (2015) *Frequently asked questions.* [Online; accessed 27-November-2017]. Available at `http://www.crimesurvey.co.uk/FAQs.html`.

Kantrowitz, M. (2016) *Name corpus: List of male, female, and pet names.* [Online; accessed 27-November-2017]. Available at `https://www.cs.cmu.edu/Groups/AI/areas/nlp/corpora/names/`.

Kao, A. and Poteet, S. R. (2007) *Natural language processing and text mining.* Heidelberg: Springer.

Karstedt, S. (2001) 'Comparing cultures, comparing crime: Challenges, prospects and problems for a global criminology', *Crime, Law and Social Change*, 36(3), pp. 285–308.

Kaushik, S. (2016) *Introduction to Feature Selection methods with an example (or how to select the right variables?)* [Online; accessed 22-February-2018]. Available at `https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/`.

Kawa, A. (2014) *Introduction to YARN.* [Online; accessed 22-February-2018]. Available at `https://www.ibm.com/developerworks/library/bd-yarn-intro/`.

Kiedrowski, J., Ruddell, R., and Petrunik, M. (2017) 'Police civilianisation in Canada: a mixed methods investigation', *Policing and Society*, pp. 1–19.

King, C. R. et al. (2014) *Beyond Hate: White Power and Popular Culture*. Farnham: Ashgate Publishing, Ltd.

Kitchin, R. (2014) *The data revolution: Big data, open data, data infrastructures and their consequences*. London: Sage.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., and Makse, H. A. (2010) 'Identification of influential spreaders in complex networks', *Nature physics*, 6(11), p. 888.

Kleinwächter, W. (2017) 'Internet Governance Outlook 2017: Nationalistic Hierarchies vs. Multistakeholder Networks?', *Circle ID*, pp. 3–4.

Kobus, C., Yvon, F., and Damnati, G. (2008) 'Normalizing SMS: are two metaphors better than one?', in: *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pp. 441–448.

Kohavi, R. et al. (1995) 'A study of cross-validation and bootstrap for accuracy estimation and model selection', in: *Ijcai*. Vol. 14. 2. Montreal, Canada, pp. 1137–1145.

Kordos, M. (2005) 'Search-based algorithms for multilayer perceptrons'. PhD thesis.

Krippendorff, K. (2004) 'Reliability in content analysis', *Human communication research*, 30(3), pp. 411–433.

Kuehl, K. S., Elliot, D. L., MacKinnon, D. P., O'Rourke, H. P., DeFrancesco, C., Miočević, M., Valente, M., Sleigh, A., Garg, B., McGinnis, W., et al. (2016) 'The SHIELD (Safety & Health Improvement: Enhancing Law Enforcement Departments) Study: Mixed Methods Longitudinal Findings', *Journal of occupational and environmental medicine/American College of Occupational and Environmental Medicine*, 58(5), p. 492.

Kuhn, M. and Johnson, K. (2013) *Applied predictive modeling*. Vol. 810. New York: Springer.

Kumar, E. (2011) *Natural language processing*. New Delhi: IK International Pvt Ltd.

Kwok, I. and Wang, Y. (2013) 'Locate the hate: detecting tweets against blacks', in: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*. AAAI Press, pp. 1621–1622.

Lam, C. (2010) *Hadoop in action*. Shelter Island, NY: Manning Publications Co.

Laskowski, J. (2017) *Mastering Apache Spark 2 (Spark 2.2+)*. [Online; accessed 27-November-2017]. Available at `https://www.gitbook.com/download/pdf/book/jaceklaskowski/mastering-apache-spark`.

Law, T. F. (2018) *21 Rape Allegations Later, Ian Connor Remains in Fashion's Buzziest Circles*. [Online; accessed 18-April-2018]. Available at `http://www.thefashionlaw.com/home/21-rape-allegations-later-kanye-west-muse-ian-connor-remains-in-fashions-inner-circle`.

Le, Q. and Mikolov, T. (2014) 'Distributed representations of sentences and documents', in: *International Conference on Machine Learning*, pp. 1188–1196.

legislation.gov *Public Order Act 1986*. [Online; accessed 27-November-2017]. Available at `http://www.legislation.gov.uk/ukpga/1986/64/contents`.

— (2018a) *Race Relations Act 1968*. [Online; accessed 27-November-2017]. Available at `http://www.legislation.gov.uk/ukpga/1968/71/pdfs/ukpga_19680071_en.pdf`.

— (2018b) *Racial and Religious Hatred Act 2006*. [Online; accessed 27-November-2017]. Available at `http://www.legislation.gov.uk/ukpga/2006/1/schedule`.

