QUEEN'S UNIVERSITY BELFAST

Rapid Hate Speech Detection

Developing a Machine Learning Method that can be deployed on Emerging Social Media Trends with Large Volumes of Hate Speech.

Terry McElroy

Word Count: 15,360

Submitted in part fulfilment of the degree of Master of Science in Business Analytics

October 2020

Queen's University Management School

Declaration

This is to certify that:

I. The dissertation comprises only my original work;

II. Due acknowledgement has been made in the text to all other materials used;

III. No portion of the work referred to in the dissertation has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Acknowledgements

Throughout the writing of this dissertation I have received a great deal of help to complete this project. I would like to take this time to thank Prof. Min Zhang for his guidance and assistance throughout this project. I would also like thank all the tutors that have taught me throughout my Masters, especially Dr. Byron Graham and Dr. Laura Steele who assisted me with resources and insights at the very beginning of the project.

Finally, I would like to give a special mention to Jess who sat down and helped me identify all the times I incorrectly used a semi-colon.

Abstract

Social media companies are constantly being scrutinised by the media and the public for their slow response times when detecting and dealing with hate speech on their platform. Larger volumes of posts and manually sorting them are a reason why they are slow to react. Therefore, machine learning techniques should be applied to improve response times. Previous research into hate speech detection on social media has centred on larger themes of hate speech and develop models that detect generalizable hate speech against different groups. However, this previous work has failed at establishing a framework that can be used to detect hate in emerging trends on social media platforms.

This work will focus on developing a machine learning strategy that can be implemented on any emerging trend with high volumes of hate speech on social media within a reasonably quick time frame. To do this, a dataset of 5,000 tweets on the COVID-19 pandemic will be used to develop a machine learning model that is able to classify if a post contains hate speech towards East-Asian people. From the analysis of the data the best method was a gradient boosted machine that had an F1-score of 0.649 on both small and large volumes of unseen data. This model was tuned for optimal performance in 90 minutes. This research has shown that it is possible to develop a machine learning model that can detect large volumes of hate speech on an emerging trend. Whilst gradient boosted was the best preforming model in this analysis, new trends result in new data and thus different methods could be more appropriate, therefore it is important for this strategy to be developed further.

Contents

Table of Figures

1. Background Information

Social media has enjoyed meteoric success since the beginning of Web 2.0. Platforms, like Facebook and Twitter, have enabled people to engage with current affairs and widen their knowledge of issues that affect different people (Thota, 2018). However, these interaction-based platforms can also work against the interests of a pluralistic society. Certain topics, such as the COVID-19 outbreak, can attract global commentary from social media users. These moments can often be a trigger for hate speech, resulting in incendiary rhetoric being spread across the platforms (Burnap and Williams, 2015). Furthermore, the rise of right-wing populism that has spread across the world has created a politically tribal environment online, where a person is either fully supportive of the ideology or they denounce anyone with the faintest association to the movement (Main, 2018). These opposite views have resulted in hate speech becoming a large problem for social media companies today as they require speed and agility to effectively manage the issue. This project will focus on how hate speech posts can be detected using machine learning methods and will use the COVID-19 pandemic as the trigger moment for analysing hate speech data.

1.1 Business Problem

Hate speech on social media platforms garners constant complaints from the media and government. These complaints normally circulate around one issue: the response time to react to hate speech. In July 2020, the Mayor of London penned an open letter to Twitter and Instagram regarding their inaction on anti-Semitism from a celebrity's posts, in which he criticised their slow reaction time and stated that, *"We see again how [social media] can be used to spread hatred and division"* (Kahn, 2020). Whilst it may seem that one person

spreading hate might not be a large issue, it influences wider society as social media allows thousands of indirect recipients to view the same posts (Ullmann and Tomalin, 2019). This can then expose people to hateful ideologies, all because the companies are too slow to react to the spread of hate speech on their platforms.

One of the main issues that social media companies face when attempting to tackle the issue on their sites is the complex nature of defining hate speech. After a sweeping analysis of legal, charities, social media and academia sources, Fortuna and Nunes (2018, p.5) have been able to derive a definition that identifies hate speech as:

> *"language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical appearance, religion, descent, national or ethnic origin, sexual orientation, gender identity or other, and it can occur with different linguistic styles, even in subtle forms or when humour is used"*

Adopting this definition of hate speech, it is often the case that social media users can be exposed to hate as secondary victims because they belong to a specific group or share a similar characteristic with the intended target (Schweppe et al., 2020). Inaction by social media companies enables the spread of hate speech on the platform as the indirect nature of the attacks can affect large swathes of the user base rather than just the initial recipient.

Triggered by the COVID-19 pandemic, there has been a significant increase in levels of Sinophobia online, with the Secretary General of the UN condemning the "tsunami of hate and xenophobia" on social media (Guterres, 2020). This rise of online hate speech has seeped into mainstream debates on these issues. Influential political figures have adopted a tone reminiscent of early propaganda techniques of Nazi Germany through using highly racist rhetoric such as "Kung Flu" and "China Virus" when discussing the pandemic (Heinze, 2016;

Coleman, 2020; Trump, 2020). Since these figures replicate the language seen online, it legitimises the hateful discourse. This becomes the responsibility of social media companies to prevent the spread of hate speech on their platform.

Since social media is facilitating the spread of hatred it is therefore their responsibility to act against it and prevent the vulnerable from being attacked. The reasons are fourfold:

1) A duty of care to protected groups: Social media companies should have a duty of care towards their users with specific characteristics and must protect them from content that is harmful.

2) Duty of care to all users: Social media moments allow large engagement on certain issues. If this issue is based in hate, then it can radicalise a person in adopting these viewpoints without understanding the full consequences of participating.

3) Loss of Revenue for the company: Social media companies rely on advertisements to provide a service for users to enjoy. However, this can leave the advertiser vulnerable to which social media moments they become associated with. If social media companies are not vigilant on hate speech, then sponsors could pull their adverts if they deem it too risky to be associated with potential hateful trends. This is already being seen in the US where companies are boycotting Facebook due to their inaction on hate speech plaguing the platform (Dwoskin and Telford, 2020).

4) Duty of care to their workers: Social media has content moderators that will examine reported posts. These posts tend to be the most graphic content on social media sites and are viewed by real people which can be detrimental to the mental health of the moderators (Dwoskin, 2019). However, this is a contentious issue given that a Facebook contractor made

it mandatory for employees to sign a PTSD waiver before starting their job, which is believed to be a protection against employee lawsuits (Murgia, 2020).

1.2 Project Aim

As social media continues to rise in popularity, instances of trigger moments are going to result in a larger spread of hate speech on the platforms. Therefore, the aims of this project are to:

1) Develop a Machine learning model that can accurately detect hate speech in a trending topic, in this instance COVID-19, as the volume of posts is too great for a person operated detection method.
2) The speed of development is key, as hate speech is instantaneous in its attack and can spread quickly amongst people interacting with these posts.
3) For successful completion of aim 1 and 2, a framework can be developed for future trigger moments for hate speech.

1.3 Literature Review

To get an understanding of the topic, the literature review will be split up into two parts. Firstly, I will examine the projects viability. This will be done by reviewing the literature to understand how Sinophobia can seep into a larger discussion of hate against East-Asian's, what social media companies currently do to counter hate speech, and discuss the ethical dilemma if censoring hate speech infringes on civil liberties. The second part will consider

what techniques and methods have been developed to automatically detect hate speech and how can these be implemented by social media companies.

1.3.1 Project Viability

Hate speech is a sub-division of hate crime. It has historically been used as political propaganda tool to incite fear in their supporters about a threat that another group poses simply by existing. However, in the age of social media it has become a lot easier for hate speech to propagate to larger audiences through direct and in-direct recipients on these platforms (Schweppe et al., 2020). As social media audiences react to current events it can easily stir up stereotypes and prejudice against groups who are being discussed in the news cycle resulting in hate speech.

Sinophobia and Anti-East Asian Sentiments

Hatred towards people of Chinese origin is not new, yet it has risen at alarming rates since the start of the COVID-19 pandemic. Since the days of Genghis Khan, the West has feared East-Asian people and used the term *Yellow Peril* to describe their anxieties (Marchetti, 1993). These ideas are firmly rooted in white supremacy, where the "yellow race" are perceived to be a threat to the Western way of life (Kawai, 2006). Over time the term Yellow Peril has been used to describe fears of many different East-Asian ethnicities. During 1800's America, Chinese immigrants were the focus of Yellow Peril with the most significant exclusionary practice coming from the US government, where it was signed into law that they could never be considered for citizenship as they were seen as an invading species (Gover et al., 2020). The Yellow Peril was regenerated after the attack on Pearl Harbor, with the main target now being Japanese people, this led to the President of the United States establishing internment

camps for Japanese-Americans during WWII (Howard, 2008; Gover et al., 2020). After the defeat of the Japanese Empire the Yellow Peril shifted to communist China, with Chinese-Americans being suspected of treason and espionage (Kawai, 2006). Also, due to the geopolitics of the Cold War and the US' military involvement in eastern Asia, we have seen many renditions of the Yellow Peril with Vietnamese and Koreans being the focus at a point (Gover et al., 2020).

However in more recent times, the focus of the Yellow Peril has shifted to people of Chinese origin due to their strengthening economic and military powers (Elwell et al., 2007). The Pew Research Center (2019) did a global survey of people's attitudes towards China and 41% of respondents have an unfavourable opinion of the country, with a further 58% of people stating China's military power is bad for their own country's national security, thus their way of life. This can be seen to be rooted in the origins of the Yellow Peril stereotype. This informs us that, in Western society, anti-Chinese sentiment is not solely focused to people from China but can have a spill over effect to other ethnicities that can embody the Yellow Peril. The implication on this project's validity means it cannot solely focus on hate against people of Chinese origin but all people of East-Asian origin. Additionally, the data being used will be in the English language and thus most likely to be through the gaze of Westerners where the Yellow Peril is most feared.

Social Media's fight against hate speech

Web 2.0 has introduced platforms where people can readily express themselves and create content that internet users can consume. Social Media falls under this umbrella term and is at the forefront of the public's preferred medium to express their views to the masses. With this ability to publish views, a minority of people publish views that are designed to be hateful.

Social Media companies do not want to be responsible for these messages propagating so they have a team that moderate their sites (Roberts, 2019). The internet being moderated is not a new concept, chat rooms were once a hotbed of controversial views and were used to harass people and thus had to be moderated. However, the volume of traffic and the variety of content was much smaller in chat rooms than it is today making the current task more difficult for a human operator (Gillespie, 2018; Roberts, 2019).

Additionally, a recent Twitter transparency report showed that Twitter had removed 2.9 Million posts from their site over a six month period because they breached the community guidelines. Of these 2.9 million posts over 80% were for hateful or abuse and harassing content (Twitter, 2020; Figure 1). The scale of these numbers on hate alone show that the job of a human content moderator is now exploitative, especially when whistle-blowers informed of PTSD waivers employees must sign before starting their job (Murgia, 2020). Therefore, it should be the duty of social media companies to improve on how they moderate content through the adoption of machine learning technologies using the vast amounts of data that they have available. For this reason, this project aims to build a framework that can be used on trending topics, to make a machine learning algorithm predict whether the post contains hate speech or not, and if so then it is recommended that the post is removed.

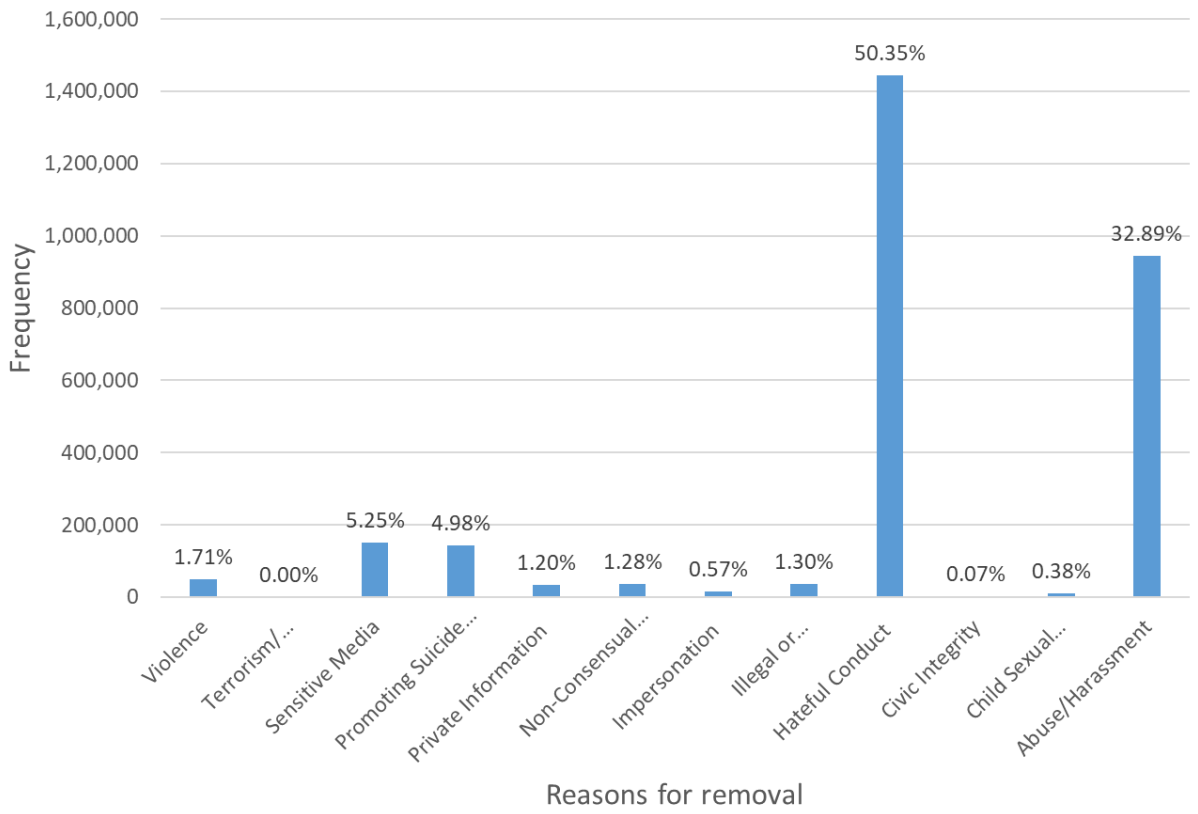

Figure 1. Content removed from twitter based on how they breach the guidelines, percentage frequency is shown above the bars. Data from July to December 2019. Figure recreated in MS Excel from Twitter (2020).

Does this censorship breach civil liberties?

