University of Tartu

Faculty of Science and Technology

Institute of Computer Science

Fortune Ikechukwu Festus

Discriminatory Speech on Digital Platform a case study of Twitter (Gender, Race, Politics, Sexuality)

Master's Thesis (30 ECTS)

Innovation and Technology Management

Supervisors:

Rajesh Sharma PhD

Christian Simon Ritter PhD

Tartu 2020

Abstract: Discriminatory speech on digital media platforms

In recent years communication via social media has become more personal and available for every individual or group of people irrespective of their interests. This has enable people express their thoughts, ideas and views freely. Though it brings lots of ease to communication, it also gives rise to discriminatory challenges. Online hate is a major example of such challenge.As these social media users grow, so does the impact of hate speech. Despite the magnitude and growth level of research in this field there is a hug gap in understanding the hate speech and how it affects certain aspects of human life's e.g race,gender,sexuality, politics. This has prompted researchers to apply techniques like social networks analysis to detect hate groups. But in this research we strongly believe that the content of hate matters as well.Thus in this paper we apply sentiment analysis and topic modelling to understand the discuss of hate as it affect race, gender, sexuality, politics. Our result shows that the content plays an important role in preventing and eradicating discrimination of these platforms.

CERCS: P160 Statistics, operation research, programming, actuarial

Keywords: Discriminatory speech, Sentiment Analysis, Topic Modelling, Hate speech

Diskrimineeriv kõne digitaalsetel platvormidel - Twitteri juhtumianalüüs (sugu, seksuaalsus, rass, poliitika)

Viimastel aastatel on suhtlus sotsiaalmeedia vahendusel muutunud isikupärasemaks ja kättesaadavaks igale üksikisikule või inimrühmale, sõltumata huvidest. See on võimaldanud inimestel oma mõtteid, ideid ja seisukohti vabalt väljendada. Ehkki see suhtlemiseks pakub palju lihtsust, tekitab see ka diskrimineerivaid väljakutseid. Veebipõhine vihkamine on sellise väljakutse peamine näide. Kuna nende sotsiaalmeedia kasutajate arv kasvab, kasvab ka vihakõne mõju. Hoolimata selle valdkonna uuringute ulatusest ja kasvutasemest, on vihakõne mõistmisel ja kuidas see mõjutab inimelu teatud aspekte, nt rass, sugu, seksuaalsus, poliitika. See on ajendanud teadlasi viharühmade tuvastamiseks kasutama selliseid tehnikaid nagu sotsiaalsete võrgustike analüüs. Kuid usume selles uurimuses, et ka viha sisu on oluline. Seetõttu rakendame selles artiklis sentimentaalianalüüsi ja teemamudelit, et mõista vihkamise arutelu, kuna see mõjutab rassi, sugu, seksuaalsust, poliitikat. Meie tulemus näitab, et sisu mängib olulist rolli platvormide diskrimineerimise ennetamisel ja likvideerimisel.

CERCS: P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Keywords: Diskrimineeriv kõne, tunnete analüüs, teema modelleerimine, vihakõne

Contents

1 Introduction

The continuous development of online communication media has significantly impacted the digital society. It has enabled individuals communicate at low cost, low supervision, enhanced circulation of ideas and enabled individuals express their likes and dislikes through various platforms [44]. However, in recent decades, social media platforms have gained more popularity as a means of communication in official and unofficial environments. While these systems have made communication easier it has also enabled certain groups or individuals circulate their preference of idea, which in turn could result in negative or positive behaviors. Media like face book, Instagram, twitter and blogs have become media of interest because of the rate at which offensive languages amongst groups can circulate and impact the society.

One feature that allows individuals express their preference of ideas is the "Direct message" on Twitter, the "groups" on facebook, this enables individuals form hate groups [48].These individuals focus on discriminatory narratives like racism, religions, gender, sexuality, ethnicity, immigration, and physical disability. As a result of its impact on the society, it has become important to identify these groups, those affected and the network behind the circulation of ideas.

Previous studies focus on [44] methods to identify hate speech in social media.In their study they analyzed data from twitter and whisper with an effort to understand the target groups of hate, the forms of hate speech and provide directions for prevention but fails to categorize target groups that are mostly affected.Also, studies carried out by [28]focused on identifying the producers of hate speech, determine their social background, explore their channels of hate speech and determine the specific groups of migrants mostly targeted and provide more information for stakeholders in Czech Republic to make better policies. In the study data was extracted from Facebook to help achieve the aim of the study. While these approaches and techniques are great, they do not provide a balance view about the problems on social media platforms, because they focus on a specific form of idea "hate groups".

The rapid growth of Social media platforms has also increased the rate of abusive and discriminatory content online, therefore it has become a critical issue in the society. As a result, researchers are making efforts to find ways to identify and detect hate groups or hateful content. However, the goal of this thesis is to characterize hate speech on twitter in other to understand the most topic of discourse amongst hate groups such as race, gender, sexuality, (politics). To do this we search for words in the (hate base, a database for hate words) with intensity greater than 50 percent . From this database we select hate words that affect (race, gender, sexuality and politics). [9] Our search criteria is based in critical race theory to identify slurs relating to (race, gender, sexuality and politics),these will be used to gather more data and address the problem of a small, but highly prolific number of hateful users. Using these words, we search and extract data from twitter, then we analyze them using an effective way to detect hate speech using sentiment analysis and topic modelling.

The project will provide new insights into;

1. Circulation of hate (bridging) speech, exclusionary (or inclusive) narratives, and anti-immigration (pro-immigration) discourses on digital media platforms.

2. We make available a new data set of user comments specifically related to Gender, Politics, Race, Sexuality. The data set includes hate words with intensity higher than 50 percent. And a more fine-grained application of sentiment analysis, and topic modelling to these data.

3. Prior Studies has focused on a fixed range of data set/ variables. However, the issues of language changing over time and users trying to evade keywords based approaches. In this research we, calculate the probability of each topic occurring on a data set of one year, using topic modelling and visualize the result for easy understanding and platform recommendation. To our knowledge this is the first study that applies these methods to this amount of data set and with variables such as gender, politics, sexuality, race.