Lemert, E. M. (1951) *Social pathology; A systematic approach to the theory of sociopathic behavior*. New York: McGraw-Hill.

Leukfeldt, E. R. and Yar, M. (2016) 'Applying routine activity theory to cybercrime: A theoretical and empirical analysis', *Deviant Behavior*, 37(3), pp. 263–280.

Leverenz, L. (2017) *LanguageManual ORC*. [Online; accessed 27-November-2017]. Available at `https://cwiki.apache.org/confluence/display/Hive/LanguageManualORC`.

Littler, M. and Feldman, M. (2015) *Tell MAMA Reporting 2014/2015: Annual monitoring, cumulative extremism, and policy implications*. Middlesbrough: Teesside University Centre for Fascist, Anti-Fascist and Post-Fascist Studies.

Lobba, P. (2014) 'Punishing Denialism beyond Holocaust Denial: EU Framework Decision 2008/913/JHA and other Expansive Trends', *New Journal of European Criminal Law*, 5(1), pp. 58–77.

Lomas, N. (2017) *Twitter's abuse problem is absolutely a failure of leadership*. [Online; accessed 27-November-2017]. Available at `https://beta.techcrunch.com/2017/10/12/twitters-abuse-problem-is-absolutely-a-failure-of-leadership/?utm_content=bufferf9bd3\&utm_medium=social\&utm_source=twitter.com\&utm_campaign=buffer`.

Long, P. M. and Servedio, R. A. (2010) 'Random classification noise defeats all convex potential boosters', *Machine learning*, 78(3), pp. 287–304.

Lotan, G., Graeff, E., Ananny, M., Gaffney, D., Pearce, I., et al. (2011) 'The Arab Spring. The revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions', *International journal of communication*, 5, p. 31.

MacGuill, D. (2017). [Online; accessed 27-December-2017]. Available at `https://www.snopes.com/fact-check/twitter-germany-nazis/`.

Madensen, T. D. (2007) 'Bar management and crime: Toward a dynamic theory of place management and crime hotspots'. PhD thesis. University of Cincinnati.

Malik, K. (1996) *The meaning of race: Race, history and culture in Western society.* New York: NYU Press.

Malik, M. (1999) ''Racist Crime': Racially Aggravated Offences in the Crime and Disorder Act 1998 Part II', *The Modern Law Review*, 62(3), pp. 409–424.

Malmasi, S. and Zampieri, M. (2017a) 'Challenges in discriminating profanity from hate speech', *Journal of Experimental & Theoretical Artificial Intelligence*, pp. 1–16.

— (2017b) 'Detecting Hate Speech in Social Media', *arXiv preprint arXiv:1712.06427*,

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014) 'The stanford corenlp natural language processing toolkit.', in: *ACL (System Demonstrations)*, pp. 55–60.

Mansour, R. F. (2016) 'Understanding how big data leads to social networking vulnerability', *Computers in Human Behavior*, 57, pp. 348–351.

Marcum, C. D., Higgins, G. E., and Ricketts, M. L. (2010) 'Potential factors of online victimization of youth: An examination of adolescent online behaviors utilizing routine activity theory', *Deviant Behavior*, 31(5), pp. 381–410.

Marsland, S. (2015) *Machine learning: an algorithmic perspective.* Boca Raton, Fl: CRC press.

Martin, C. (2018) 'Striking the Right Balance: Hate Speech Laws in Japan, the United States, and Canada', *Hastings Constitutional Law Quarterly*, 45(3), pp. 455–532.

Martinez Jr, B. A. and Selepak, A. (2014) ‘The sound of hate: exploring the use of hatecore song lyrics as a recruiting strategy by the White Power Movement’, *Intercom: Revista Brasileira de Ciências da Comunicação*, 37(2), pp. 153–175.

Marwick, A. and Miller, R. (2014) *Online harassment, defamation, and hateful speech: A primer of the legal landscape.* [Online; accessed 27-December-2017]. Available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2447904`.

Massey, C. R. (1992) ‘Hate Speech, Cultural Diversity, and the Foundational Paradigms of Free Expression’, *UCLA L. Rev.* 40, p. 103.

Matsuda, M. J. (1989) ‘Public response to racist speech: Considering the victim’s story’, *Michigan Law Review*, 87(8), pp. 2320–2381.

Mayer-Schönberger, V and Cukier, K. (2013) *Big Data–A Revolution That Will Transform How We Live, Think and Work.* London: John Murray.

Mayor of London (2017) *Mayor launches new unit to tackle online hate crime.* [Online; accessed 27-December-2017]. Available at `https://www.london.gov.uk/press-releases/mayoral/mayor-launches-unit-to-tackle-online-hate-crime`.