The main opposition to the removal of hate speech is that this act infringes on human rights, and therefore unethical. The international covenant on civil and political rights (ICCPR), in which the majority of countries base their human rights legislation off, established that freedom of expression is a fundamental right that everyone is bestowed (UN, 1966). Dworkin (2009) agrees that this is an infringement of rights, arguing that for free speech, thereby democracy, to succeed then everyone should have the right to voice their opinion on a given subject no matter how abhorrent they may seem. It is only after these fringe viewpoints are considered that democracy can be legitimate.

However, Waldron (2012) disagrees with Dworkin, claiming that laws are already in place for extremely heinous expression, undermining Dworkin's point. Further to this, freedom of expression is not the only right that people are bestowed, and to argue that one right is greater than another, is a debate of moral pluralism (UN, 1966; Heinze, 2016). In fact, it is written into the ICCPR that people's freedom of expression should be subject to restrictions, allowing other people to fulfil their own rights of equality and life without harassment and intimidation (UN, 1966; Henize, 2016).

This leaves the idea that people are having their rights infringed upon is one of nonsense. During a trigger moment for hate, it is perfectly legitimate for the restriction of expression, as the target group is unable to practice their other human rights. Therefore, this makes the project ethically viable to target hate speech.

Conclusion

This section of the literature review examined the legitimacy of the project. We can see that the hate directed towards Chinese people is often incorrectly targeted towards other groups of East-Asian people, therefore they should be included in the project scope. We have also seen how the volume of posts that human content moderators have to examine is too large and negatively affects their health, thus a machine learning method should be applied. Finally, we have seen that the ethical argument against censoring hate speech is misleading because uncensored speech affects numerous rights of the targeted individual. From this review we can say that the project is viable.

1.3.2 Efforts to detect hate speech using Machine Learning

Machine learning research has been focused on text analytics for many decades. Tennant (1981) states that the reason to do so are twofold: 1) to provide summaries from a large

swathe of textual sources 2) to further understand the development of human language and how minds work. However, rising popularity of social media has shifted the motivation of research to predict the context behind the text and therefore be able to classify if it has hateful intent or not. This section of the literature review will examine the techniques that have been developed by previous studies in the automatic detection of hate speech and will answer the questions of data source, classification methods, and how could these models be implemented on social media?

What data is used to detect hate speech?

To detect hate speech, data containing hateful words must be used. Fortuna and Nunes (2018) conducted an analysis of papers that examined hate speech detection on social media and found that the overwhelming choice of source was Twitter followed by general websites (Table 1). The reasons that Twitter data is used so commonly in these cases is because they offer a popular platform for users to instantly share their viewpoint on topics free of charge (Twitter, n.d.; Fox, 2014). This mission statement enables users to publicly react to current events thus allowing researchers to gain up to date opinions that people are expressing including hateful ones (Pitsilis et al., 2018). Looking into hate speech due to the COVID-19 pandemic it is paramount for this research to use the most recent opinions from people. Furthermore, language is a complex system that it is ever evolving (Fitch, 2010), thus hate speech is constantly evolving and Twitter will be the best place to see this evolution in real time.

Social Media	Frequency
Twitter	16
General Sites	5
YouTube	3
Yahoo! Finance	2
American Jewish Congress sites	1
Ask.fm	1
Blogs	1
Documents	1
Facebook	1
formspring.me	1
myspace.com	1
Tumblr	1
Whisper	1
White Supremacist Forums	1
Yahoo News	1

Yahoo!	1

Table 1. Social media sites used for hate speech detection in previous studies. Table replicated from Fortuna and Nunes (2018).

How can hate speech be classified?

There are many different methods to detecting hate speech, a popular method in the wider natural language processing literature, for language modelling and document classification, is the Bag of words model. This method takes words from text and translates them into a binary vector so they can be affectively modelled (Brownlee, 2019). However, the length of the vector is depended on the number of different words in the text being analysed, this leaves the length of the vectors to be very long and quite difficult to understand (Ma, 2018) and it also loses the context of the text (Fortuna and Nunes, 2018). To reduce the size of these vectors words are often combined with one or two words surrounding it resulting in more clarity for the subject being discussed, this is called n-grams. Whenever examining a large volume of text any repeated phrases will be only be assigned a binary indicator once. Looking into anti-East-Asian sentiment, this n-gram method would be very useful; phrases such as *"corona virus"* or *"Chinese Communist Party"* will be counted and give a better understanding of the text making it easier to classify whether the tweet is hateful or not. This method has been used in several previous studies. Davidson et al. (2017) uses the bag of words and n(=2/3)-gram method for racist words and phrases allowing the model to better understand the significance of having certain words before and after the racist remark. This method can be adapted for this study by better identifying the context and nuance that people use in their tweets when discussing COVID-19.

The heart of this problem is a classification method, either a post will contain hate speech or it will not. Whilst using n-grams for this task is useful for visualising and

understanding context of a post it is not a robust means for prediction and can lead to misclassification (Burnap and Williams, 2015; Schmidt and Wiegand, 2017). Therefore it would be more appropriate to use a traditional classification model to build a robust model. The most modern techniques use different applications of Neural Networks (Mehdad and Tetreault, 2016; Al-Makhadmeh and Tolba, 2020). These methods boast high accuracy and are able to work even if a post has characters omitted to hide offensive terms, for example "*a$$h*le*" can still be detected by these algorithms. However, the aims of this project are clear, the model must be quick to implement and also work on a trending topic. Neural networks take quite a long time to be trained and also require a large amount of data so they would not be appropriate for this project (Hinton et al., 2012).

A more classical approach to hate speech detection is through using a support vector machine (SVM). These methods have been used to detect racist text messages in the infancy of hate speech detection by Greevy and Smeaton (2004). From this early work they have been improved upon through increased computational power and larger interest in the topic (Schmidt and Wigand, 2017; Fortuna and Nunes, 2018). This has led to work by MacAvaney et al. (2019) claiming to rival the neural network method for results but has the additional benefit of being easier to understand thus making it an optimal approach to the topic.

Further to this, random forests and gradient boosted machines are two techniques that are commonly used for classification problems, however, they are not commonly used in hate speech detection (Fortuna and Nunes, 2018). This could be down to them overfitting the training data making them unsuitable for prediction. To understand if this is the case they should also be used as a means of detection.

Implementation on social media

The objective of this research project is to detect hate speech in a trending topic. However, whilst there is research on how this can be achieved, very few actually suggest how to implement. Currently, social media is dependent on users reporting posts and then going to a moderator to see if it breaches terms of service or not (Laub, 2019; Murgia, 2020). Ullmann and Tomalin (2019) suggest social media should use a quarantining method. They suggest that if a post is flagged by an algorithm to contain potential hate speech then it should be delayed in it posting and should await for a third person/a more advance screening process to determine the hatefulness of the post. The idea is an interesting one, however, social media is used because of the lack of friction between publishing a piece of content. Ullman and Tomalin (2019) recognise this point and suggest that social media companies could introduce filters for a user's feed. They state that Twitter and Facebook already do this for potential graphic images or videos and say that by implementing it in this way it can protect the intended and unintended recipient of the post but also protect the sender's freedom of expression. Both these methods are good and valid, however, recent censorship of controversial persons has resulted them moving their views to fringe social media sites, such as Parler (Forster, 2020). These sites boast to not censor any views a user may have; this could create an echo chamber hate resulting in hate speech becoming a greater problem on social media than what it once was, showing that a balance should be struck when it comes to dealing with hate speech on the social media.

Conclusion

This part of the literature review has shown that for our study into hate speech detection we should be using data from Twitter due to its popularity for expressing public opinion. Tweets should be examined using an N-gram method for a better understanding of the context. This

then should be incorporated into SVM, random forest, and gradient boosted machine methods to classify the tweet as hateful or not. The implementation is an interesting area in the literature, however, research into the topic is sparse. This is likely due to social media companies being a private enterprise that make their decisions in-house rather than looking for answers in academia.

1.4 Overview of the Dissertation

The business problem and the review of relevant literature has identified the need for this research. Chapter Two will discuss the technical side of the techniques that will be applied in the project. Chapter Three will display the results from implementing these techniques, and will relay back to the literature review and discuss the how the results advance the current understandings surrounding the topic. Finally, Chapter Four will be a conclusion of the project, where limitations and future works will be discussed.

2. Methodology

As mentioned in 1.1 the research is focused in detecting anti-east Asian sentiment on the social media platform Twitter during the COVID-19 pandemic. The main aim of the study is to use a limited dataset for an accurate representation of a trending topic, and from this develop a robust model that can detect hate speech. A model will allow the development of a framework for future projects/problems that a social media company could face when hate speech trends on their platform. This section will be discussing the analytics methodology that will be followed throughout the project. It will also discuss data source and how the data is to be prepared before modelling section, and then further understanding of the technical parts of the analysis that were highlighted in section 1.3.3 are also discussed.

2.1 Analytics Methodology

To achieve the goals of this project a methodology similar to CRISP-DM will be followed. The CRISP-DM methodology is an industry standard for an analytical task to solve business problems, Figure 2 shows this methodology from start to finish.

Figure 2. The CRISP-DM cycle. Image reproduced from Vorhies, 2016.

Business Problem: This is detailed in section 1.1 of this report, traditionally this would come from a management position in an organisation that needs a solution to a business problem.

Data Understanding: This is exploratory data analysis to understand the structure and nature of the data.

Data Preparation: This stage will change the data from its raw nature before it is used for modelling.

Modelling: This stage can only happen after the following three stages have been complete or the analysis may not be successful. This is where different machine learning methods will be used to make various solutions to the problem.

Evaluation: This stage uses the models that have been developed and will test them against each other to see which is the most accurate and suitable for the problem.

Deployment: The selected model will be used to solve the business problem and will be using new and unseen data.

The version I will use will have a data preparation stage before data understanding because I am taking a fraction of a published dataset and will be editing it before the understanding phase. Final preparations will be made after the understanding phase before modelling, Figure 3 below. Also, the deployment phase is beyond the scope of this dissertation and instead will be the recommendations from the project.

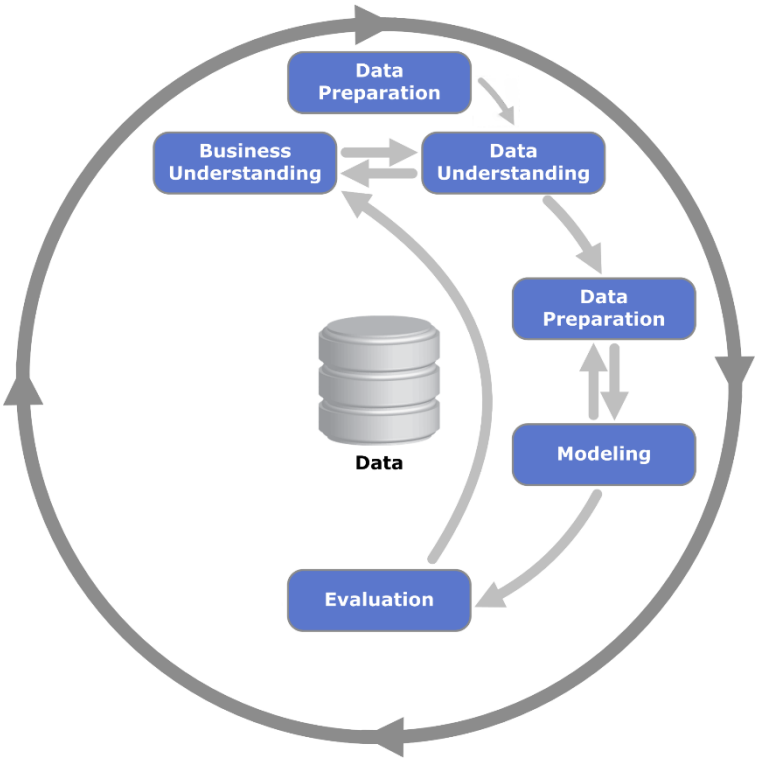

Figure 3. Edited CRISP-DM cycle that I will follow for my project.

2.2 Data

2.2.1 Source

To complete this analysis data from tweets are required. This can be achieved from using Twitter's API to search terms that will bring results of hate speech and classify them manually by reading each one to determine if it is hateful or not. However, the project is being carried

out six months into the pandemic, during the US election race and the issue has become politicised making the issue more towards political ideology rather than hate speech. Therefore, to save time and reduce the bias that I would have, I will be using data from Vidgen et al. (2020) study into detecting East-Asian prejudice on the site during the early months of the COVID-19 pandemic. They have published all their data online and it has allowed me to use their 20 thousand annotated tweet dataset for my research. They classified their tweets in five mutually exclusive categories; hostility against an East-Asian entity, criticism of an East-Asian entity, counter speech, discussion of East-Asian prejudice, and neutral. The order shown is also the hierarchy of classification, in the event a tweet falls under more than one category (Vidgen et al., 2020). They have then used two trained academics to classify all the tweets and if there is any discrepancy then it is adjudicated by a third person with more experience in the field to determine which is correct. It should be noted that since the data is a secondary source no further ethical considerations have to take place.

2.2.2 Pre-processing Data Selection

The data from Vidgen et al. (2020) is a large dataset with many additional features to it, such as the individual that is being targeted, since this project is an examination of hate speech against all East-Asian's this data is not needed for the study and is removed. Also, since the adjudicator has the final say on the classification of the tweet their opinion will be used as the only classification indicator for the data. The data had the original purpose of identifying prejudice, yet it is an appropriate dataset to use in this context due to the similarities in the research focus.

In line with the needs of the project, the tweets will be reclassified to hate or neutral speech using the Fortuna and Nunes (2018) definition of hate speech. The results of this data reclassification are shown in table 2.

Hate	Neutral
Hostility against an East-Asian entity	Counter speech
Criticism of an East-Asian entity	Discussion of East-Asian prejudice
	Neutral

Table 2. This shows how the tweets will be reclassified.

Vidgen et al.'s (2020) classification model ensures that the data is mutually exclusive, therefore the discussion of prejudice cannot be in the hate speech category. Also, because hate speech can include diminishing language and the hierarchy that has been used in the previous study, criticism can be classified as hate speech. This will leave the data frame with 20,000 tweets and the binary classification of hate or neutral speech.

However, the first aim of this study is to replicate a trending topic, so the volume of data must be limited. Lotan (2015) argues that for something to trend then there must be a large spike in traffic on the subject and it is not necessarily a result of the overall volume of tweets. To replicate this a stratified sample of 5,000 tweets from the 20,000 will be used to carry out this analysis, the remaining data can then be used as unseen data to test the models for overfitting.

2.2.3 Pre-processing Data Quality

Text is an unstructured form of data, and results in significant amounts of pre-processing to make it structured. This section will discuss several of the pre-processing steps that are carried out on the corpus (all text) before a document (text for a single tweet) can be used in a machine learning model.