4. The researcher also provides stakeholder with recommendation on better platform politics, having identified the effect of hate speech the lives of people.

This thesis is organized as follows: The first part in the introductory and background of the study. In section 2, related works/ literature about discriminatory speech, text mining, Social networks, Topic modelling and sentiment analysis will be reviewed. In section 3 we will discuss the research methodology used to carry out the research. In section 4, a set data extracted from twitter, will be analyzed to discover and detect hate speech. In section 5, we will discuss our finding results and further recommendation on strategies for inclusive platform politics.

2 Background of study

Social media platforms have been under increasing public scrutiny because of inconsistencies in how they apply their policies with respect to cultural difference, hate speech and issues relating to discrimination. The definition of hate speech is very connected with freedom of expression, group rights, as well as concepts of dignity, liberty, and equality [20].The impact of the Social media on the circulation of hate speech and practices has been a complex and on-going field of research. Therefore, detecting and regulating hate speech has become more important as a result of its strong connections to actual hate crimes [52]. Early identification of hate speech could help enable the outreach programs that attempt to prevent escalation from speech to action. In this field, researchers like [44] have analyzed the targets of online hate with an attempt to unveil a set of important patterns, used during hate interaction, and provide directions for prevention and detection approaches. In their study, they focused on analyzing data from whisper and twitter, to help them understand the various forms of online hate and categorize hate speech. [28]took a different approach to identify the producers of hate speech and determine their social background, explore the main channels of hate proliferation, determine the specific groups of migrants targeted by hate speech. In the study, the author combined random sampling and insider data collection tool as well as facebook data to help achieve the goal. While these efforts are of great importance, [28]] study focuses on immigrants in Czech Republic, the result cannot be generalized because of its specific geographical issue. They also fail to provide specific forms of hate about the problems in the popular social media, as they are focused on generic forms or issues.

Early work on hate speech specifically racism and the Social media pointed to inequality on the web, because people poses different levels of access to the internet, this results to racial inequalities [27],this study later stressed the unevenness in digital literacy and skills [25] as important factors considering digital inequality.From a discursive perspective, Social media creates an opportunity to perform racial identity [35]and a forum to replicate power relations and hierarchies [31] or strengthen racism[11].Currently social media platforms are the current mediators of most of the online sociability and creativity [49], are also tools for both pro-social and antisocial uses. For example, the movement created by indigenous activists in 2015 about sosblackaustralia – to stop the foreclosure of Aboriginal remote communities – as found on Twitter and Facebook a space for advocating for the rights of Black people in Australia. However, Twitter has been a platform where hate speech and harassment thrive [43],including racial, political, gender and sexist abuse [24].

3 Related works

In this section the author hopes to define hate speech according to [8],also we shall discuss various approaches, methods and findings that surrounds the area of hate speech on digital platforms and the need for media platforms especially Facebook and Twitter to adopt better platform policies, in order to reduce or eliminate hate speech discussion. In addition, we segment the related works based on techniques used by other researchers to approach the problem of hate speech detection, this includes; web mining and social network analysis, sentiment analysis, sentence-level subjectivity. In this research we will use sentiment analysis and topic modelling to approach this problem, this will enable us to achieve the goal of identifying most discoursed topics and recommending these topics to affected stakeholders (Twitter). Though this technique was used on twitter the result could be applied to similar platforms like Facebook.

Social media platforms allow users to communicate at nearly marginal costs. Most users leverage these platforms not only to interact with each other, but also share opinions and ideas. While these systems allow users to express themselves, there is also a negative impact to these systems. Particularly, these social media have become a fruitful ground for inflamed discussions, that usually creates a separation between 'us' against 'resulting to the use of offensive language. Unfortunately, hate speech is becoming an increase topic for discussion in social media[18]and most times it turns out to threaten human lives. Social media platforms therefore find it difficult to identify and censor hateful comments.

3.1 Overview of Discriminatory Speech on social media

It has become important to study and discuss the impact of hate speech on human lives, and its correlation to actual hate crimes. It impact creates direct and indirect effect on its victims, ranging from anger , depression, insecurity, suicidal thoughts etc [7]. It can be used as a tool to harass people, it can also be used to commit opportunistic crimes. As a result it is a power tool to propagate an ideology. Social media could have a direct impact in offline communities and also trigger violence [17]. [1] found that online and offline anti-muslim hate crime impacts upon peoples live causing anxiety, low self esteem, depression and isolation etc. This is particularly strong because of the deliberate effort aimed at threatening and inciting violence.

The Association of Chief Police Officers (ACPO) (2013) noted how online hate activities can damage community cohesion and peace they state that "We understand that hate material can damage community cohesion and create fear, so the police want to work alongside communities and the Internet industry to reduce the harm caused by hate on the Internet" (cited online, 2013). Hate crimes on the internet can also be used to create cloud storage and communicative messages that go beyond the physical or online space[29]. For [38], this means that the scope of hate crime has crossed into the virtual world, therefore [30] argues that hate crimes are capable of sending a message to a particular community.

This opportunity has been greatly explored by far right groups who have engaged in Cyber squatting and google bombing. The refers to when anti-hate pages, google searches and wider online sources are used to create content to target a group of people. This method has been used by far-right groups such as the Anti-Defamation League, the English Defence League and now Britain First, who have used the Internet to successfully create some level of intolerance and spread hate. For example, Social media sites like Facebook and twitter where been used to facilitate hate ideologies [23] argues that:"The increase in the use of the Internet as a vehicle for hate is therefore seemingly undeniable, be it organized hate groups or those expressing prejudice and hostility in a more casual manner." Such degree of online hate has triggered events like that of Paris attacks in 2015 and Brussels terrorist attacks in 2016 [2]. [53] argues that such incidents could have a recursive effects on people with similar characteristics, background or identity with the offenders. Social media as platform has been very profitable in terms of its increase in number of users and financial growth. However, despite these impressive growth level, in may 2013, platform like Facebook was forced to reply series of letters by some prominent female celebrities who were threatened with violence. This resulted in a demand that Facebook take action towards gender based discrimination or stereo-type and threats of violence against women. The letter stated that Facebook should include a zero tolerance to its algorithm towards jokes about rape. In addition, due to the recent refugee crisis, a lot of immigrants choose Europe as a safe haven. Meanwhile Germany began to notice patterns of racist abuse and hate speech towards immigrants both online and offline. Mark Zuckerberg who met with the German chancellor Angela Merkel stated: "Hate speech has no place on Facebook and in our community... until recently in Germany I don't think we were doing a good enough job, and I think we will continue needing to do a better and better job" (Associated Press, 2016). Facebook operates using a set of community standards which includes and defines hate speech as "a content that attacks people based on their actual or perceived race, ethnicity, national origin, religion, sex, gender, sexual orientation, disability or disease" is not allowed. We do, however, allow clear attempts at humour.