McEnery, T., McGlashan, M., and Love, R. (2015) ‘Press and social media reaction to ideologically inspired murder: The case of Lee Rigby’, *Discourse & Communication*, 9(2), pp. 237–259.

McGarry, A. (2017) *Romaphobia: the last acceptable form of racism.* London: Zed Books Ltd.

Mead, G. H. (1967) *Mind, self, and society: From the standpoint of a social behaviorist (Works of George Herbert Mead, Vol. 1).* Chicago: University of Chicago Press.

Mehdad, Y. and Tetreault, J. (2016) ‘Do Characters Abuse More Than Words?’, in: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 299–303.

Meier, R. F. and Miethe, T. D. (1993) ‘Understanding theories of criminal victimization’, *Crime and justice*, 17, pp. 459–499.

Melero, M., Ruiz Costa-Jussà, M., Domingo, J., Marquina, M., and Quixal, M. (2012) ‘Holaaa!! writin like u talk is kewl but kinda hard 4 NLP’, in: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC’12)*, pp. 3794–3800.

Merton, R. K. (1938) ‘Social structure and anomie’, *American sociological review*, 3(5), pp. 672–682.

Mills, C. W. (2014) *The racial contract.* Ithaca, NY: Cornell University Press.

Miró, F. (2014) ‘Routine activity theory’, *The Encyclopedia of Theoretical Criminology*,

Mohammed, M., Khan, M. B., and Bashier, E. B. M. (2016) *Machine Learning: Algorithms and Applications.* Boca Raton, Fl: CRC Press.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012) *Foundations of machine learning.* Boston, MA: MIT press.

Mondal, M., Silva, L. A., and Benevenuto, F. (2017) ‘A Measurement Study of Hate Speech in Social Media’, in: *Proceedings of the 28th ACM Conference on Hypertext and Social Media.* ACM, pp. 85–94.

Moran, M. (1994) ‘Talking About Hate Speech: A Rhetorical Analysis of American and Canadian Approaches to the Regulation of Hate Speech’, *Wis. L. Rev.* P. 1425.

— (2003) *Rethinking the reasonable person: an egalitarian reconstruction of the objective standard.* Oxford University Press on Demand.

Morning, A. (2011) *The nature of race: How scientists think and teach about human difference*. Oakland, CA: Univ of California Press.

Morstatter, F. and Liu, H. (2017) 'Discovering, assessing, and mitigating data bias in social media', *Online Social Networks and Media*, 1, pp. 1–13.

Moule Jr, R. K. and Powers, R. A. (2019) 'An experimental assessment of third parties as potential guardians: victim gender, conflict, and individual perceptions of social situations', *Journal of interpersonal violence*, p. 0886260519827664.

Mueller, M. L. (2010) *Networks and states: The global politics of Internet governance*. Boston, MA: MIT press.

Murthy, D. (2012) 'Towards a sociological understanding of social media: Theorizing Twitter', *Sociology*, 46(6), pp. 1059–1073.

— (2013) *Twitter: Social communication in the Twitter age*. Chichester: Wiley.

Murthy, D. and Bowman, S. A. (2014) 'Big Data solutions on a small scale: Evaluating accessible high-performance computing for social research', *Big Data & Society*, 1(2), p. 2053951714559105.

Myers, S. A. and Leskovec, J. (2014) 'The bursty dynamics of the twitter information network', in: *Proceedings of the 23rd international conference on World wide web*. ACM, pp. 913–924.

Nagel, E. van der and Frith, J. (2015) 'Anonymity, pseudonymity, and the agency of online identity: Examining the social practices of r/Gonewild', *First Monday*, 20(3), issn: 13960466. doi: `10.5210/fm.v20i3.5615`. Available at `http://www.ojphi.org/ojs/index.php/fm/article/view/5615`.

Navarro, J. N., Clevenger, S., Beasley, M. E., and Jackson, L. K. (2017) 'One step forward, two steps back: Cyberbullying within social networking sites', *Security Journal*, 30(3), pp. 844–858.

Nelson, R. A., McCarthy, D. D., Malys, S., Levine, J., Guinot, B., Fliegel, H. F., Beard, R. L., and Bartholomew, T. (2001) 'The leap second: its history and possible future', *Metrologia*, 38(6), p. 509.

Neumayer, R. (2006) *Clustering based ensemble classification for spam filtering.* na.

Newburn, T. (2017) *Criminology.* London: Routledge.

Newman, M. E. (2008) 'The mathematics of networks', *The new palgrave encyclopedia of economics*, 2(2008), pp. 1–12.