General formatting

There are four major formatting issues when it comes to analysing a corpus:

1. **Letter-case:** each character will have its own code so a computer can understand it. This results in uppercase and lowercase letters being interpreted two separate entities; a computer will think the words of the same spelling but in different cases will be different words. This is problematic for building a machine learning model as the data will be duplicated. To correct for this the entire corpus will be put into one case, following the traditional method of selecting lowercase text.
2. **Punctuation:** a computer will interpret a word with punctuation, such as a comma, in a different way to unpunctuated words. Thus, it is necessary to remove all punctuation from the corpus.
3. **Stop-words:** these are words that are present in a sentence that add little meaning, such as 'the'. These words are used in abundance when writing, causing the data to be skewed towards these high-volume words that have little relevance. This would affect the efficacy of a machine learning model and therefore they must be removed from the corpus.

4. **Suffixes:** the suffix of a word will interfere with a computer's understanding of a document. Including words in different tenses will result in duplication of data. Therefore, the words will be stemmed to their root to prevent the issue.

Twitter Related Issues

This project is using data from Twitter posts. One of the main issues that victims of hate speech face, is that they can be targeted from bot accounts (Albadi et al., 2019). These accounts can post the same tweet several times, but change elements to maximise the post's reach, such as the tweet's recipient, the attached link, and the associated hashtag (Pew Research Center, 2018). This is a problem for the analysis because it can skew the data to over represent the views of these accounts. To combat this, all Twitter usernames, attached links, associated hashtags, and emoticons are removed from the corpus. This allows any duplicate documents to be removed with ease, limiting the influence of bot accounts on the modelling.

2.2.4 Structuring Data

After these steps have been completed, the data can be made into a structured form. To do this, the corpus will be tokenised to allow each word or n-gram to be a row in the dataset (Silge and Robinson, 2020). Structuring the data in this way allows summary counts of the most common words/n-grams. Moreover, the data can be visualised allowing the classification to be checked against the word/n-grams that are associated with each class.

After this a document term matrix (DTM) will be formed. This is a two-dimensional matrix where the columns represent the words/n-grams and the rows represent the documents (Ananderajan et al., 2018/9). Each element in the array will then be weighted based on the frequency of a term appearing in the corpus. A method known as term frequency inverse document frequency (TFIDF) will be used to calculate this weighting. This is how often

a term will appear in a single document multiplied by the inverse of how common that term appears over all documents. If a term has a high frequency over relatively few documents, then it will have a large weighting because it has significance in these documents. Whereas, a term that features regularly over the entire corpus will have a small weighting as it does not aid the contextualisation of the document (Ananderajan et al., 2018/9).

To lower the dimensionality of the matrix sparse terms will be removed, these are the terms that appear in less than 1% of all documents and will add no additional insight. Once these steps have been completed, the DTM can be joined into a data frame with the corresponding classification, allowing machine learning models to be developed.

2.3 Technical solutions

Undergoing this analysis I will discuss three machine learning methods, Support Vector Machines (SVM), Random forests, and Gradient Boosting Machines (GBM). I will also discuss resampling methods that I will use to optimize the models predictive powers and the associated statistics for evaluating the model performance.

2.3.1 k-fold Cross validation

This is a resampling technique used to train and test the model from the training dataset. Using a k-fold method, the training data will be split into k-parts, where k-1 parts will be used to train the algorithm. The fold that is not being used will test this model. This process will be repeated until each fold has had the opportunity to be a test for the model (Figure 4).

Furthermore, machine learning algorithms have many parameters that can be changed manually. However, if cross validation is applied then the parameters can be tuned to develop the best model (Hastie et al., 2017; James et al., 2018).

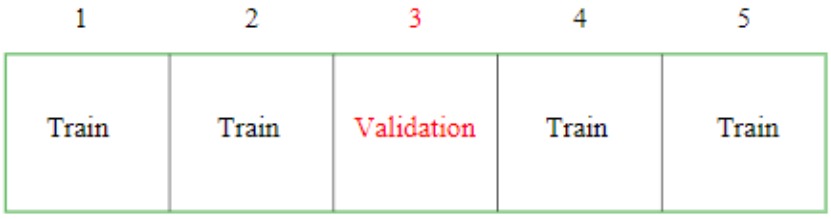

Figure 4. This is an example of how a dataset is split into k = 5 folds for cross validation. Figure reproduced from Hastie et al., (2017, p.242).

This resampling technique is widely used in industry and is appropriate for this project because it is less computationally straining than other cross-validation methods, such as leave one out cross validation. Since the second aim is about speed of production it has a good trade-off between time and accuracy.

2.3.2 Comparison Statistics

After the best models for each method are found the best method must be determined. This is done using several different statistics from the model prediction. This section will explain them.

Confusion Matrix

A confusion matrix allows access to how the model preforms when classifying the test data. For the binary classification that this project will use they are very simple to interpret (Figure 5). Performance statistics that can be calculated from a confusion matrix are: accuracy, precision, recall, and F1-score. These statistics are key to choosing the optimal model.

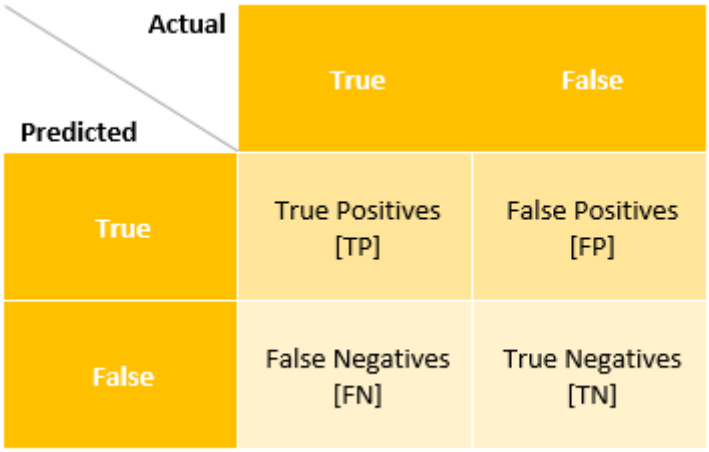

Figure 5. This is a confusion matrix for binary classification. This shows how a machine learning model compares to the original data when making predictions.

Accuracy and Cohen's Kappa

The accuracy statistic is easy to calculate (Equation 1). It is a measure of how accurate the model is, however, in imbalanced classification then high accuracy is not necessarily a sign of an accurate model. This is where Cohen's Kappa is needed (Equation 2). This statistic is a measure of whether a prediction was by chance or not, values for Cohen's kappa are explained in Table 3 (McHugh, 2012).

$$Acc = \frac{\Sigma(TP, TN)}{\mathrm{AD}}$$

Equation 1. This is how the accuracy is calculated, AD is the sum of the entire confusion matrix.

$$\kappa = \frac{(Acc - AC)}{(1 - AC)}$$

Equation 2. This is how Cohen's Kappa is calculated, where AC is the agreement by chance.

$$AC = \left(\frac{\Sigma(TP, FN)}{AD} \times \frac{\Sigma(TP, FP)}{AD}\right) + \left(\frac{\Sigma(FP, TN)}{AD} \times \frac{\Sigma(FN, TN)}{AD}\right)$$

Equation 3. This is how the agreement by chance is calculated.

Band	Meaning
<0	No agreement (random guess)
0-0.2	None to slight agreement
0.21-0.39	Fair agreement
0.4-0.59	Moderate agreement
0.6-0.79	Substantial agreement
0.8-1	Almost perfect agreement (perfect model)

Table 3. This is how the Kappa statistic is interpreted for a machine learning model. Recreated from McHugh (2012).

Precision, Recall and F1-score

These statistics are an industry standard when evaluating a classification model with imbalanced classes. The precision ratio calculates how accurate a model is at correctly predicting the minority class (Equation 4); optimising precision is most appropriate when trying to minimise false positives. The recall statistic is a ratio of correct positive predictions to all possible positive predictions (Equation 5); optimising recall is most appropriate when trying to minimise false negatives. A combination of these ratios is the F1-score, it captures both precision and recall in a single value (Equation 6), and is the most common statistic used when grading models with imbalanced data (Brownlee, 2020a).

$$Precision = \frac{TP}{\Sigma(\mathrm{TP}, \mathrm{FP})}$$

Equation 4. This is how precision is calculated for binary classes.

$$Recall = \frac{TP}{\Sigma(\mathrm{TP}, \mathrm{FN})}$$

Equation 5. This is how recall is calculated for binary classes.

$$F1 = \frac{2 \times Precision \times Recall}{\mathrm{Precision} + \mathrm{Recall}}$$

Equation 6. This is how the F1-score is calculated.

ROC and AUC

The ROC is a probability curve that is used as a performance measure for classification models. It is plotted using the True-Positive rate (Recall) and False-Positive rate (Equation 7) statistics. The AUC is the area under this curve and signifies how good a model is a correctly predicting classes, the higher the AUC the better the prediction of TP and TN (James et al., 2018).

$$FPR = \frac{FP}{\Sigma(\mathrm{FP}, \mathrm{TN})}$$

Equation 7. The False-positive rate is the proportion that the model incorrectly predicted as true; this makes them false positives. This equation shows how this value is calculated.

How to compare models

From this we will use the accuracy and kappa statistics to compare models of the same method. When comparing different methods, the AUC statistic will be used. If there are any duplicated AUC values, then the F1-score will be used to select the best method and model.

2.3.3 Support Vector Machines (SVM)

SVM is a supervised learning classification method. It creates an optimal hyperplane between data points to best classify them according to the training data (James et al., 2018). However,

the hyperplane that is selected should allow for misclassification or the model that would be developed would have high variance and not be able to be used for prediction (Figure 6 a and b). By allowing misclassification it creates a support vector classifier and will give data points a more appropriate prediction despite which class might be closer to the data point (Figure 6 c). However, if the data cannot use a simple support vector classifier to separate the data (Figure 6 d), then an SVM is used. The SVM will add further dimensions to the data to determine the best hyperplane to separate the data (Yadav, 2018; James et al., 2018). It does this by using cross validation to resample the test data to determine which dimension is most appropriate for the data, creating a hyperplane that is best for predicting the data class.

This is an appropriate method for this project because in our data classification the labelling of criticism as hate speech from the raw data could bring about some misclassifications in the data. Furthermore, SVM works well with high dimension problems and a binary class imbalance (Yadav, 2018); the DTM will create huge dimensions for the dataset and hate speech is also minority class.

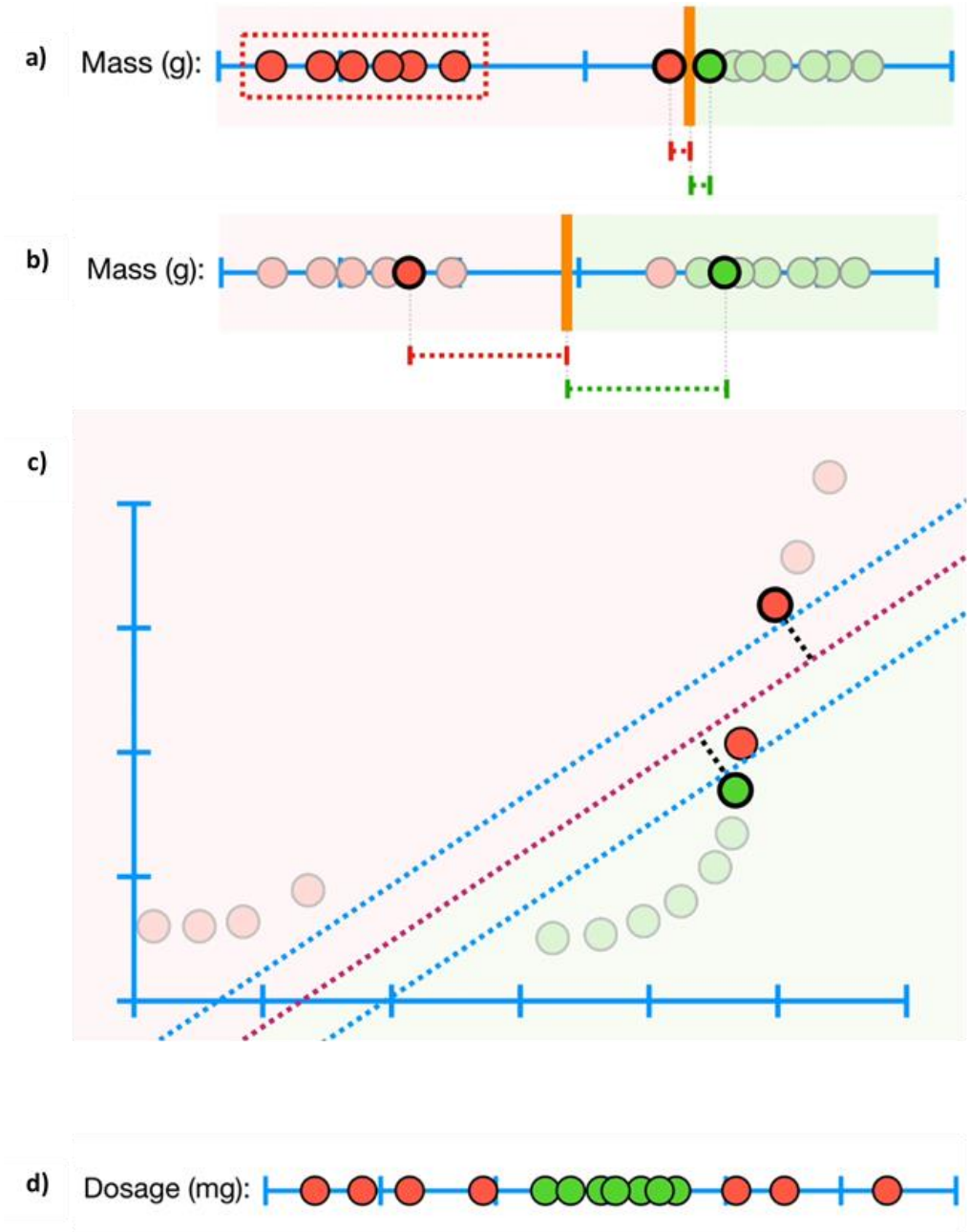

Figure 6. This is an illustration of how SVM determines which class a data point should belong to. Images reproduced from StatQuest with Josh Starmer (2019).

2.3.4 Random Forests

Random forests are an extension of decision trees, however, unlike decision trees a random forest can be used to accurately predict the outcome when given unseen data (James et al., 2018). It does this by using a bootstrapped dataset and will make splits on a random predictor variable. It will repeat this selection step until it can accurately reach a terminal node. This is

repeated several times, this is known as bootstrapped aggregation (Hastie et al., 2017; Brownlee, 2020b). To make this a random forest, the predictor variables that the algorithm can choose from is much smaller than the total predictor variables available. When developing several different trees the chosen predictor variables will change (Figure 7). This will reduce the correlation between the different trees that are built by the algorithm and thus reducing the variance when all the trees are averaged to determine the model's accuracy (Hastie et al., 2017; James et al., 2018).