[41] A key researcher in the field also studied the characteristics and connectivity patterns in online far right protest against refugee housing, with an aim to understand the temporal behaviour and social media activities of the protest movement as well as the correctness of the interactions of the participants. They analyze data from 2015 which contain over one million interaction form over 200,000 users. They found several activity metrics like number of posts issued, negative polarity in comment , and user engagement to the pick in late 2015, coinciding with chancellor Angela Merkel's much criticized decision of September 2015 to temporarily admit the entry of Syrian refugees to Germany. This decision by the chancellor resulted in a step rise in the far right groups mostly in form of political parties;it includes AfD – Alternative für Deutschland meaning "Alternative for Germany". Since inception in 2012 until 2015 they managed to enter the European Parliament. Meanwhile other far right group took to social media to propagate their ideologies. The researcher observed a close relationship between there

gain in high followers to the rise in user engagement, and continuous interaction on Facebook. More so, these refugees were housed and registered in various camps around Germany, this lead to a rise in refugee attacks, such as arson of buildings, use of explosives on building set aside for refugees. Making theses refugee the target of hate crimes in 2015.[45] argues that communication between these anti- refugee housing movements were carried out via a dedicated Facebook page, which there most topics of discuss includes racist, xenophobic and Islamophobic views. [45] aims to answer two research questions

(i) What are the temporal characteristics of the social media activities of this protest movement? (ii) What is the degree of connectedness and cooperation in this protest movement?

The aim of the first question is to get insight on the pattern of activity of the protest movements, and the general content of posted on these pages, and somehow find hints that could relate to their external activities. The second questions focus on the connectedness of individuals on these groups , and their relationship with other pages belonging to political parties. The researcher analysed the time course of pages posts, to understand which time of the year where there so much post or activity on the page, they observed a pick in the 3rd quarter of 2015 which correlate with the admission of Syrian refugees to Germany.[50] similar observation have been made from previous research of far right user engagements on social media [42].

Also they studied the polarity in users comment for this the research used part of speech tagging to all comments to understand the positivity or negativity level in the comment section

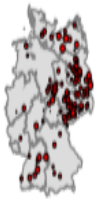

Figure 1: Geolocation of the pages.

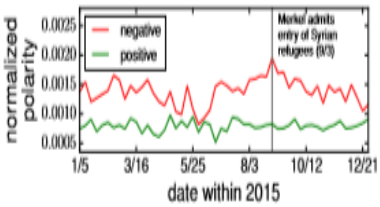

Figure 2: Time course of normalized polarity in user comments.

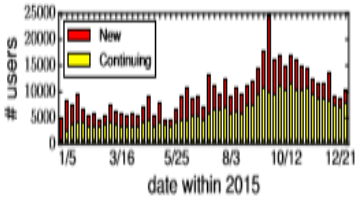

Figure 3: Weekly active users on the protest pages in 2015.

Figure 1: Retrieved from Research

In Figure 2, they observed an spike in negative sentiment leading to September when the chancellor confirmed the movements of refugees into Germany which confirms that the decision provoked wide spread anger in the far-right political group. In addition to this the researcher studied,the ability of this Facebook page to attract users over time,the users where split into 2 categories new users and continuing users, the corresponding user for all the weeks in 2015 is shown in fig 3, they observed a slight increase in the number of new users and existing users.

However this numbers began to drop towards the end of the year.Also the number of average weekly active users diminished (9,935). They encounter a strong correlation of the numbers of continuing users with the time index, but fail to determine a significant correlation for new users, in addition they found that the protest page was successful in attracting new users but maintained a low constant growth of users. They concluded that they encounter several peaks in activity metric which proved that chancel mikel's decision to allow refugees angered several far right organisations. However despite the mobilization effect carried out by these protest pages , there evidence suggests low growth of users in these pages.Which contradicts previous claims that AFD is not a far right organisation. The relevance of this experience in Germany to work is the fact that discriminatory speech can be easily spread via social media platforms and these can have direct and indirect impact on the lives of people, both offline and online.

[12] A key researcher in the field of digital discrimination discussed the challenges of hate speech detection from offensive languages, in his work the researcher tried to distinguish certain hate words from offensive words. Firstly, he defines hate speech as "a language used to express hatred towards a targeted group of people or is intended to be derogatory, to humiliate or to insult members of a group. With this in mind, he went further to explain the difference and context with which hate speech differs from offensive words e.g most African Americans use the term n*gga during their day to day interaction, which could mean "a person or my friend" while other use the term "b*tch or h*e" when referring to rap songs e.t.c this languages are often used in social media, as a result of this contextual effect to hate speech detection, his research to creating a usable hate speech detection system was inspired. They trained a model to detect the difference between these categories and analyze the results to understand how these categories can be distinguished. Their data set is made up of 25k tweets with terms from a lexicon, these text were manually coded into 3 categories; hate speech, offensive language , neither. After a series of careful coding, it resulted in 24,802 labelled tweets. Only 5 percent of the tweets were labeled hateful,1.3 percent was coded anonymously, this was much lower than a comparable study using tweeter which 11.6 percent of speech were flagged hateful [4]. This could be a result of the strict measures used in categorizing hate speech. Similarly, greater percentage of tweets were categorized as offensive language (73 percent at 2/3 and 53 percent at 3/3). They created features using bigram, unigram, trigram features, each was measure by its TF-IDF. In other to capture the syntactic structure there used NLTK to construct the Penn Part-of-speech.Also d Flesch-Kincaid Grade Level and Flesch Reading Ease scores was used to capture the quality of each tweet.