Newton, A. D., Partridge, H., and Gill, A. (2014) 'Above and below: measuring crime risk in and around underground mass transit systems', *Crime science*, 3(1), p. 1.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., and Chang, Y. (2016) 'Abusive language detection in online user content', in: *Proceedings of the 25th International Conference on World Wide Web.* International World Wide Web Conferences Steering Committee, pp. 145–153.

noswearing.com (2018) *Swear Word List & Curse Filter.* [Online; accessed 22-February-2018]. Available at `https://www.noswearing.com/dictionary`.

O'Brien, D. T., Sampson, R. J., and Winship, C. (2015) 'Ecometrics in the age of big data: Measuring and assessing 'Broken windows' using large-scale administrative records', *Sociological Methodology*, 45(1), pp. 101–147.

O'Dea, C. J. and Saucier, D. A. (2017) 'Negative emotions versus target descriptions: Examining perceptions of racial slurs as expressive and descriptive', *Group Processes & Intergroup Relations*, 20(6), pp. 813–830.

Office for National Statistics (2015) *Improving crime statistics in England and Wales*. [Online; accessed 27-November-2017]. Available at `http://webarchive.nationalarchives.gov.uk/20160105191801/http://www.ons.gov.uk/ons/rel/crime-stats/crime-statistics/year-ending-june-2015/sty-fraud.html`.

Olteanu, A., Talamadupula, K., and Varshney, K. R. (2017) 'The limits of abstract evaluation metrics: The case of hate speech detection', in: *Proceedings of the 2017 ACM on Web Science Conference*. ACM, pp. 405–406.

O'Neill, A. (2017) 'Hate Crime, England and Wales, 2016/17, Statistical Bulletin 17/17', *Crime and Policing Statistics. London: Home Office*,

Parekh, B. et al. (2012) 'Is there a case for banning hate speech?', *The content and context of hate speech: Rethinking regulation and responses*, pp. 37–56.

Park, J. H. and Fung, P. (2017) 'One-step and Two-step Classification for Abusive Language Detection on Twitter', in: *ALW1: 1st Workshop on Abusive Language Online*. Association for Computational Linguistics. Bristol.

Pei, S., Muchnik, L., Andrade Jr, J. S., Zheng, Z., and Makse, H. A. (2014) 'Searching for superspreaders of information in real-world social media', *Scientific reports*, 4, p. 5547.

Pei, S., Morone, F., and Makse, H. A. (2017) *Theories for influencer identification in complex networks*. Springer Nature.

Pentreath, N. (2015) *Machine Learning with Spark*. Birmingham: Packt Publishing Ltd.

Perez, S. (2017) *Twitter adds more anti-abuse measures focused on banning accounts, silencing bullying*. [Online; accessed 27-November-2017]. Available at `https://techcrunch.com/2017/03/01/twitter-adds-more-anti-abuse-measures-focused-on-banning-accounts-silencing-bullying/`.

Peterson, J. and Densley, J. (2017) ‘Cyber violence: What do we know and where do we go from here?’, *Aggression and violent behavior*, 34, pp. 193–200.

Pitsilis, G. K., Ramampiaro, H., and Langseth, H. (2018) ‘Detecting Offensive Language in Tweets Using Deep Learning’, *arXiv preprint arXiv:1801.04433*,

Podgorelec, V., Kokol, P., Stiglic, B., and Rozman, I. (2002) ‘Decision trees: an overview and their use in medicine’, *Journal of medical systems*, 26(5), pp. 445–463.

Ponterotto, J. G. (2005) ‘Qualitative research in counseling psychology: A primer on research paradigms and philosophy of science.’, *Journal of counseling psychology*, 52(2), p. 126.

Powers, D. M. (2011) ‘Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation’, *Journal of Machine Learning Technologies*, 2(1), pp. 37–63.

Pramanik, M. I., Lau, R. Y., and Chowdhury, M. K. H. (2016) ‘Automatic Crime Detector: A Framework for Criminal Pattern Detection in Big Data Era’, in: PACIS, p. 311.

Qian, H. and Scott, C. R. (2007) ‘Anonymity and self-disclosure on weblogs’, *Journal of Computer-Mediated Communication*, 12(4), pp. 1428–1451.

Quantcast (2018) *twitter.com.* [Online; accessed 22-February-2018]. Available at `https://www.quantcast.com/measure/twitter.com`.

Quraishi, M. and Philburn, R. (2015) *Researching racism: A guidebook for academics and professional investigators.* London: Sage.

R v Sheppard and Whittle (2010). EWCA Crim 65.