This method is advantageous for the project. The way in which the data is structured will leave many input variables; random forests are known to work best with large numbers of input variables. However, this method can also suffer from taking a long time to arrive at an optimal model, thus making it harder to achieve the second aim of the project.

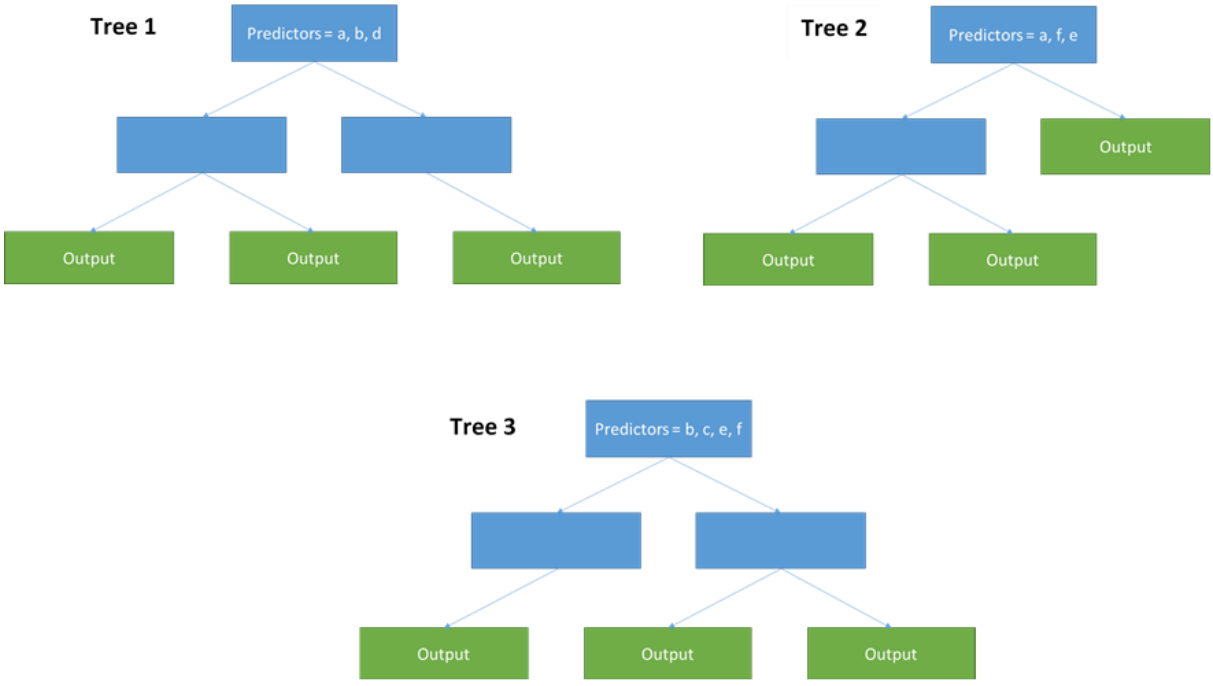

Figure 7. A visualisation of how a random forest builds trees. Note how the predictors change at the root node.

2.3.5 Gradient Boosted Machines (GBM)

This is another tree-based method that normally outperforms random forests. However, unlike random forests the trees that are produced are dependent on each other. For a classification problem, GBM will make a random guess as to what the base probability will be for each data entry where all entries belong to one class. Calculating the log odds for the classes provides an initial prediction (Equation 8) which is then converted into a probability using the logistic function (Equation 9). This value will be a number between 0 and 1 and depending on the cut-off probability it will predict that every entry belongs to the class closest to this number. If the number is 0.7 then all entries will be defaulted to 1, or if 0.3 to 0 assuming the cut-off probability is 0.5. The residuals are then calculated (Equation 10). With this information a tree is built based on the training data, this will input the residuals at the terminal node with the aim of transforming these values (Equation 11) to get a new log odd prediction (Equation 12) and thus probability from the logistic function. This process will be repeated until the prediction probability converges for each observation or until the number of trees requested is built (Hastie et al., 2017).

$$LogOdds = \log\left(\frac{yes}{no}\right)$$

Equation 8. Log odds formula for a yes/no outcome

$$logistic\ function = \frac{e^{LogOdds}}{1 + e^{LogOdds}}$$

Equation 9. Logistic function to calculate probability form the Log odds

$$psuedo\ Residual = \ observed - predicted$$

Equation 10. How the residuals are calculated for each observation in the data. Observed will be 1 or 0 depending on the class.

$$OP_{trans} = \frac{\Sigma Residuals_i}{\Sigma[PrevP_i \times (1 - \ PrevP_i)]}$$

Equation 11. Odds prediction transformation. Where P is the previous probability and i is the observation in the data.

$$LogOddsPred = \ LO_0 + (l \times OP_{trans})_1 + (l \times OP_{trans})_2 + \cdots + (l \times OP_{trans})_n$$

Equation 12. Equation for the Log Odds Prediction, LO is in the initial log odds prediction and l is the learning rate (a constant between 1 and 0). n is the number of trees that are developed.

This method is advantageous over the random forest method because it will learn from mistakes that are being made throughout model development. Gradient boosting also combines the results as it runs instead of taking an average at the end thus reducing bias. However, gradient boosting can over fit the training data leaving it inaccurate to unseen data (Ravanshad, 2018).

3. Results and Discussion

To complete the task of developing a machine learning model to categorise tweets as hate or neutral speech a brief exploratory analysis of the data is carried out. Once this has been completed it will allow a better understanding of why an algorithm will predict the classification of the tweets. This section will present the results of these analyses, and discuss the implication they have on the theory and business problem.

3.1 Exploratory Analysis

To complete this stage of the analysis the data is visualised to understand how a word may affect the prediction for the machine learning model. To begin with, we need to know how the hate speech compares to the neutral. Figure 8, the proportion of the tweets that are classified as hate is less than 30%, therefore there is a class imbalance for the data being used. This will affect the base accuracy of the machine learning models, where if it predicts all tweets as neutral speech the accuracy will be greater than 70%. Therefore, cross-validation and statistics from the confusion matrix will be used to identify this and stop it from occurring.

Examining further, the individual words and phrases of the tweets must be studied to understand how the textual data affects class prediction. This section will also discuss why some of these features are present in the data.

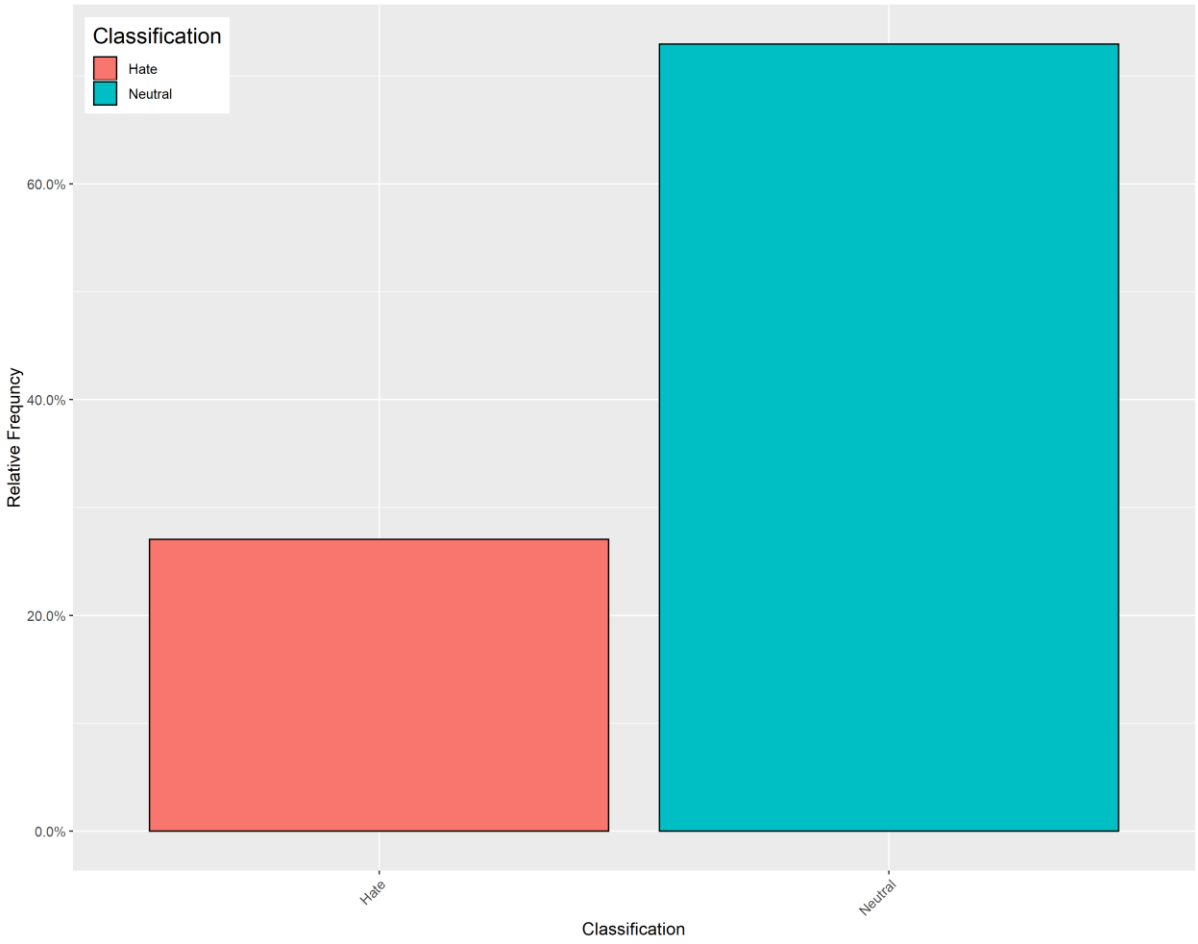

Figure 8. This shows the relative percentage frequency for tweets belonging to different classes. Hate = 27.04% and Neutral = 72.96%

3.1.1 Unigram Analysis

This examines how a single word gets classified in the dataset, Figure 9 shows the top 20 most used words in the corpus and breaks them down into the count between hate and neutral speech. This graph shows that "China" is overwhelmingly the most common word in the corpus, with most occurrences in tweets that are classified as hate. This gives a glimpse into what the tweets show, indicating that somehow China is to blame for the pandemic. Additionally, this also shows how the yellow peril is again featuring in this pandemic; China has been the first country to report the virus therefore they are to blame. This is an interesting feature of the data because the word "China" has nothing inherently linked to it that would

automatically assume that the context of the speech is hateful. Therefore, to improve the understanding of why "China" is the word most associated with hate speech in the data a similar analysis of bigram and trigram phrases should happen.

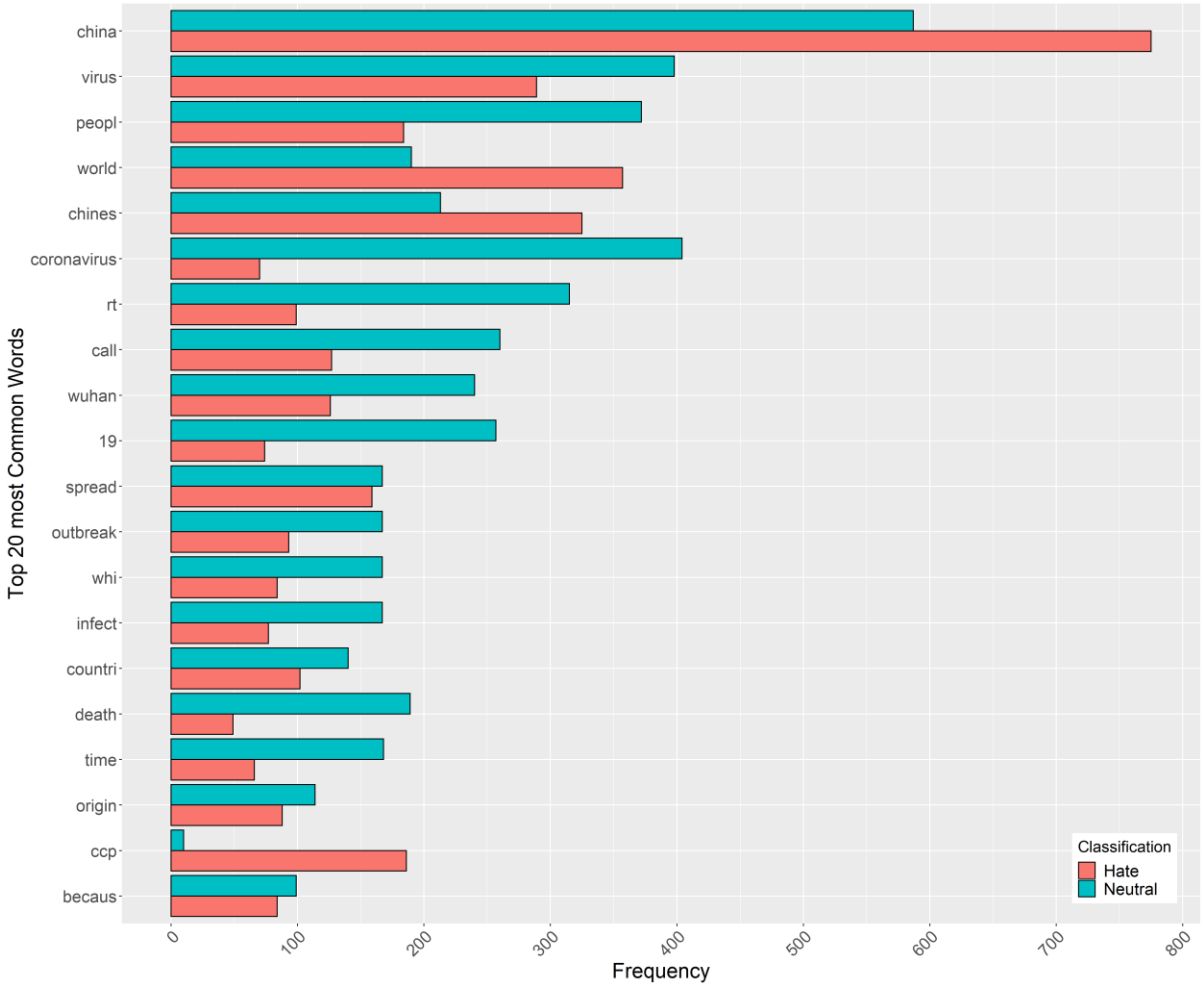

Figure 9. This shows the breakdown of unigrams between hate and neutral speech.