A variety of models used from previous detection tasks were tested including;logistic regression, naıve Bayes, decision trees, random forests, and SVMs. Each model was tested using 5 fold cross validation technique to prevent overfitting while 10 percent of the sample for evaluation was held. After several iterations they observed that linear regression model had better performance with L2 regularization. They trained the entire data set using this model and used it to predict the label for each tweet. They concluded that, when carrying out hate speech detec-

tion task, it is important to adequately differentiate between hate speech and offensive language due to the legal implication of hate speech. He noted that, using lexical method are effective in identifying possible term but are not accurate enough to identify hate speech, only a small amount of tweets were considered hate speech by human coders. Meanwhile automated classification method is efficient in differentiating between these classes, a closer analysis of the result shows that the presence or absence of particular hateful or offensive terms can both help and/or hinder the accuracy of classification. There result also show how hate-speech can be used in different contexts,it can be directed to a person or group of people, it can also be directed towards nobody in particular and i can be used during communication between 2 people. He advises that further work should be focused on the context in which hate speech is used and the individuals motivation for using it. While these research has uncovered interesting findings, it fails to visualize insights on the most topics of discuss, because when you understand what hate speech users talk about, you are one step closer to their personality and motivation. That is why we use sentiment analysis and topic modelling to achieve this goal in this research.

3.2 Far-Right Political Discrimination and Ideology

"The continuous increase in right wing politics has created a political and social environment that has made the extreme far-right feel more comfortable in expressing their narrative and cause publicly" [33]. These narratives and discuss which are being inspired by social media post have been on the rise. In June 2016, the violence resulting from hate crime in the UK that led to the killing of politician Jo Cox by Thomas Mir who was influenced by Noe-Nazi ideology, to assault gay men,Jews, Asians, and Muslims. These acts of murder lead the UK government to declare 3 extreme far-right groups as terrorists organisations, namely; National Action in December 2016, Scottish Dawn in September 2017. Their actions included continuous support for terrorists acts leading to the death of Jo Cox, former UK Home Secretary, Amber Rudd said: "National Action is a vile racist, homophobic and anti-Semitic group which glorifies violence and stirs up hatred while promoting their poisonous ideology and I will not allow them to masquerade under different names. ... Our priority as a government will always be to maintain the safety and security of families and communities across the United Kingdom and we will continue to identify and ban any terrorist group which threatens this, whatever their ideology".

Consequently, another extreme far-right group emanated from the national action group with the same socialist ideology called System Resistance Network (SRN) at the point of writing his literature, [33] said this organisation was one year old. Formed by the leader of National Action Alex Davies they focused on posting hate narratives around the UK, but recently encouraged his followers to read Hitler's Mein Kampf and other national socialist literature as a result the people in the UK are calling that this group be prescribed as a terrorist group. One major problem social media platform faces is the legislative rules set by states, for example the Christchurch shooting which was recorded on Facebook, there has been several petitions advo-

cating that these platforms regulate the spread of such violence and hate. For example, while consistently monitoring and closing Islamic State accounts, Twitter closed the first Britain's far-right account UK including its founder Paul Golding and Jeyda Fransen's accounts and more recently those associated with Tommy Robinson. While these companies have algorithms to identify hate posts, many are not picked up by the algorithms. Companies such as Facebook, YouTube and Twitter have internal regulation regarding the content they post as hate or violence , however more effort needs to be put into the supervision of hate speech online.

3.3 Hate speech

Hate speech can be defined as a narrative that is capable to cause harm or emotional damage to an individual. It is a bias-driven, aggressive, hateful speech aimed at a group of people because of some of their actual or supposed distinctive appearances [8]. It is intended to cause harm and promote discrimination. The social media environment has created and opportunity for the sharing and exchange of hate to thrive.

3.3.1

Understanding hate speech Hate speech has been an interesting area of research in the field of sociology, [13]. Particularly, [34] claims that some forms of hate is not close to being solved in the society of, mostly hate towards gender and race. Hate speech resulting from such biases are moderately popular and the system has created policies to solve them. However, there has been a several undesirable consequences of these policies (e.g., incivility, tension, censorship, and reverse discrimination) as a result of protecting hate victims and prosecuting haters.Overtime, this tension has driven the evolution of standard policies to regulate hate speech. Studies supported by [20], UNESCO, reviews the increasing issues of hate speech on media platforms from a legal and social standpoint. They report that platforms like Facebook and Twitter have primarily adopted only a reactive approach to deal with hate speech reported by their users, but these platforms need to do more to curtail such re-occurrence. Their study reports "These platforms have access to a lot of data that can be analyzed and combined with real life events that would enable a better understanding of the consequences of hate speech activity online". Back in 2004, there has been attempts study websites to understand the level of racist or political extremism activities going on in them. [22]. Nowadays, there are a lot of issue under study related to social media platforms and racism or discrimination. [6],[16]. Though, these techniques don't give a data driven view of hate speech in online media today, we aim to bridge this gap.

3.3.2 Hate group detection using web mining and Social networks analysis

With this context, it is not surprising that most current efforts are motivated by the impulse to measure, detect and eliminate hateful messages or hate speech. Researchers like, [48] has focused on detecting hate speech through web mining, natural language processing and social

networks analysis, with an attempt to propose an approach (architecture) for hate group detection. In this study, the researcher takes an experimental approach, to measure the performance of hate groups detection under different environments data from Facebook. During the research he came up with a method to detect hate speech on facebook platform that involves text mining and social networks.

While these detection approaches are great it only considers media like face book, which is one amongst many social media of high impact, also the research provides a narrow implementation of such approach to discriminatory topics like racism. This research aims to expand on the topics and also provides suggestions for better platform politics.