Ramos, J. et al. (2003) ‘Using tf-idf to determine word relevance in document queries’, in: *Proceedings of the first instructional conference on machine learning.* Vol. 242, pp. 133–142.

Ratcliffe, J. (2010) 'Crime mapping: spatial and temporal challenges', in: *Handbook of quantitative criminology.* New York: Springer, pp. 5–24.

Ratcliffe, J. H. (2016) *Intelligence-led policing.* London: Routledge.

Rattansi, A. (2007) *Racism: A very short introduction.* Vol. 161. Oxford: Oxford University Press.

Razavi, A. H., Inkpen, D., Uritsky, S., and Matwin, S. (2010) 'Offensive language detection using multi-level classification', in: *Canadian Conference on Artificial Intelligence.* New York: Springer, pp. 16–27.

Remrey, L. P. (2016) 'Surveillance in Cyberspace: Applying Natural and Place Manager Surveillance to System Trespassing'. PhD thesis.

rich (2016) *Hive Orc table with text file.* [Online; accessed 27-November-2017]. Available at `https://community.hortonworks.com/questions/24039/hive-orc-table-with-text-file.html`.

Riloff, E., Qadir, A., Surve, P., De Silva, L., Gilbert, N., and Huang, R. (2013) 'Sarcasm as contrast between a positive sentiment and negative situation', in: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 704–714.

Riquelme, F. and González-Cantergiani, P. (2016) 'Measuring user influence on Twitter: A survey', *Information Processing & Management*, 52(5), pp. 949–975.

Ritter, A., Clark, S., Etzioni, O., et al. (2011) 'Named entity recognition in tweets: an experimental study', in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics, pp. 1524–1534.

rjurney (2012) *Why isn't Hadoop implemented using MPI?* [Online; accessed 27-November-2017]. Available at `https://stackoverflow.com/questions/4590674/why-isnt-hadoop-implemented-using-mpi`.

Roberts, C., Innes, M., Preece, A., and Rogers, D. (2017) ‘After Woolwich: Analyzing open source communications to understand the interactive and multi-polar dynamics of the arc of conflict’, *The British Journal of Criminology*, 58(2), pp. 434–454.

Rogers, R. et al. (2014) ‘Foreword: Debanalising Twitter: The transformation of an object of study’, *Twitter and society*, pp. 9–27.

Romero, D. M., Galuba, W., Asur, S., and Huberman, B. A. (2011) ‘Influence and passivity in social media’, in: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 18–33.

Rosemond, E. (2017) *This stereotype is killing black children.* [Online; accessed 7-March-2018]. Available at `https://www.washingtonpost.com/opinions/this-stereotype-is-killing-black-children/2017/02/10/2c06fa14-e249-11e6-a547-5fb9411d332c_story.html?utm_term=.bef8159e8962`.

Rosen, A. (2017) *Tweeting Made Easier.* [Online; accessed 7-December-2017]. Available at `https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html`.

Ross, B., Rist, M., Carbonell, G., Cabrera, B., Kurowsky, N., and Wojatzki, M. (2017) ‘Measuring the reliability of hate speech annotations: The case of the european refugee crisis’, *arXiv preprint arXiv:1701.08118*,

Ruhnau, B. (2000) ‘Eigenvector-centrality - a node-centrality?’, *Social networks*, 22(4), pp. 357–365.

Sae-Bae, N., Sun, X., Sencar, H. T., and Memon, N. D. (2014) ‘Towards automatic detection of child pornography’, in: *Image Processing (ICIP), 2014 IEEE International Conference on.* IEEE, pp. 5332–5336.

Safavian, S. R. and Landgrebe, D. (1991) 'A survey of decision tree classifier methodology', *IEEE transactions on systems, man, and cybernetics*, 21(3), pp. 660–674.

Saif, H., Fernández, M., He, Y., and Alani, H. (2014) 'On stopwords, filtering and data sparsity for sentiment analysis of twitter', in: *LREC 2014, Ninth International Conference on Language Resources and Evaluation.* Reykjavik, Iceland, pp. 810–817.

Sammut, C. and Webb, G. I. (2017) *Encyclopedia of Machine Learning and Data Mining.* Springer. New York.

Sampson, R., Eck, J. E., and Dunham, J. (2010) 'Super controllers and crime prevention: A routine activity explanation of crime prevention success and failure', *Security Journal*, 23(1), pp. 37–51.

Samuel, A. (1959) 'Some Studies in Machine Learning Using the Game of Checkers', *IBM Journal of Research and Development*, 3(3), p. 210.

Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., and Schwartz, H. A. (2014) 'Developing age and gender predictive lexica over social media', in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1146–1151.