3.1.2 Bigram Analysis

Now that we know China is the most commonly used word, it is important to understand the context behind why this might be the case. Figure 10 shows the top 20 most common bigrams used in the corpus. The use of "China" is not as prevalent as it was in the unigrams, this could be down to China being used in many ways, thus why it has large frequency's in both hate and neutral speech. However, "Chinese" appears several times with "people",

"virus", "government", and "communist" accompanying the word in hate speech. From understanding the literature about the yellow peril this could be another example of how this ideology has reincarnated to blame the Chinese for the pandemic, thereby labelling people of Chinese origin a threat to life. Interestingly, the term "COVID 19" is the most commonly used bigram for neutral speech. This would show that people who follow the WHO nomenclature are less likely to be offensive than those that use a term like "Chinese Virus", where the evidence is less conclusive. The frequency of bigrams is much smaller than what has been seen in the unigrams, when trying to model for these terms it could severally reduce the accuracy of the predictions.

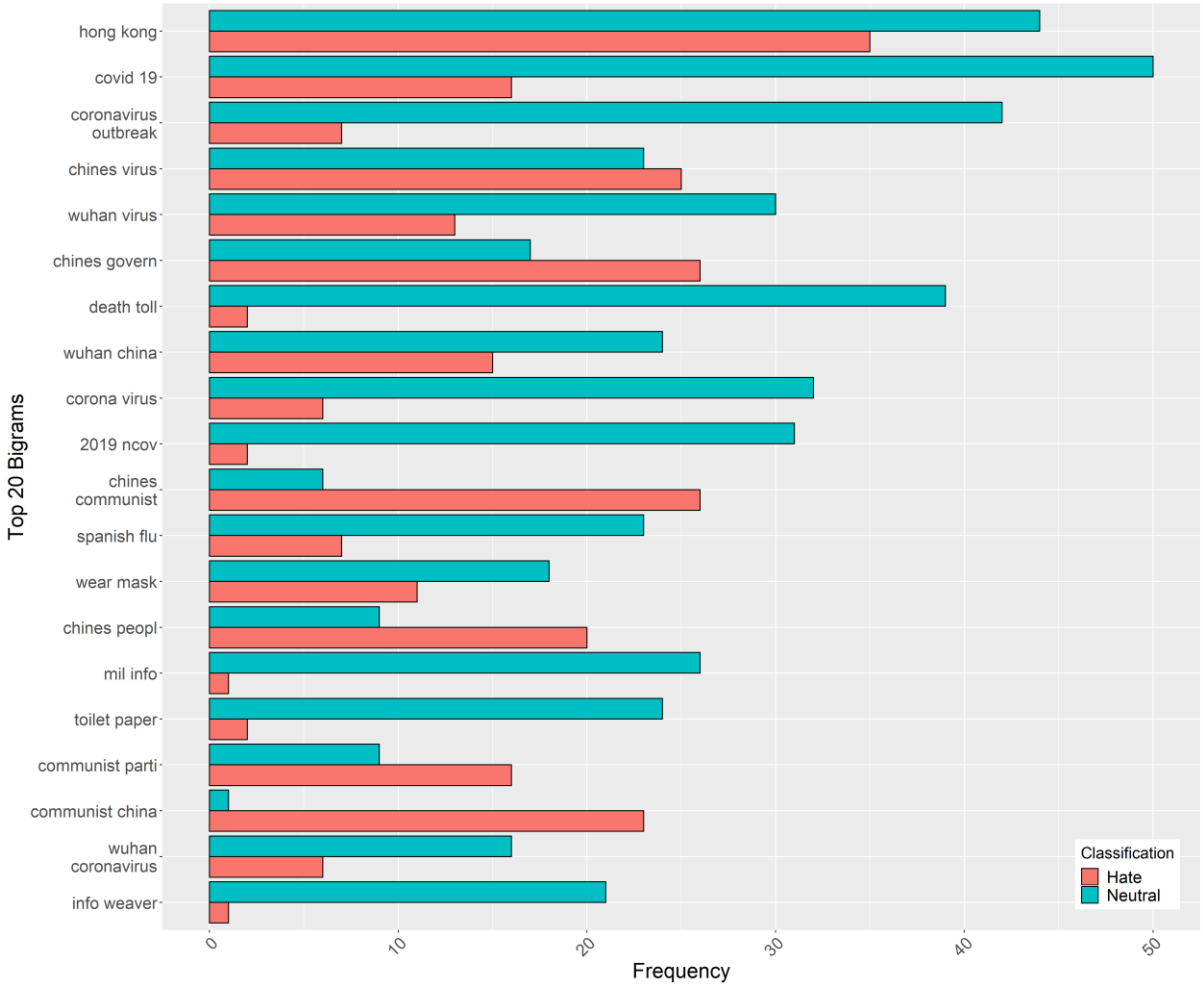

Figure 10. This shows the breakdown of bigrams between hate and neutral speech.

3.1.3 Trigram Analysis

To get further context out of the data the trigrams should be examined. Figure 11 shows that these phrases are dominated by neutral speech. This represents a cohesive way to phrase the pandemic for news headlines, “post coronavirus update” and “morning post coronavirus” have the same frequency implying that the two data points are directly related to one another. Additionally, “live break news” supports the claim that these are commonly used phrases for news updates.

The term “spread fake news” is the second most common trigram for hate speech. As a historic trope of fascism, this term implies that legitimate news organisations cannot be trusted as they are acting in the interests of the enemy. This is a notable phrase to discover within this analysis, given that the COVID-19 pandemic has triggered a significant rise in levels of distrust towards the media (Depoux et al., 2020). As a result, there has been rise in the popularity of conspiracy theories regarding the origin of the pandemic, with most popular ideas centring around it being human engineered or the introduction of 5G technologies (McLaughlin, 2020). However, this data was collected at the start of the pandemic when these conspiracy theories were in their infancy, thus explaining the small incidence that fake news has been mentioned in the data. Again, we see that the frequency in which these terms occur are very small in comparison to unigrams and bigrams. Consequently, it would not be possible to build a comprehensive machine learning model from this type of data.

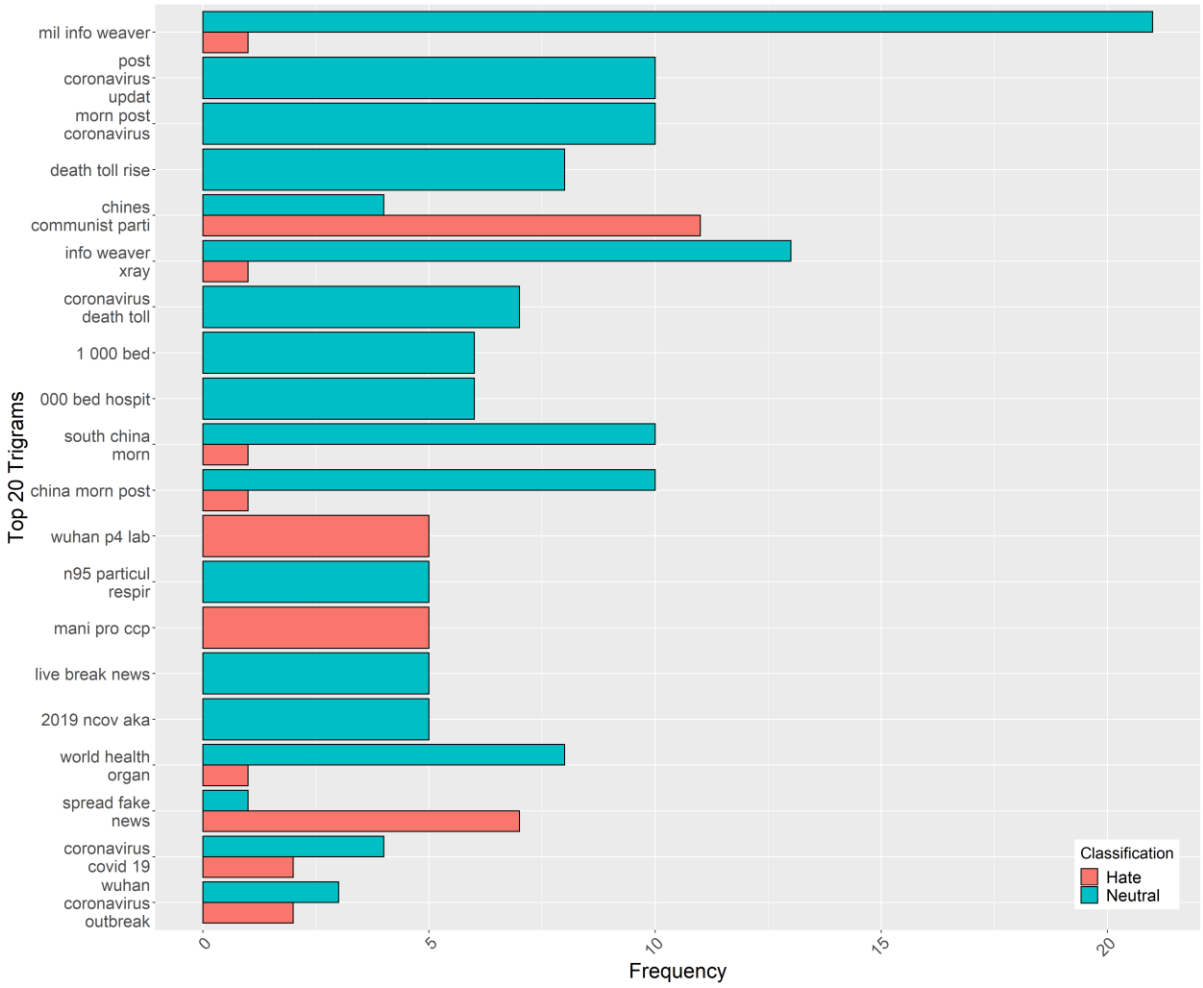

Figure 11. This shows the breakdown of trigrams between hate and neutral speech.

3.2 Machine Learning

Using machine learning methods allows a model to be built and make predictions from the data inputted. The methods mention the methodology section are used to build models for both unigrams and bigrams. A DTM is built for both unigrams and bigrams, and terms that appear in less than 1% of the entire corpus have been removed. The Caret package has been used throughout the model development with Table 4 displaying the algorithm method used in each machine learning approach. The Caret package is a wrapper package that allows easy and consistent implementation for different machine learning approaches in R, thus why it has been used.

Machine Learning Approach	Caret Method
SVM	SVMLinear2
Random Forest	ranger
Gradient Boosted Machine	gbm
Extreme Gradient Boost	xgBoost

Table 4. The Caret methods used for each Machine learning approach.

Continuing from the methodology, the optimal model for each approach is assessed on the Accuracy and Cohen's Kappa coefficient. Once the optimal model for each approach is found then the methods are compared using ROC-AUC, and the F1-score. The highest AUC or F1-score is then selected as the optimal model.

3.2.1 Unigram Modelling

SVM

The SVM method was tuned using a 10-fold cross validation method that has been explained in Chapter 2. The only parameter that needed to be optimised with this method is the cost parameter, the optimal model had a cost value of 0.5. To make sure that this model could be used for predicting outcomes a testing validation dataset was withheld from being used in building the model, this allows the accuracy statistics (Table 5) and confusion matrix statistics (Table 6 and 7) to be obtained, and with the confusion matrix statistics the ROC curve can be plotted (Figure 12).

These statistics show that there has been a slight improvement from the base accuracy of 72.96% to 78.55%, however these results need to be compared to other methods before it is possible to choose the optimal model.

Training Data Accuracy	80.65%
Training Data Kappa	0.45
Testing Data Accuracy	78.55%
Testing Data Kappa	0.41

Table 5. Optimal SVM model Accuracy and Kappa scores for training and test validation.

Actual / Predicted	Hate	Neutral
Hate	155	74
Neutral	193	823

Table 6. The confusion Matrix on the test data for the optimal SVM model.

Precision	0.677
Recall	0.445
F1-score	0.537

Table 7. The precision, recall, and F1-score for the optimal SVM model.

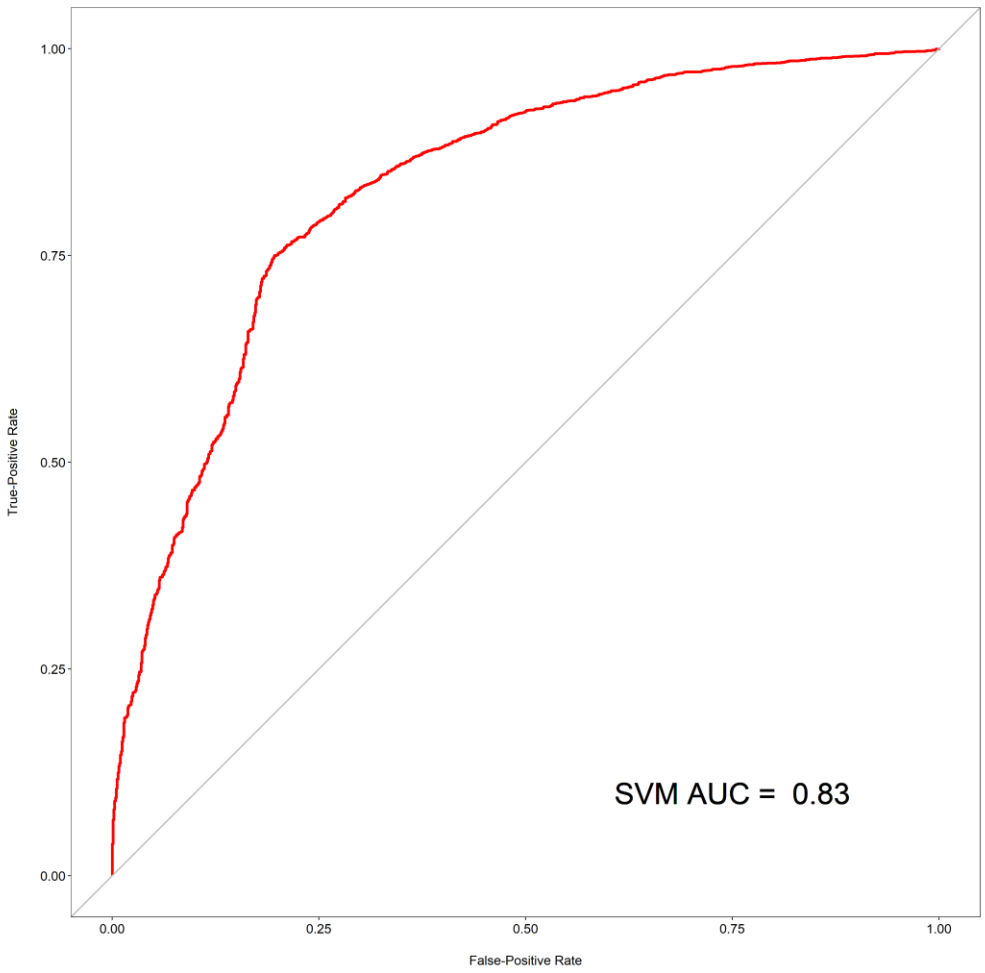

Figure 12. The ROC curve for the optimal SVM model.