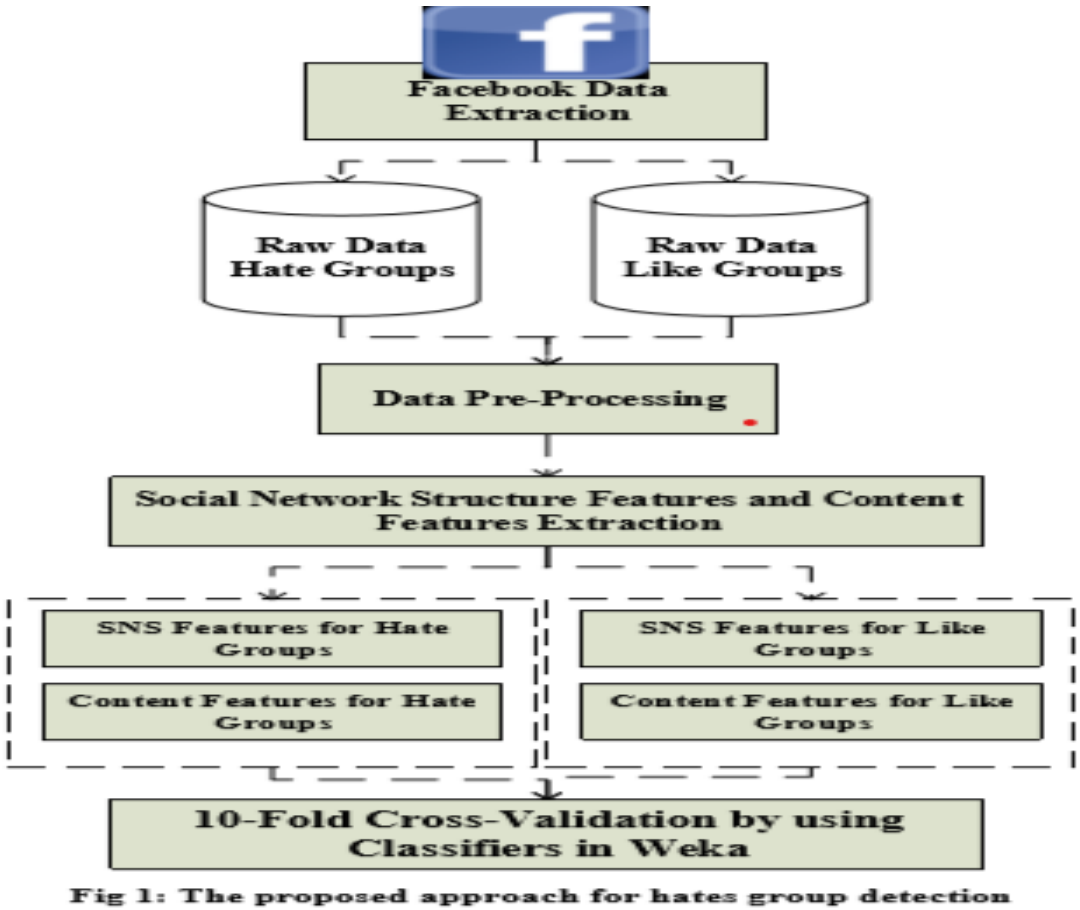

Figure 2: Web mining and Social networks Analysis

3.4 Sentiment Analysis

Previously, studies have analyzed data from social media platforms using sentiment analysis to come up with opinions about individuals on such platforms [21] applied a lexicon-based approach hate speech detection, with an aim to create a classifier that detects hate speech (racism, religion, nationality), to achieve this he analyzed 8000 words from blog post, new comment section. The researcher takes a series of steps to achieve their goal;

Step 1: They used a rule-learning approach to extract subjective sentences.

Step 2: Use subjective sentences identified in step 1 above, to extract semantic and subjective word features.

Step 3: Use bootstrapping, to augment the lexicon in step 2 with noun patterns based on the semantic classes of religion, ethnicity and race and hate-related verbs.

Step 4: Build and test the classifier with the annotated corpus based on the features identified in Step 2 and Step 3. From there efforts a classifier was built to help predict and identify a subjective sentence.

```
Predict subjSent  algorithm

Input: d: Text Document, sl: SUBJCLUE lexicon
Output: count: count of subjective words
//initialize list and count
Subjsentlist: list of subjective sentences
count←0
sentence ← ""
Begin
While (d! = null)
        sentence ← split (d)
        word← ""
        lex ← ""
                For each sentence ∈d
                        For each word ∈sentence
                        word←CRFtagger(word)
                          If word matches lex then
                              Count++
                                  If count >=2 then
                                    Output count
                                  End if
                                  addSubjsentlist(sentence)
                        End if
                         End for
                End for
End while
```

Figure 1. Subjective Sentence Prediction

Similarly, studies by [44] focused on methods to measure, identify and prevent hate speech. In the study, the researcher applies a quantitative and qualitative approach to analyze twitter data. [44] defines hate speech as any offense motivated, in whole or in a part, by the offender's bias against an aspect of a group of people. Therefore, they propose a simple way to detect hate speech using sentence structure; I <intensity><userintent><hatetarget> Thus, in other to find hate targets from their datasets, two templates were designed. The first template for the <hate target> token as simply "<one word> people". For example, search for patterns like "Pregnant " or "Indian People". As result this template helps capture when hate is directed to a group of people. From the application of the template on the dataset, they found the top ten hate targets groups on twitter.

Though, sentences structures and sentiment analysis approaches have yielded result in the field of Discriminatory speech [22,3] they spend a lot of effort applying such techniques only

on nationality, racism, and sexuality.

3.5 Sentence-level Subjectivity Detection

With sentence level subjectivity, each sentence is a given document which is analyzed and checked to be subjective. The sentence can be classified as positive or negative. [36] use a subjectivity detector to remove objective sentences from a given document. Then, they use minimum cuts formulation, they integrate inter sentence level information with bag-of-words features. They also report considerable improvements over a baseline word vector classifier. To learn subjective sentences [39] use two bootstrapping algorithms [40], [47] to learn lists of words from a collection of texts. Then they train a subjectivity classifier on a small set of annotated data using the words as features along with some other previously identified subjectivity features. A sentence is classified as subjective if it contains an expression with a medium to high intensity otherwise it is classified as objective. This ensures only those sentences that are clearly subjective are classified as so. Besides identifying subjectivity and polarity of a sentence, [54] classify the strength of the opinions and emotions being expressed in individual parts, considering clauses down to four levels. They take advantage of several syntactic clues as well as subjectivity features tested in past research to recognize the subjectivity strength of a clause. [14] explored the idea of inter-sentential and intra-sentential sentiment consistency using natural language processing. Instead of finding field dependent view words, they showed that the same word could indicate different orientations in different contexts even in the same field. Thus, they proposed to use aspect and opinion word pair to capture the context of the sentiment.