Schmidt, A. and Wiegand, M. (2017) 'A survey on hate speech detection using natural language processing', in: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pp. 1–10.

Schneider, J. (1997) *Cross Validation.* [Online; accessed 1-March-2018]. Available at `https://www.cs.cmu.edu/~schneide/tut5/node42.html`.

Sebastiani, F. (2002) 'Machine learning in automated text categorization', *ACM computing surveys (CSUR)*, 34(1), pp. 1–47.

Sellars, A. (2016) *Defining Hate Speech.* Tech. rep. Berkman Klein Center for Internet and Society at Harvard University.

Shalev-Shwartz, S. and Ben-David, S. (2014) *Understanding machine learning: From theory to algorithms.* Cambridge: Cambridge University Press.

Sherman, L. W., Gartin, P. R., and Buerger, M. E. (1989) 'Hot spots of predatory crime: Routine activities and the criminology of place', *Criminology*, 27(1), pp. 27–56.

Shin, Y. (2015) 'Application of boosting regression trees to preliminary cost estimation in building construction projects', *Computational intelligence and neuroscience*, 2015, p. 1.

Shvachko, K., Kuang, H., Radia, S., and Chansler, R. (2010) 'The hadoop distributed file system', in: *Mass storage systems and technologies (MSST), 2010 IEEE 26th symposium on.* Ieee, pp. 1–10.

Silva, L. A., Mondal, M., Correa, D., Benevenuto, F., and Weber, I. (2016) 'Analyzing the Targets of Hate in Online Social Media.', in: *ICWSM*, pp. 687–690.

Silverman, D. (2015) *Interpreting qualitative data.* London: Sage.

Simi, P. and Futrell, R. (2015) *American Swastika: Inside the white power movement's hidden spaces of hate.* Lanham, MD: Rowman & Littlefield.

Smith, D. T. (2017) 'On the therapeutic uses of racism in other countries', in: Epstein, C. (eds.) *Against International Relations Norms: Postcolonial Perspectives.* London:Routledge, pp. 123–137.

Sooben, P. N. (1990) *The Origins of the Race Relations Act.* 12. Centre for Research in Ethnic Relations, University of Warwick Coventry.

Sood, S., Antin, J., and Churchill, E. (2012a) 'Profanity use in online communities', in: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, pp. 1481–1490.

Sood, S. O., Churchill, E. F., and Antin, J. (2012b) 'Automatic identification of personal insults on social news sites', *Journal of the Association for Information Science and Technology*, 63(2), pp. 270–285.

Spark (2015a) *Spark ML Programming Guide*. [Online; accessed 22-February-2018]. Available at `https://spark.apache.org/docs/1.2.2/ml-guide.html`.

— (2015b) *Spark Release 1.3.0*. [Online; accessed 22-February-2018]. Available at `https://spark.apache.org/releases/spark-release-1-3-0.html`.

— (2016a) *Dataset API on top of Catalyst/DataFrame*. [Online; accessed 22-February-2018]. Available at `https://issues.apache.org/jira/browse/SPARK-9999`.

— (2016b) *Spark Release 1.6.0*. [Online; accessed 22-February-2018]. Available at `https://spark.apache.org/releases/spark-release-1-6-0.html`.

— (2017a) *Evaluation Metrics - RDD-based API*. [Online; accessed 16-December-2017]. Available at `https://spark.apache.org/docs/2.2.0/mllib-evaluation-metrics.html`.

— (2017b) *Machine Learning Library (MLlib) Guide*. [Online; accessed 22-February-2018]. Available at `https://spark.apache.org/docs/2.2.0/ml-guide.html`.

— (2017c) *stopwords*. [Online; accessed 22-February-2018]. Available at `https://github.com/apache/spark/tree/master/mllib/src/main/resources/org/apache/spark/ml/feature/stopwords`.

Spark (2018a) *Multilayer perceptron classifier.* [Online; accessed 22-February-2018]. Available at `https://spark.apache.org/docs/2.2.0/ml-classification-regression.html#multilayer-perceptron-classifier`.

— (2018b) *TF-IDF.* [Online; accessed 22-February-2018]. Available at `https://spark.apache.org/docs/2.2.0/mllib-feature-extraction.html#tf-idf`.

Spertus, E. (1997) 'Smokey: Automatic recognition of hostile messages', in: *AAAI/IAAI*, pp. 1058–1065.

Spider-Byte (2016) *Alex Is A Stupid Nigger.* [Online; accessed 26-February-2018]. Available at `http://knowyourmeme.com/memes/alex-is-a-stupid-nigger`.