Random Forests

Again the random forest model is optimised with the cross validation approach mentioned previously. This method there are four parameters that need to be optimised; the number of trees, the split rule, the number of random variables that can be considered for each node, and the minimum number of observations in the terminal node; the optimal model developed had values of 250, "Gini", 5, and 7 respectively. Again, the test validation data has been withheld to get the accuracy statistics (Table 8), the confusion matrix statistics (Table 9 and 10), and the ROC curve can be plotted (Figure 13).

This model demonstrates a higher accuracy than the base again and by using the ROC curve we can see that it outperforms the SVM model. This is currently the most optimal model for our data.

Training Data Accuracy	83.15%
Training Data Kappa	0.53
Testing Data Accuracy	82.73%
Testing Data Kappa	0.53

Table 8. Optimal Random Forest model Accuracy and Kappa scores for training and test validation

Actual / **Predicted**	**Hate**	**Neutral**
Hate	190	57
Neutral	158	840

Table 9. The confusion Matrix on the test data for the optimal random forest model

Precision	0.769
Recall	0.546
F1-score	0.639

Table 10. The precision, recall, and F1-score for the optimal random forest model

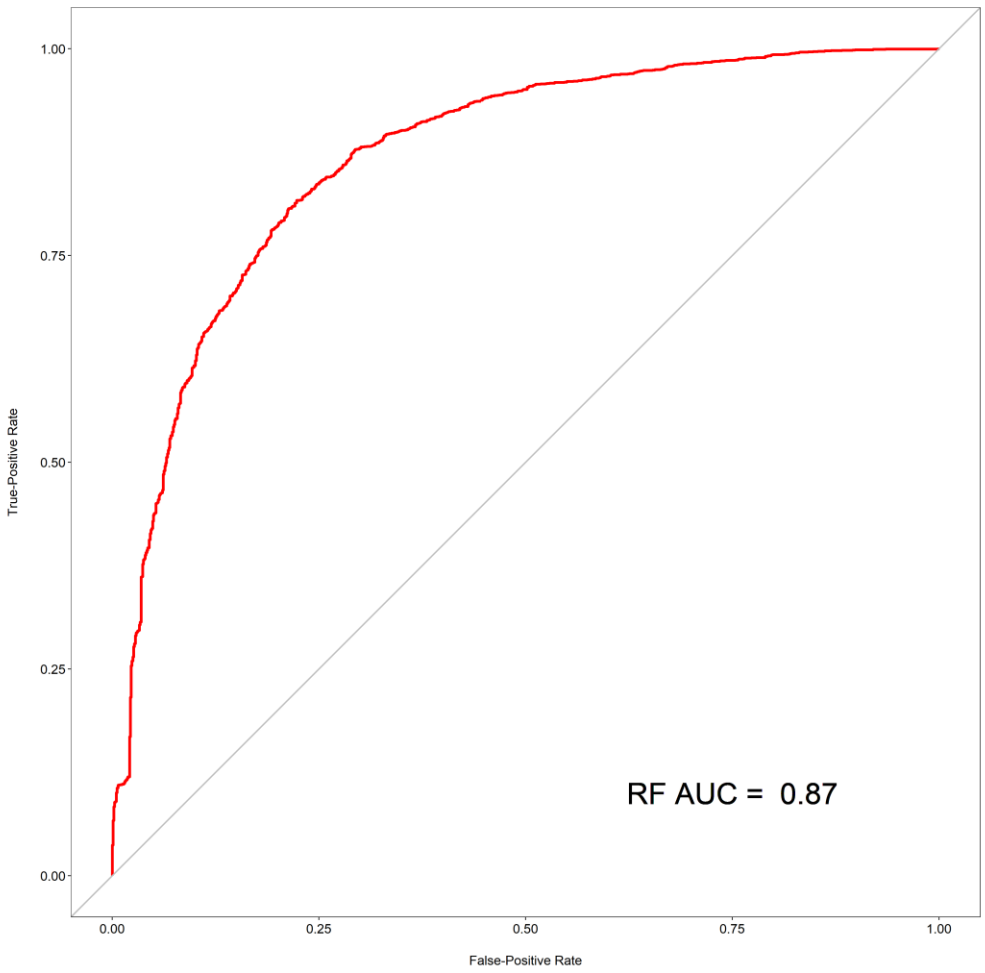

Figure 13. The ROC curve for the optimal random forest model

Gradient Boosted Machines

The GBM also has four parameters that need to be optimised through cross validation. These parameters are; number of trees, minimum observations at terminal node, the interaction depth, and the learning rate; the optimal model had values of 150, 10, 3, and 0.1 respectively. Using the same test data for validation the accuracy statistics (Table 11), confusion matrix statistics (Table 12 and 13), and the ROC curve can be obtained (Figure 14).

This model has over a 10% increase in the accuracy statistic when predicting outcomes compared to the base accuracy. Furthermore, it also out preforms the SVM model when

comparing ROC curves, however, it matches the ROC of the random forest. Examining the F1-score of both random forest and GBM, the GBM model is the most appropriate model so far.

Training Data Accuracy	83.37%
Training Data Kappa	0.54
Testing Data Accuracy	83.05%
Testing Data Kappa	0.54

Table 11. Optimal GBM model accuracy and kappa scores for training and test validation.

Actual / Predicted	Hate	Neutral
Hate	195	58
Neutral	153	839

Table 12. The confusion Matrix on the test data for the optimal GBM model.

Precision	0.771
Recall	0.560
F1-score	0.649

Table 13.The precision, recall, and F1-score for the optimal GBM model.

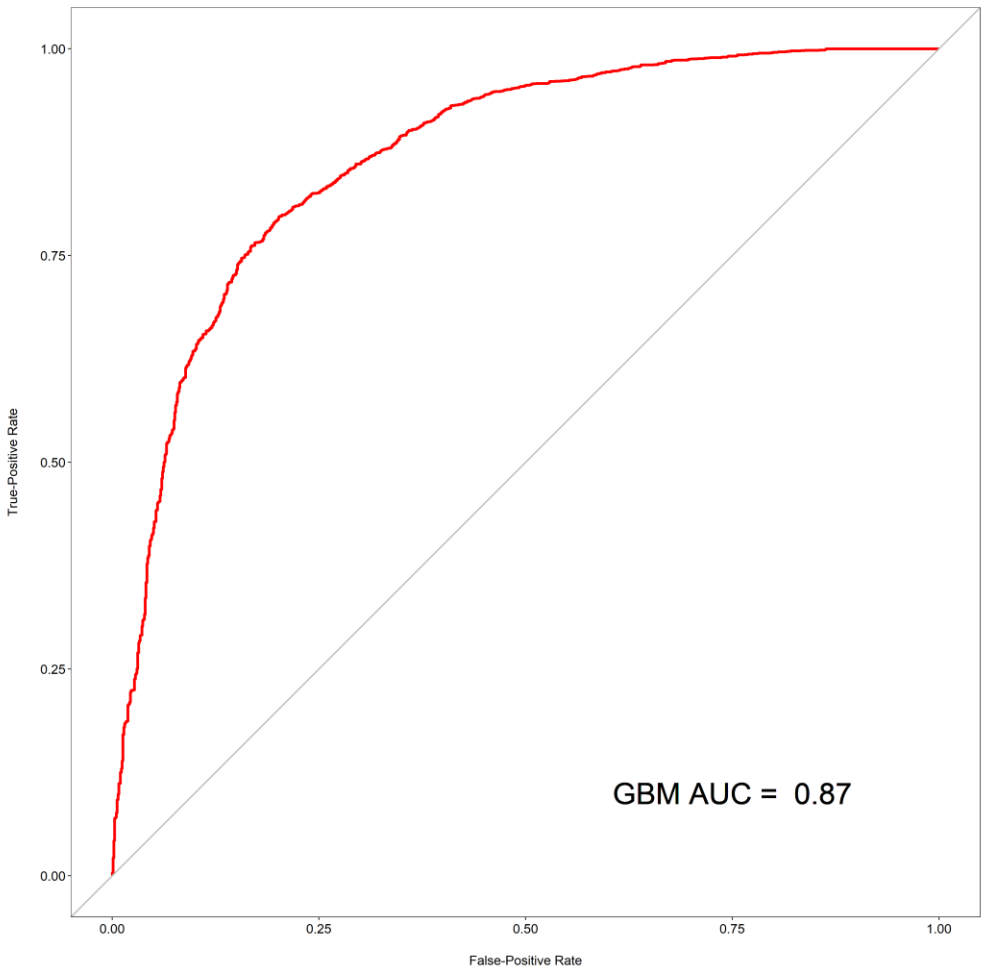

Figure 14. This is the ROC curve for the optimal GBM model.

Extreme Gradient Boost (XGB)

Since the GBM model outperformed both the SVM and random forest models, it is important to see if more complex methods are used will it result in better performance. Therefore, an extreme gradient boost method is being used to see if will out preform previous models with this data. There are three parameters that are optimised through cross validation; number of trees, learning rate, and the proportion of variables considered for each tree; this give the optimal values of 150, 0.3 and 0.6. Again using the same test data to validate the effectiveness of the model, accuracy statistics (Table 14), confusion matrix statistics (Table 15 and 16), and ROC curve can be obtained (Figure 15).

Once again this model out preforms the baseline accuracy, and matches both random forest and GBM in ROC curve performance. Examining the F1-score and the XGB method preforms worse than both the random forest and the GBM models.

Training Data Accuracy	83.58%
Training Data Kappa	0.54
Testing Data Accuracy	82.41%
Testing Data Kappa	0.53

Table 14. Optimal XGB model Accuracy and Kappa scores for training and test validation.

Actual / Predicted	Hate	Neutral
Hate	193	64
Neutral	155	833

Table 15. The confusion Matrix on the test data for the optimal XGB model.

Precision	0.751
Recall	0.555
F1-score	0.638

Table 16.The precision, recall, and F1-score for the optimal XGB model.

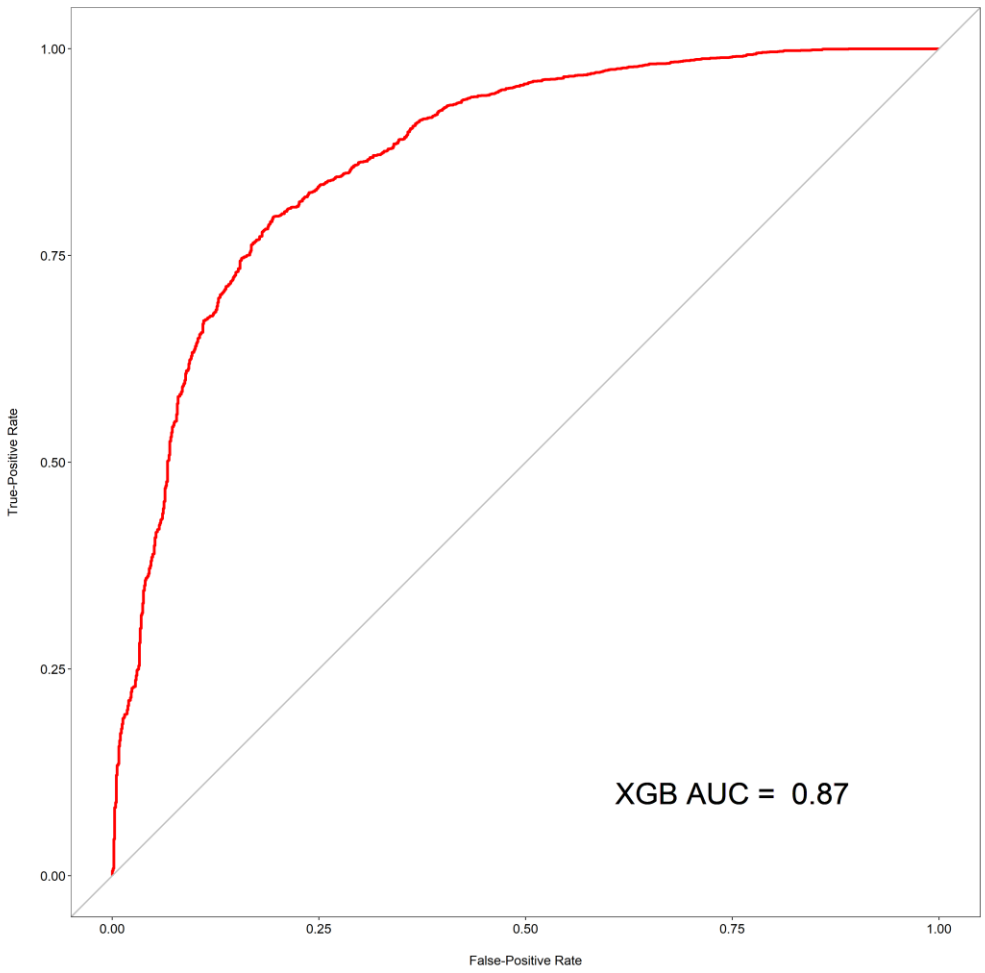

Figure 15. The ROC curve for the optimal XCB model.

Testing the Models with a large amount of unseen data

In the methodology section it was mentioned that the data would be sampled from a dataset containing approximately 20 thousand entries. This leaves a large chunk of the original data untested on the models. Table 17, shows the accuracy and confusion matrix statistics associated with each model when the unseen data is used. From this we have nearly identical statistics for all of the models, this is evidence that the models do not over-fit the training dataset and can be used for future predictions with confidence.

	SVM	RF	GBM	XGB
Accuracy	80%	83%	84%	83%
Kappa	0.432	0.523	0.547	0.534
Precision	0.704	0.774	0.781	0.771
Recall	0.456	0.528	0.555	0.545
F1-score	0.553	0.628	0.649	0.638

Table 17. This shows the summary statistics when the remaining data is used as a test for each model.

3.2.2 Bigram Modelling

The bigram models use two-word phrases to build the TDM. However, this matrix has lower dimensions than that of the unigram data when the sparse terms are removed. This reduces the predictive power from this data compared to the unigram analysis. This section will compare the unigram and bigram models to each other using ROC-AUC. These results show that no bigram model should be chosen, and because of the poor performance from these models, a trigram solution will not improve predictability (Figure 16,17,18, and 19).