3.6 Lexicon Building

Previous researcher has also approached the problem of hate speech using the lexicon building , they attempt to create sentimental words that represent negative and positive. [10],[37] There are 2 major categories for creating such opinion lexicon, the corpus-based and dictionary-based approach. The later involves creating a static dictionary of semantically related words tagged with both semantic label and polarity orientation score or reliability label. [46], [55] The dictionary approach is primarily generated using a bootstrapping strategy that uses a set of seed opinion words and an online dictionary like WordNet [19].There are lots of relevant resources on opinion lexicons built from mainly adjectives, but also from adverbs, nouns, and verbs.[46], [26], Esuli et al in [19] used semi supervised approach with WordNet term relationships such as hyponymy, antonym and synonym to automatically create a lexical resource that assigns each set of words from WordNet into 3 sentiment scores adding up to one, namely; objectivity, positivity ,negativity. They used core seed of words that where previously known to carry positive, negative or objective sentiment and gradually add new synset using WordNet relations.

Dictionary based approaches generally find it difficult to find words with context specific

orientation. While corpuses-based approaches use domain corpus to reflect opinion words with a preferred pattern. In other to determine the semantic orientation of words, Natural language processing rule-based techniques, syntactic, structural and sentence level are used to select the words and phrases to be included in the opinion lexicon. With this method a lexicon is filled with more useful words that can be incorporated contextually to the features that could amplify or reduce the intensity of close lexicon items. In [55] Wilson et al used a phrase-level sentiment analysis approach that first determines whether an expression is neutral or polar and then disambiguate the polarity of the polar expression.

3.7 Deep Learning technique

Similarly, Researchers like [3]have explored deep learning architectures to address the issue of hate speech on tweeter. In there work they experiment using methods such as SVMs Random Forest, Logistic Regression, Deep Neural Networks (DNNs) and Gradient Boosted Decision Trees (GBDTs). These classifiers are in turn defined by tasks specific embeddings learned using a combination of 3 deep learning methods, namely; Fast Text, Convolutional Neural Networks (CNNs), Long Short-Term Memory Networks (LSTMs). As baseline, they compared char n-grams [51]Tf-IDF vectors and bag of words. They focus on using these methods to examine the application of deep learning on the task of hate speech detection. Secondly, they explored various tweets semantics embedded in char n-grams, bag of words etc. Having provided a baseline approach, they investigate using neural network architecture, described as follows; Random Embeddings or GloVe embeddings [15]CNN: Inspired by Kim et al [32]worked on using CNNs for sentiment classification, leveraged CNN for hate speech detection. They used LSTMs to capture long range dependencies in tweets, which plays a role in hate speech detection.

We have discussed approaches and techniques used to identify and detect hate speech on social media platforms. Some have proposed an architecture to detect hate speech on facebook platform that involves web/text mining and social networks. Others used sentiment analysis a lexicon-based approach to understand individuals' opinions, they also create a classifier the helps detect hate speech from the texts data set. Also, sentence level subjectivity was discussed, an approach used to learn subjective sentences with the use two bootstrapping algorithms and learn lists of words from a collection of texts. Then they train a subjectivity classifier on a small set of annotated data using the words as features along with some other previously identified subjectivity features. While these efforts are contributory, firstly we focus on discriminatory categories such as Racism, Gender, Political extremism and sexuality. Secondly, we will use topic modeling and sentiment analysis techniques because it helps us understand the most discuss topics under these hate categories and their emotions behind them.[45] argues that most far right groups communicate via social media platforms, therefore the best way to identify these groups and begin a digital transformation, is to understand there topics of discuss before categorizing them into groups. The project will provide new insights into the circulation of hate

(bridging) speech and recommending strategies for more inclusive government and platform politics.

4 Data Set

Now we briefly describe the methodology used to gather data from a popular social media; Twitter. Twitter is an American micro blogging and social networking service launched in June 16th 2006, it enables users to post messages and interact, these short messages are called "tweets". There are two categories of users: the registered users and unregistered users. Registered users can tweet, like, retweet tweets but unregistered users can only read them. Within a short time-span twitter has become a very popular social networking service with over 330 million monthly active users.34 percent of twitters users are females,66 percent are males, 22 percent of US adults use twitter while 21 percent of all internet female users use twitter also 42.5percent.[44]of Twitter users are on the platform daily, which makes it a valuable venue for studying hate speech. We found that despite the anonymity of twitter theirs existed traced of hate discuss [5]] and we decided to use twitter in this study. We begin with hate speech lexicon identified by users as hate speech and compiled by hatebase.org, we search for words in the (hate base, a database for hate words) with intensity greater than 50 percent.

From our definition of hate speech, a word is said to be hateful or discriminatory if its;

- Uses a sexist, gender or racial slur.
- Attacks the minority
- Criticizes the minority without a well founded argument.
- Promotes hate or violent crime.
- Initiate inferiority in a group of people.
- Supports a problematic hashtag.
- Negatively stereotypes minority
- Defends xenophobia or sexism

From this database we select hate words that affect (race, gender, sexuality and politics). Using twitter API , we search for tweets that contain terms from this lexicon and extract data from twitter, then we analyze them using sentiment analysis and topic modelling to identify the emotions and the most popular words of discourse. We study 4 variable which includes; Gender, sexuality, race and politics for us to extract data from twitter we use some key search words see table below. Meanwhile for politics variable we study a group of individuals or groups who identify with different ideologies, both far right and far left.