Strauss, A. and Corbin, J. M. (1990) *Basics of qualitative research: Grounded theory procedures and techniques.* Newbury Park, CA: Sage Publications, Inc.

Strickland, P. and Douse, D. (2012) *Insulting words or behaviour: Section 5 of the Public Order Act 1986.*

Stryker, S. (1980) *Symbolic interactionism: A social structural version.* San Francisco, CA: Benjamin-Cummings Publishing Company.

Sugiyama, M. (2015) *Introduction to statistical machine learning.* Burlington, MA: Morgan Kaufmann.

Suler, J. (2004) 'The online disinhibition effect', *Cyberpsychology & behavior*, 7(3), pp. 321–326.

Sutherland, E. (1947) *Principles of criminology (4th ed.)* Philadelphia: J. B. Lippincott.

Sykes, G. M. and Matza, D. (1957) 'On Neutralizing Delinquent Self-images', *American Sociological Review*, 22, pp. 667–670.

Taylor, R. C. (2010) 'An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics', *BMC bioinformatics*, 11(12), S1.

tfidf (2018) *What does tf-idf mean?* [Online; accessed 22-February-2018]. Available at `http://www.tfidf.com/`.

Tierney, J. and O'Neill, M. (2013) *Criminology: Theory and context.* London: Routledge.

Tuli, F. (2011) 'The basis of distinction between qualitative and quantitative research in social science: Reflection on ontological, epistemological and methodological perspectives', *Ethiopian Journal of Education and Sciences*, 6(1),

Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., and Daelemans, W. (2016) 'A dictionary-based approach to racism detection in Dutch social media', *arXiv preprint arXiv:1608.08738*,

Twitter (2017) *Tweet object - Twitter Developers.* `https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object`. [Online; accessed 21-November-2017].

— (2018) *Twitter Health Metrics Proposal Submission.* [Online; accessed 8-March-2018]. Available at `https://blog.twitter.com/official/en_us/topics/company/2018/twitter-health-metrics-proposal-submission.html`.

Twomey, A. (2017) 'A Civil Society Perspective on Anti-Traveller and Anti-Roma Hate: Connecting Online to On the Street', in: *Critical Perspectives on Hate Crime.* Springer, pp. 355–366.

Van Hee, C., Lefever, E., Verhoeven, B., Mennes, J., Desmet, B., De Pauw, G., Daelemans, W., and Hoste, V. (2015) 'Detection and fine-grained classification of cyberbullying events', in: *International Conference Recent Advances in Natural Language Processing (RANLP)*, pp. 672–680.

Vapnik, V. (2013) *The nature of statistical learning theory.* New York: Springer.

Vavilapalli, V. K., Murthy, A. C., Douglas, C., Agarwal, S., Konar, M., Evans, R., Graves, T., Lowe, J., Shah, H., Seth, S., et al. (2013) 'Apache hadoop yarn: Yet another resource negotiator', in: *Proceedings of the 4th annual Symposium on Cloud Computing.* ACM, p. 5.

Vazsonyi, A. T., Machackova, H., Sevcikova, A., Smahel, D., and Cerna, A. (2012) 'Cyberbullying in context: Direct and indirect effects by low self-control across 25 European countries', *European Journal of Developmental Psychology*, 9(2), pp. 210–227.

Vosoughi, S., Zhou, H., and Roy, D. (2016) 'Enhanced twitter sentiment classification using contextual information', *arXiv preprint arXiv:1605.05195*,

Walden, I. (2007) *Computer crimes and digital investigations.* Oxford: Oxford University Press.

Wall, D. (2007) *Cybercrime: The transformation of crime in the information age.* Vol. 4. Polity.

Wall, D. S. (2017) 'Towards a Conceptualisation of Cloud (Cyber) Crime', in: *International Conference on Human Aspects of Information Security, Privacy, and Trust.* Springer, pp. 529–538.

Wall David, S (2004) 'The Internet as a Conduit for Criminal Activity', *Information Technology and the Criminal Justice System*, p. 77.

Wallace, B. C., Small, K., Brodley, C. E., and Trikalinos, T. A. (2011) 'Class imbalance, redux', in: *Data Mining (ICDM), 2011 IEEE 11th International Conference on.* IEEE, pp. 754–763.

Walters, M and Brown, R (2016) 'Preventing Hate Crime: Emerging Practices and Recommendations for the Improved Management of Criminal Justice Interventions', *Brighton: University of Sussex*, Ward, J. S. and Barker, A. (2013) 'Undefined by data: a survey of big data definitions', *arXiv preprint arXiv:1309.5821*,

Ward, K. D. (1997) 'Free Speech and the Development of Liberal Virtues: An Examination of the Controversies Involving Flag-Burning and Hate Speech', *U. Miami L. Rev.* 52, p. 733.