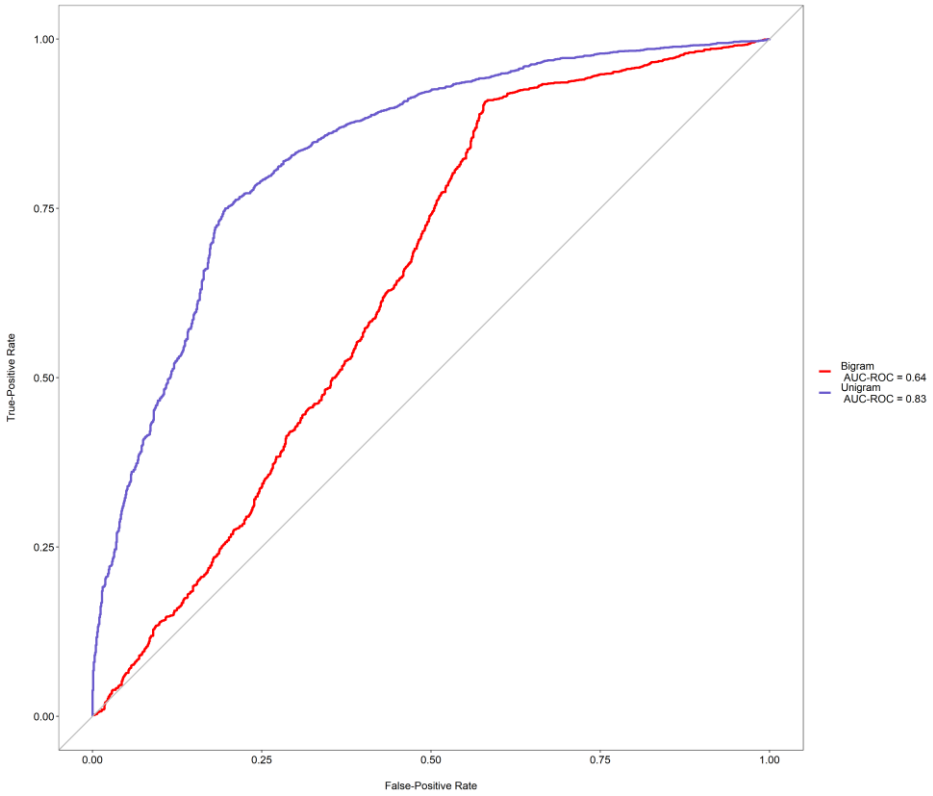

Figure 16. Comparison between SVM models.

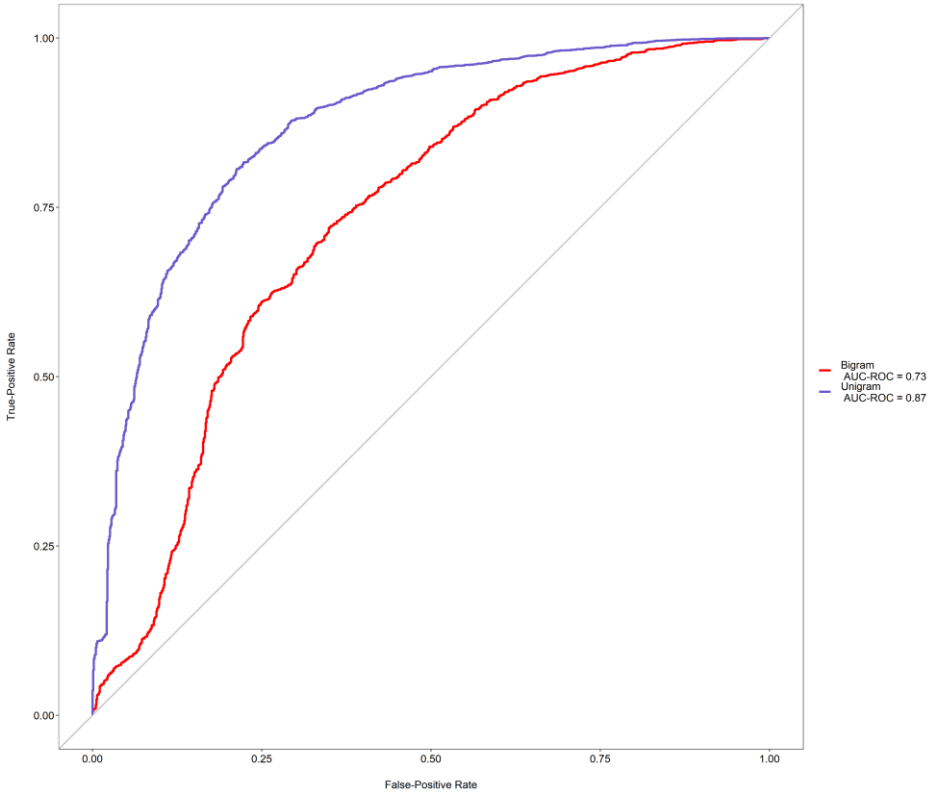

Figure 17. Comparison between Random Forest models.

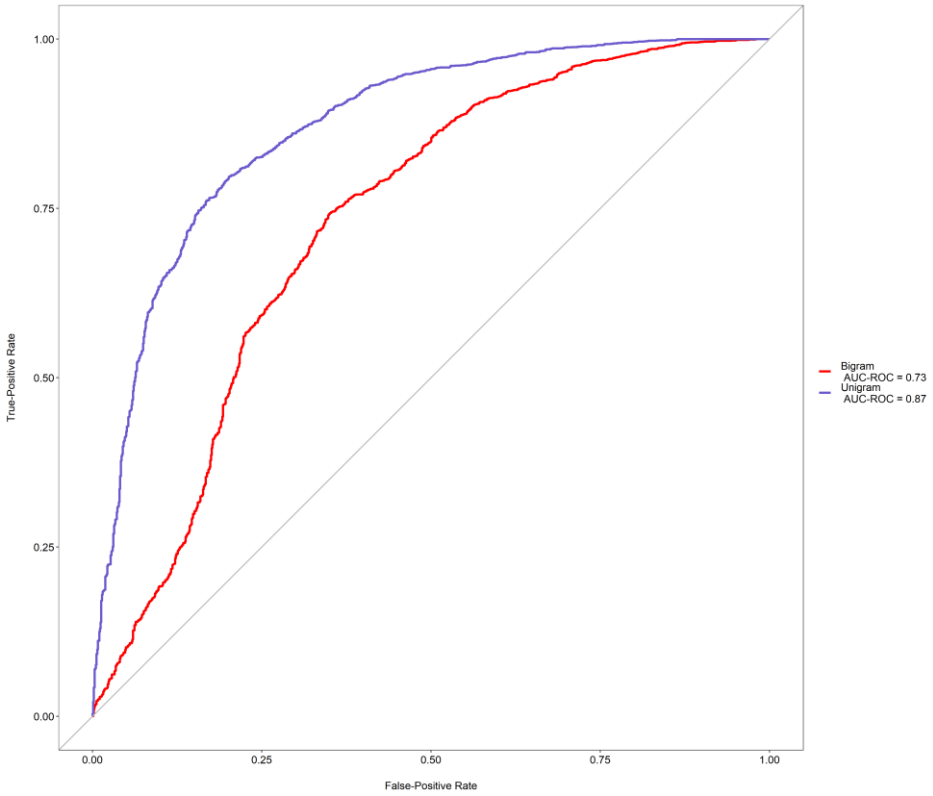

Figure 18. Comparison between GBM models.

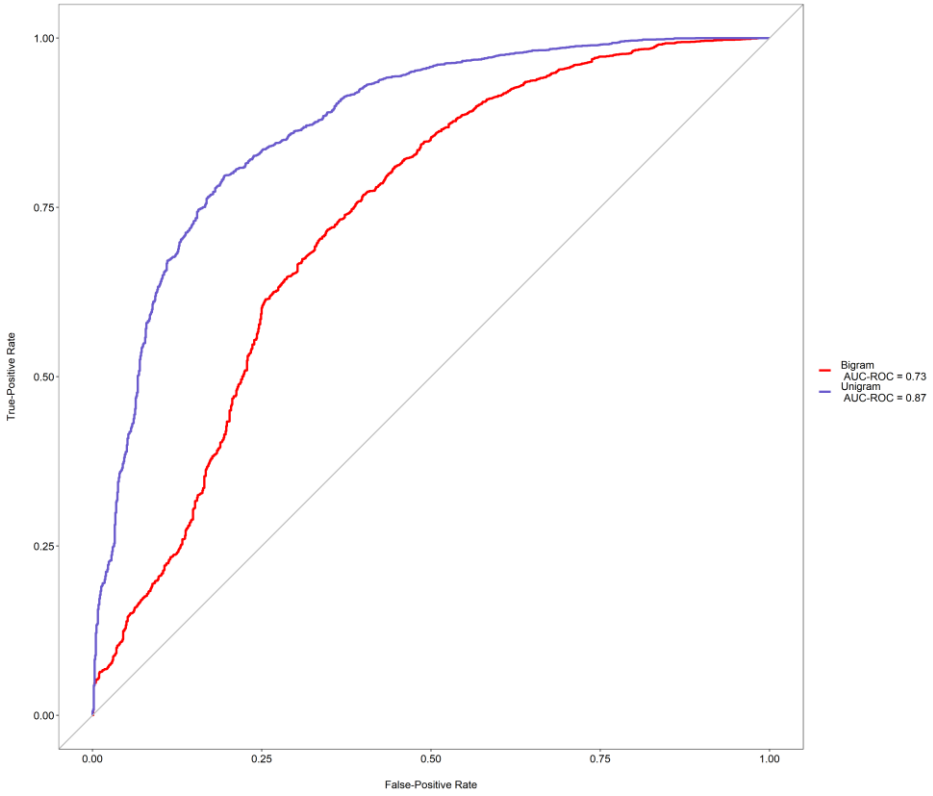

Figure 19. Comparison between XGB models.

3.3 Discussion of Modelling

The main aim of this project was to use machine learning techniques to detect hate speech in trending topics on social media. By using a trending topic the speed that a model can be implemented is key to a successful framework being developed. This section will discuss the implications on the theory for hate speech detection, how this study could affect social media companies, and I will present a way of implementing this method into a platform.

The results section proposed four different methods for achieving this goal: SVM, random forest (RF), gradient boosted machine (GBM), and extreme gradient boost (XGB). Of the four methods tested for the data used, GBM out preformed the other methods as evidenced by the statistics used to rate the models performance, scoring highest in each category, thus making it the optimal model that should be used for implementing on the data (Table 18). Further to this, whilst GBM is not the quickest to develop it has significantly better performance statistics than that of SVM therefore it should be preferred to this method.

	SVM	RF	GBM	XGB
Tuning Times* (approximate)	30mins	140mins	90mins	105mins
Accuracy	79%	83%	83%	82%
Kappa	0.405	0.529	0.541	0.525
ROC-AUC	0.830	0.870	0.870	0.870
Precision	0.677	0.769	0.771	0.751

Recall	0.445	0.546	0.560	0.555
F1-score	0.537	0.639	0.649	0.638

*Table 18. This is a table showing all the statistics associated with the models performance including time to tune. Statistics are from the validation dataset. *These times are dependent on the computer that was used to develop the models.*

3.3.1 Implication on Theory

This project has built upon the current literature by taking the idea of a trending topic and applying machine learning methods to automatically detect hate speech. The success of this project has resulted in a novel approach to rapidly detect hate speech in emerging trends that can be implemented when researching trends on social media. This can be used to gauge a reaction to current events where the subject is highly emotive and results in a trigger moment for hate speech against a single group. Building a robust model for a future trend would enable live tracking of hate online and allows for further understanding as to why it is occurring.

Due to the novel approach of this project comparisons to previous literature are difficult to make. Previous studies have primarily focused on detecting many types of hate speech in the one study and due to the large umbrella term that hate speech is, it will encompass everything from racism to sexism and beyond. However, in a normal survey of data on social media these topics would be in a large minority if there is no trigger moment (Burnap and Williams, 2015), for example Davidson et al. ’s (2017) hate speech class was 5% of the total data making a large imbalance between hate speech and other classifications. This results in a large default accuracy of 95% and when the F1-score for hate speech is calculated in this study the value is 0.531 (Recall: 0.61; Precision: 0.44) a result that is comparable to the SVM model for this project.

Likewise, Burnap and Williams (2015) use the idea of a trigger moment to examine trending hate on social media. Their hate speech class accounts for 11.7% of the data, where hate includes race and religion. In comparison, this project solely focuses on race with hate speech accounting for 27.04% of the data. As such, this lower ratio is evidence that social media interactions with current events have grown over the past decade, as they gathered their data in 2013. Moreover, whilst this study uses a limited dataset like ours, its focus is in policy and decision making while this study focuses on building a robust model to improve detection in trending topics. This explains the difference in results; a quick implementation will result in smaller F1-scores because the situation is to outperform a human operator in terms of speed, whilst decision and policy making models need very high F1-scores, 0.95 as they are used as evidence that can affect people's lives for many years into the future. Therefore, this novel approach opens a whole new area of hate speech detection for trending topics, where the focus is a trade-off between speeds of implementation against overall accuracy of the model.

3.3.2 Implications for Business

The project shows an effective means of detecting hate speech on an emerging trend. The gradient boosted model gives a solution to social media companies that are accused of being too slow to act against hate speech on their platforms. Table 18 displays the approximate tuning times of the models, it shows that it does not take a very long time to derive an optimal model, therefore this approach would allow social media companies to respond quicker to trigger moments.

If social media companies continue to ignore their response time to hate speech, then they risk their reputation. This could result in the company suffering financially, ensuing in

them losing out to competitors that appear to take the issue more seriously. This could happen in two ways. The first is from advertisers; Mark Zuckerberg famously testified to congress that "we run ads" when asked how Facebook make money (NBC News, 2018). However, advertisers have reacted to Facebooks inaction over hate speech on their platform by pulling their adverts in the wake of the Black Lives Matter movement (Dwoskin and Telford, 2020). Paying for advertisements on a slow to react platform is a way of condoning these actions or not caring. Therefore, advertisers will see continued advertising as too risky; fearing a public backlash. This results in them pulling their adverts and the associated investment too.

The second way to lose revenue is by losing users. Again the success of social media is dependent on people engaging with the platform. If users deem engaging with the platform to be too toxic then they are likely to move to a different form of social media. This has most notably been seen on Twitter, where celebrities were the victim to constant trolling from anonymous accounts forcing them to leave the platform (Lapowsky, 2015). Whilst one deleted account might not seem like a large loss, it is the following of the person that can have a great impact on the platform, with many of them choosing to follow them to a competitor's platform. This inaction resulted in Twitter being forced to change their terms of use to prevent future loss.