Table 1 Race

S/N	Variable	Search Words	Intensity
1	Race	White Supremacist's	Extremely offensive
2	Race	Nigga	Extremely offensive
3	Race	Nigga	Extremely Offensive
4	Race	Dindu	Extremely Offensive

Table 2 Sexuality

S/N	Variable	Search Words	Intensity
1	Sexuality	Gay People	Extremely offensive
2	Sexuality	Faggot	Extremely offensive

Table 3 Gender

S/N	Variable	Search Words	Intensity
1	Gender	Cunt	Extremely offensive
2	Gender	Bitch	Extremely offensive
3	Gender	whore	Extremely offensive
4	Gender	Tranny	Extremely offensive
5	Gender	Sexiest	Extremely offensive
6	Gender	conchuda	Extremely offensive

Table 4 Politics

S/N	Variable	Search Handles
1	Politics	farright
2	Politics	robjeffecology
3	Politics	American3rdP
4	Politics	Gothamist
5	Politics	ProhibitionUS
6	Politics	neenCa
7	Politics	farleft
8	Politics	ProhibitionUS
9	Politics	neenCa
10	Politics	Black Riders
11	Politics	Communist Party USA
12	Politics	FSPUS
13	Politics	LiberationFL
14	Politics	SocialistAct

Our data set consist of 1 year period for politics and 6 months for other variables which in total 108,890 tweets.

4.1 Methodology

In this section we explain the approach used for analysing the data set.

4.2 Data collection

Firstly, data was collected from twitter using twitter API and data scrapping techniques .The data set consists of 108,890 tweets within a period of 6months- 1 year period, in a csv format and directly import to R studio for analysis. After the collection of the data another important step in the process in cleaning and pre processing.

4.3 Pre-processing

In order to analyze these text we removed, punctuation's, stop words, white spaces, links, numbers, strange symbols from my texts. Also letters were converted into small letter to make them the same word. I applied the package tm-map(). Also we created a custom function to remove links and strange symbols. Gsub('http//[[alum]] [[punct]]).These cleaned text where then transformed using content-transformer so that tm-map () can understand it

- To lower- used this to convert to lowercase letters
- Remove punctuations-used this to remove dashes, dots, commas
- Remove numbers- used this to remove number
- Strip white space- used this to removes spaces
- Removed words- this was used to remove specific words by me.

4.4 Data Analysis

In order to analyze the words originating from the tweets data set, we created a corpus. These corpuses was converted to document term matrix which enabled us to check the most frequent words used in our data set. Also, we used lexicon based on unigram to assign score negative, positive sentiment and emotions such as , joy, disgust, anticipation, fear, sadness, surprise, and trust. The sentiment method used for our analysis was NRC.

4.4.1 Technique

Topic modeling technique was used, also it is an unsupervised classification approach used cluster numeric data, which helps find natural groups of items. In the case of words, we use topic modeling to categories the most topics of discussion in our data set. LDA was used for the purpose of this analysis. Latent Dirichlet allocation (LDA) is a particularly popular approach for fitting topic models. Here it divides each document to a combination of topics, and each topic to a mixture of words. Without diving into the mathematical application topic model can be understood in two ways.

Every Document is a mixture of topics- that is in 2 document we could say this document 1 contains 80 percent of topic A and 20 percent of topic B, while document 2 contains 30 percent of topic A and 70 percent of topic B.

Every topic is a combination of words- that is for example topics in politics could include “far right”, “far left”, “Donald trump” etc. LDA is mathematical model used to detect most relevant topics in a document. In this analysis we used topic model to understand people’s discussion on our data set.

4.5 Data Visualization

At the end of our analysis we visualize of results which include sentiment analysis, NRC emotions, word cloud showing topics of discussion as well as topic models. this in turn help us visually understand the narrative going on in our data set.

5 Result

This section will discuss the findings from our data set in 4 categories namely; Gender, Race, Sexuality, Politics. The aim of this work was to apply sentiment analysis and topic modelling to detect and gain deep insight on the presence of hate speech on twitter.And provide recommendations for better platform policies. Also main advantage of this insight is to understand what individuals are saying on these platforms. [1] argues that these hateful comments have significant effects on the lives of people there for our results and methods enables stakeholders to be more aware of the words and targets of hate online.

5.1 Gender

The data set for the Gender variable consists of 41,787 tweets. During the analysis we apply sentiment analysis on our data set, to understand the sentiment of individuals discus and found that there is a significant amount of gender discrimination and hate speech on twitter. Considering our data set and time frame which data was collected as in *figure 3*. We found that alot of words used to describe gender on twitter containing negative sentiments. this supports [30] augments which states that online hate has crossed the digital world, and now has effect on people. In 2013, several female celebrities filed reports and letter to facebook regarding abusive and life threatening comments made about them [53].

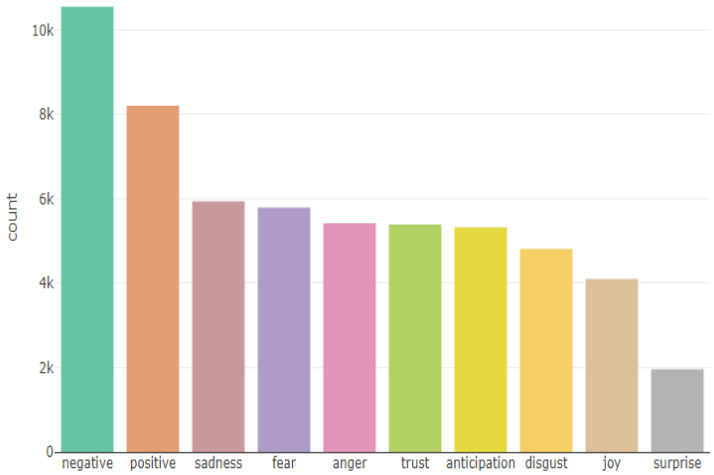

Figure 3: Gender Sentiment Analysis

In *figure 3* over 20 percent of our data consist of negative sentiment. while about 12.5 percent positive sentiment gotten from these comments. In view of this, we visualized a word cloud to better understand the words of discussion in our data set as seen below.

From *figure 3* i observe that the most highlighted words appears to be the most frequent therefore representing negative sentiment. [30] Argues that hate speech on digital platforms can have significant effects on its victims like; depressions, anxiety, even suicidal thoughts.