Warner, W. and Hirschberg, J. (2012) 'Detecting hate speech on the world wide web', in: *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, pp. 19–26.

Waseem, Z. and Hovy, D. (2016) 'Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter.', in: *SRW@ HLT-NAACL*, pp. 88–93.

Watt, J., Borhani, R., and Katsaggelos, A. (2016) *Machine Learning Refined: Foundations, Algorithms, and Applications*. Cambridge: Cambridge University Press.

Webster, C. (2007) *Understanding race and crime*. New York: McGraw-Hill.

White, T. (2012) *Hadoop: The definitive guide*. Sebastopol, CA: O'Reilly Media, Inc.

Wikipedia (2017) *List of ethnic slurs — Wikipedia, The Free Encyclopedia*. [Online; accessed 21-November-2017]. Available at `https://en.wikipedia.org/w/index.php?title=List_of_ethnic_slurs\&oldid=810910051`.

Williams, B. N. and Stahl, M. (2008) 'An analysis of police traffic stops and searches in Kentucky: a mixed methods approach offering heuristic and practical implications', *Policy Sciences*, 41(3), pp. 221–243.

Williams, M. L. and Burnap, P. (2015) 'Cyberhate on social media in the aftermath of Woolwich: A case study in computational criminology and big data', *British Journal of Criminology*, 56(2), pp. 211–238.

Williams, M. L. and Pearson, O. (2016) *Hate crime and bullying in the age of social media.* Tech. rep. Welsh Government.

Woodie, A. (2017a) *Anatomy of a Hadoop Project Failure.* [Online; accessed 27-November-2017]. Available at `https://www.datanami.com/2017/03/17/anatomy-hadoop-project-failure/`.

— (2017b) *Hadoop Has Failed Us, Tech Experts Say.* [Online; accessed 27-November-2017]. Available at `https://www.datanami.com/2017/03/13/hadoop-failed-us-tech-experts-say/`.

WPformers.com (2017) *How much data does Google handle??* [Online; accessed 22-January-2018]. Available at `https://www.wpformers.com/google-datacenter-capacity/`.

Wulczyn, E., Thain, N., and Dixon, L. (2017) 'Ex machina: Personal attacks seen at scale', in: *Proceedings of the 26th International Conference on World Wide Web.* International World Wide Web Conferences Steering Committee, pp. 1391–1399.

Xiang, G., Fan, B., Wang, L., Hong, J., and Rose, C. (2012) 'Detecting offensive tweets via topical feature discovery over a large scale twitter corpus', in: *Proceedings of the 21st ACM international conference on Information and knowledge management.* ACM, pp. 1980–1984.

Xiangrui (2016) *Switch RDD-based MLlib APIs to maintenance mode in Spark 2.0.* [Online; accessed 22-January-2018]. Available at `http://apache-spark-developers-list.1001551.n3.nabble.com/Switch-RDD-based-MLlib-APIs-to-maintenance-mode-in-Spark-2-0-td17033.html`.

Xu, J.-M., Jun, K.-S., Zhu, X., and Bellmore, A. (2012) 'Learning from bullying traces in social media', in: *Proceedings of the 2012 conference of the North American chap-*

ter of the association for computational linguistics: Human language technologies. Association for Computational Linguistics, pp. 656–666.

Yar, M. (2005) ‘The Novelty of ‘Cybercrime’ An Assessment in Light of Routine Activity Theory’, *European Journal of Criminology*, 2(4), pp. 407–427.

Ye, J., Chow, J.-H., Chen, J., and Zheng, Z. (2009) ‘Stochastic gradient boosted distributed decision trees’, in: *Proceedings of the 18th ACM conference on Information and knowledge management.* ACM, pp. 2061–2064.

Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., and Edwards, L. (2009) ‘Detection of harassment on web 2.0’, *Proceedings of the Content Analysis in the WEB*, 2, pp. 1–7.

Zaharia, M., Xin, R. S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M. J., et al. (2016) ‘Apache spark: a unified engine for big data processing’, *Communications of the ACM*, 59(11), pp. 56–65.

Zhang, D. and Lee, W. S. (2003) ‘Question classification using support vector machines’, in: *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.* ACM, pp. 26–32.

Zhong, H., Li, H., Squicciarini, A. C., Rajtmajer, S. M., Griffin, C., Miller, D. J., and Caragea, C. (2016) ‘Content-Driven Detection of Cyberbullying on the Instagram Social Network.’, in: *IJCAI*, pp. 3952–3958.