3.3.3 Implementing on a Platform

The research on the implementation phase for social media companies is very sparse with Ullmann and Tomalin (2019) being one of the few researchers to discuss the issue. Trending topics have a very short timeframe of being highlighted on social media, however COVID-19 is an exception to this rule as the situation is constantly evolving and it is globally relevant. It is this evolution of a topic that can result in changing language and thus new hate speech

terms being identified. For a social media company to competently police this issue then they would need to always be reassessing what is hate speech and what is not. This would allow the algorithm to evolve as quickly as language can.

However, the question of what should be done with posts that are detected as hate speech still remain unanswered. I propose that if a social media company wants to halt the spread on social media then a tweet that they have classified as hate be removed from indirect receivership. This would essentially mean that if trending topics were examined by a person who does not follow the account then their hate message would not appear in the trending section for a given topic. For example, if in the '#COVID_19' trend for Figure 20 there was a post that was openly hateful then this would be removed from the view of indirect recipients and left solely for people who follow the account. By implementing this type of strategy, it could also reduce the effect of hateful bot accounts that are on the platform as their message would not be seen, thus rendering them ineffective and a waste of money to develop.

Figure 20. Screenshot of the trending section for Twitter Tuesday 22nd September 2020.

4. Conclusion

The aim of this project was to establish a quick response to hate speech in emerging trends on social media through using machine learning methods. The success of the project has resulted in a robust gradient boosted machine to detect hate speech against people of East-Asian origin at the start of the COVID-19 pandemic on Twitter. This has therefore allowed a framework to be developed for future emerging trends with high levels of hate speech that can be implemented by social media companies to address the issue of hate speech within a timely manner to prevent loss of revenue. This section will cite limitations that this project faced, detail the framework for social media companies, and will suggest future areas of research.

4.1 Limitations of the Research

This research has provided a novel approach to rapidly detect hate speech on social media platforms. However, the project has a number of limitations. To begin with the data being used is for detecting prejudice on social media and not hate speech. Whilst I have detailed how the data has been re-categorised it could be argued that the definition of hate speech has been interpreted very liberally. Thus, explaining why the class imbalance is not as extreme as previous studies, such as Burnap and Williams (2015).

Along with the interpretation of hate speech the data also suffered from a self-imposed limitation on the volume of data. This was to try and replicate the nature of a trending topic, however, the data source did contain 20 thousand posts. Whilst a trending topic relies on rate of interaction with a key word, it is possible for a trending topic to contain 20 thousand posts, as evidenced by Figure 20. Adopting all the data as an acceptable number

for a trending topic could allow for a more accurate model for classifying posts for hate speech.

A further limitation of the study is resource based. The computational power that I have access to is limited to a basic computer that is not optimised for developing mathematical models through extensive memory or large processing power. The idea of the project is speed of implementation, and if it takes a basic computer 90 minutes to tune the optimal model then it would take a more powerful machine, that are likely to be available to social media companies, a much shorter time to run. Improved machine power would also allow for more variable tuning, over a larger possibilities of parameter values thus making the model more robust and better at classifying user's posts.

The final limitation limits all hate speech classification; the evolution of language. In the case of COVID-19 there are old tropes and stereotypes that were associated with people from East-Asia and thus make it very easy to identify hate speech with much older models. However, COVID-19 also ushered in a new wave of hate against this group of people and thus what were once neutral speech statements have become associated with hate speech. The data that has been used to build this model was mined at the very start of the pandemic, therefore the evolution of hateful terms that has developed during COVID-19 may not be accounted for in the data. This could result in the model already being out of date for classifying hate speech against East-Asian people.

4.2 Framework for Social Media Companies

To begin with, a trigger moment for hate speech can happen at any time. Therefore, social media companies should be vigilant and aware of current event issues. Once a trigger moment has been identified, social media companies should deploy their teams that

moderate posts to focus on the trigger issue to identify both hate and neutral speech. Due to their large resources for moderation this task should take no longer than an hour to identify a few thousand posts. At this point the data should be processed in a similar way to this project before being ready for modelling.

In the modelling section, an optimal model on whatever method being tested should be completed within a reasonable timeframe of 2 hours. Again, due to these company's large resources it will be possible to have a robust model in this time frame. The model should be optimised to whatever the company views as more important, either minimising false-positives or false-negatives. Once these step are completed and the model is tested thoroughly, the model should be deployed on trends associated to the trigger moment. If these steps are followed then it should not take a social media company any longer than six hours to react to a hate speech on their platform.

4.3 Future Work

An obvious step to further develop this type of analysis is to take a new trending topic and preform a similar approach of classification. It is hard to say what topic should be examined because this analysis relies on a trigger moment for sufficient hate speech data to be available. Further resampling of COVID-19 for hate speech could help update the model that has currently been developed, however, the pandemic has become a political football so hate speech may not be as rife as it was at the start of the pandemic, limiting further improvement.

Similar techniques could be deployed to categorise social media posts in different ways. As mentioned previously, the COVID-19 pandemic has become a political issue, therefore resampling the data could be used to understand a person's voting intention and use this modelling to predict the outcome of elections. This would be useful for social media

companies or any organisation as it could allow them to prepare with more confidence for any potential changes that a new government will bring.

5. References

Albadi, N., Kurdi, M. and Mishra, S., 2019. Hateful People or Hateful Bots? Detection and Characterization of Bots Spreading Religious Hatred in Arabic Social Media. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), pp.1-25.

Al-Makhadmeh, Z. and Tolba, A., 2020. Automatic hate speech detection using killer natural language processing optimizing ensemble deep learning approach. *Computing*, *102*(2), pp.501-522.

Anandarajan, M., Hill, C. and Nolan, T (2019) 'Term-Document Representation', in Anandarajan, M., Hill, C. and Nolan, T (ed.) *Practical Text Analysis.* Switzerland: Springer, pp. 61-73.

Brownlee, J. (2019) *A Gentle Introduction to the Bag-of-Words Model,* Available at: *https://bit.ly/340HrnR* (Accessed: 13th August 2020).

Brownlee, J. (2020a) *How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification,* Available at: *https://bit.ly/3cb6nv5* (Accessed: 3rd September 2020).

Brownlee, J. (2020b) *Bagging and Random Forest Ensemble Algorithms for Machine Learning,* Available at: *https://bit.ly/2Rz2BlU* (Accessed: 3rd September 2020).

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.

Coleman, J. (2020) *Trump again refers to coronavirus as 'kung flu',* Available at: *https://bit.ly/3kRQvBR* (Accessed: 3rd August 2020).

Davidson, T., Warmsley, D., Macy, M. and Weber, I., 2017. Automated hate speech detection and the problem of offensive language. *arXiv preprint arXiv:1703.04009*.

Depoux, A., Martin, S., Karafillakis, E., Preet, R., Wilder-Smith, A. and Larson, H., 2020. The pandemic of social media panic travels faster than the COVID-19 outbreak.

Dworkin, R. (2009) 'Foreword', in Hare, I. and Weinstein, J. (ed.) *Extreme Speech and Democracy.* New York: Oxford University Press, v-ix.

Dwoskin, E. and Telford, T. (2020) *Facebook is working to persuade advertisers to abandon their boycott. So far, they aren't impressed.,* Available at: *https://wapo.st/30V7f2O* (Accessed: 28th July 2020).

Dwoskin, E., 2019. *Philippine Workers Are On The Front Lines Of The Battle To Keep The Internet Safe, But At What Cost?*. [video] Available at: *https://wapo.st/30XoL6G* (Accessed: 28th July 2020).

Elwell, C. K., Labonte, M. and Morrison, W. M. (2007) 'Is Chian a Threat to the U.S. Economy?', in Finn, J. D. (ed.) *China-U.S. Economic and Geopolitical Relations.* New York: Nova Science Publishers, pp. 31 - 84.

Fitch, W. T. (2010) *The Evolution of Language*, New York: Cambridge University Press.

Forster, K. (2020) *Parler: Katie Hopkins and Laurence Fox flee to Twitter's anything-goes rival,* Available at: *https://bit.ly/2Y1WslB* (Accessed: 13th August 2020).

Fortuna, P. and Nunes, S., 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), pp.1-30.

Fox, J. (2014) 'Why Twitter's Mission Statement Matters', *Harvard Business Review Digital Articles*, pp. 2–4.

Gillespie, T., 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.

Gover, A.R., Harper, S.B. and Langton, L., 2020. Anti-Asian hate crime during the CoViD-19 pandemic: exploring the reproduction of inequality. *American journal of criminal justice*, *45*(4), pp.647-667.

Greevy, E. and Smeaton, A.F., 2004, July. 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* (pp. 468-469).

Guterres, A. (2020) 8 May, Available at: https://twitter.com/antonioguterres/status/1258613180030431233?s=20 (Accessed: 3rd August 2020)

Hastie, T., Tibshirani., R, and Friedman, J. (2009) *The Elements of statistical learning: Data Mining, Inference, and Prediction*, 2nd edn., New York: Springer Science+Business Media .

Heinze, E. (2016) *Hate Speech and Democratic Citizenship*, New York: Oxford University Press.

Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Howard, J., 2009. *Concentration camps on the home front: Japanese Americans in the house of Jim Crow*. University of Chicago Press.

United Nations, (1966) International Covenant on Civil and Political Rights, Available at: *https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx* (Accessed: 3rd August 2020)

James, G., Witten, D., Hastie,T. and Tibshirani, R. (2013) An Introduction to Statistical Learning with Applications in R, New York: Springer.

Kahn, S. (2020) 26 July, Available at: https://twitter.com/SadiqKhan/status/1287470218277683201 (Accessed: 28th July 2020)

Kawai, Y., 2005. Stereotyping Asian Americans: The dialectic of the model minority and the yellow peril. *The Howard Journal of Communications*, *16*(2), pp.109-130.

Lapowsky, I. (2015) *Why Twitter Is Finally Taking a Stand Against Trolls,* Available at: *https://www.wired.com/2015/04/twitter-abuse/* (Accessed: 1st October 2020).

Laub, Z. (2019) *Hate Speech on Social Media: Global Comparisons,* Available at: *https://on.cfr.org/3iJfkxA* (Accessed: 13th August 2020).

Lotan, G. (2015) *#FreddieGray — is not trending on Twitter?,* Available at: *https://bit.ly/2H7maQd* (Accessed: 3rd September 2020).

Ma, E. (2018) *3 basic approaches in Bag of Words which are better than Word Embeddings,* Available at: *https://bit.ly/2PRPtY7* (Accessed: 13th August 2020).

MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N. and Frieder, O., 2019. Hate speech detection: Challenges and solutions. *PloS one*, *14*(8), p.e0221152.

Main, T. J. (2018) *The Rise of the Alt-Right*, Washington, D.C.: Brookings Institution Press.

Marchetti, G., 1994. *Romance and the yellow peril: race, sex, and discursive strategies in Hollywood fiction*. Univ of California Press.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica*, *22*(3), pp.276-282.

McLaughlin, A. (2020) *Investigating the most convincing COVID-19 conspiracy theories,* Available at: *https://www.kcl.ac.uk/investigating-the-most-convincing-covid-19-conspiracy-theories* (Accessed: 5th October 2020).

Mehdad, Y. and Tetreault, J., 2016, September. 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).

Murgia, M. (2020) *Facebook content moderators required to sign PTSD forms,* Available at: *https://on.ft.com/2FisG5C* (Accessed: 28th July 2020).

NBC News (2018) *Senator Asks How Facebook Remains Free, Mark Zuckerberg Smirks: 'We Run Ads' | NBC News,* Available at: *https://www.youtube.com/watch?v=n2H8wx1aBiQ&ab_channel=NBCNews* (Accessed: 5th October 2020).

Rew Research Center (2018) *5 things to know about bots on Twitter,* Available at: *https://pewrsr.ch/3hBe5Qg* (Accessed: 3rd September 2020).

Pew Research Center (2019) *2. Attitudes towards China ,* Available at: *https://www.pewresearch.org/global/2019/12/05/attitudes-toward-china-2019/* (Accessed: 10th September 2020).

Pitsilis, G.K., Ramampiaro, H. and Langseth, H., 2018. Effective hate-speech detection in Twitter data using recurrent neural networks. *Applied Intelligence*, *48*(12), pp.4730-4742.

Ravanshad, A. (2018) *Gradient Boosting vs Random Forest,* Available at: *https://bit.ly/33sDRkw* (Accessed: 3rd September 2020).

Roberts, S.T., 2019. *Behind the screen: Content moderation in the shadows of social media*. Yale University Press.

Schmidt, A. and Wiegand, M., 2017, April. 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).

Schweppe, J., Haynes, A. and MacIntosh, E.M., 2020. What is measured matters: The value of third party hate crime monitoring. *European Journal on Criminal Policy and Research*, *26*(1), pp.39-59.

Silge, J. and Robinson, D. (2017) *Text Mining with R: A Tidy Approach*, Sebastopol: O'Reilly Media.

StatQuest with Josh Starmer (2019) *Support Vector Machines, Clearly Explained!!!,* Available at: *https://www.youtube.com/watch?v=efR1C6CvhmE&ab_channel=StatQuestwithJoshStarmer* (Accessed: 3rd September 2020).

Tennant, H. (1981) *Natural Language Processing* , New York: Petrocelli Books.

Thota, S.C., 2018. Social Media: A Conceptual Model of the Why's, When's and How's of Consumer Usage of Social Media and Implications on Business Strategies. *Academy of Marketing Studies Journal*, *22*(3), pp.1-12.

Trump, D. J. (2020a) 25 May. Available at: https://twitter.com/realDonaldTrump/status/1265013797334507521 (Accessed: 4th August 2020)

Twitter (2020) *Rules Enforcement,* Available at: *https://transparency.twitter.com/en/reports/rules-enforcement.html#2019-jul-dec* (Accessed: 10 September 2019).

Twitter. (n.d.) FAQ, Available at: https://bit.ly/2FsEse1 (Accessed: 13th August 2020).

Ullmann, S. and Tomalin, M., 2020. Quarantining online hate speech: technical and ethical perspectives. *Ethics and Information Technology*, pp.1-12.

Vidgen, B., Botelho, A., Broniatowski, D., Guest, E., Hall, M., Margetts, H., Tromble, R., Waseem, Z. and Hale, S., 2020. Detecting East Asian Prejudice on Social Media. *arXiv preprint arXiv:2005.03909*.

Vorhies, W. (2016) *CRISP-DM – a Standard Methodology to Ensure a Good Outcome,* Available at: *https://bit.ly/3msbpIj* (Accessed: 3rd September 2020).

Waldron, J. (2012) *The Harm in Hate Speech*, Cambridge: Harvard University Press.

Yadav, A. (2018) *Support Vector Machines (SVM),* Available at: *https://bit.ly/3hw3cPG* (Accessed: 3rd September 2020).