Figure 4: Gender Word cloud

On the word cloud we have words like small b**bs, b*tch, girl,ridicule, enormous. This help explains why we have negative sentiment from our analysis.We went a step further to count the number of words and see the most frequent.In order to clarify this assumption we visualize the most frequent words used over 10000 times, to clearly understand the topic of discussion. We can observe that word bitch,girl,small,boobs happen to be the most frequent words. Such expression to the female gender is highly offensive, this gives a need to prevent such words on digital platforms like twitter.This presences such gender abuse on social media supports [53] research where Facebook was forced to respond to letter from female celebrities regarding threats.

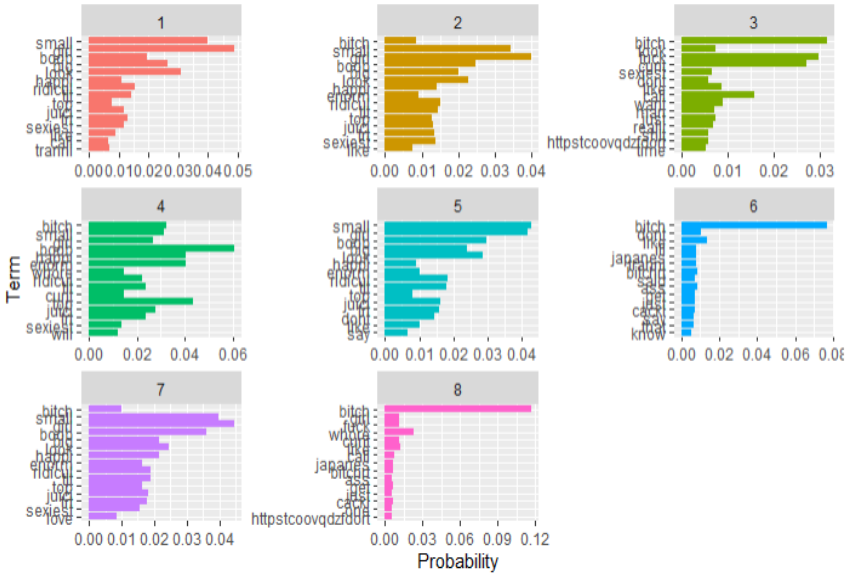

Figure 5: Gender Topics (X-axis probability, Y-axis Terms)

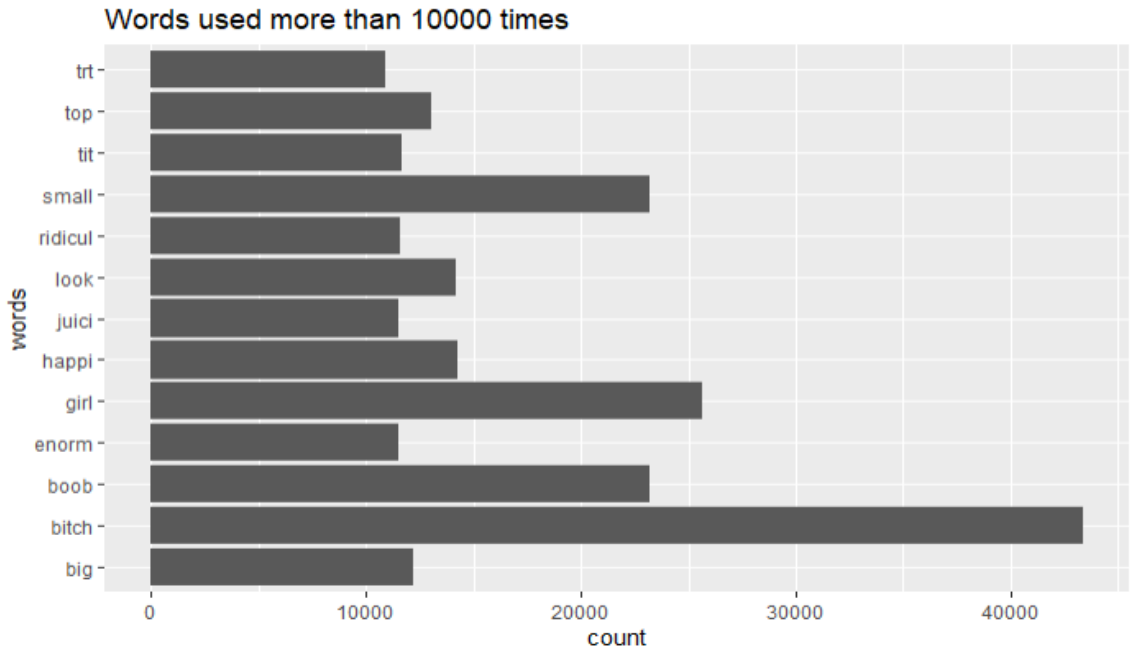

Figure 6: Most Frequent Words

The purpose of a topic model in this research is to understand and calculate the probabilities of a word occurring from our data set. therefore, a topic model was carried out on our data set to see the most discussed topics and the probability of it being discussed or used. As seen above .From this visual representation we found that words like b*tch,small,b**bs,whore, c*nt. Have very high probabilities of occurring. this confirms that one the the most used gender slurs were just mentioned. If platforms like twitter can pay attention to this words or add word detectors. it is evident that gender discrimination or abuse will reduce drastically from the platform. thereby saving lives and improving the sanity of social media.

5.2 Politics

The politics variable consist of 29,710 tweets which was studied from both far right and far left political ideologies. [42] studied the connectivity between characteristics and connectivity patterns in online far right protest against refugee housing and found using some metrics like user engagement and user activity , to understand that the decision of chancellor merkel in 2015 to admit immigrants from Syria, raise a lot of far right groups to protest. Surprisingly, after carrying out sentiment analysis on these tweets, we found out that a positive sentiment exists within their discussion. But we observed narrative and words such as; people, coronavirus, parti, report,health worker. While these findings might appear random , they represent the series of events that occurred during the data collection period. Also the prevalence of positive sentiment to negative, might be because the political persons under study where tweeting from their personal accounts. In the future studies should be carried out on individuals in a group. *Figure 7* shows that 25.966 words from our data set represents a positive sentiment while 24.005 represents negative sentiment, which not a huge difference.This is particular an interesting finding considering the series of negative event that occurs politically during this time frame.

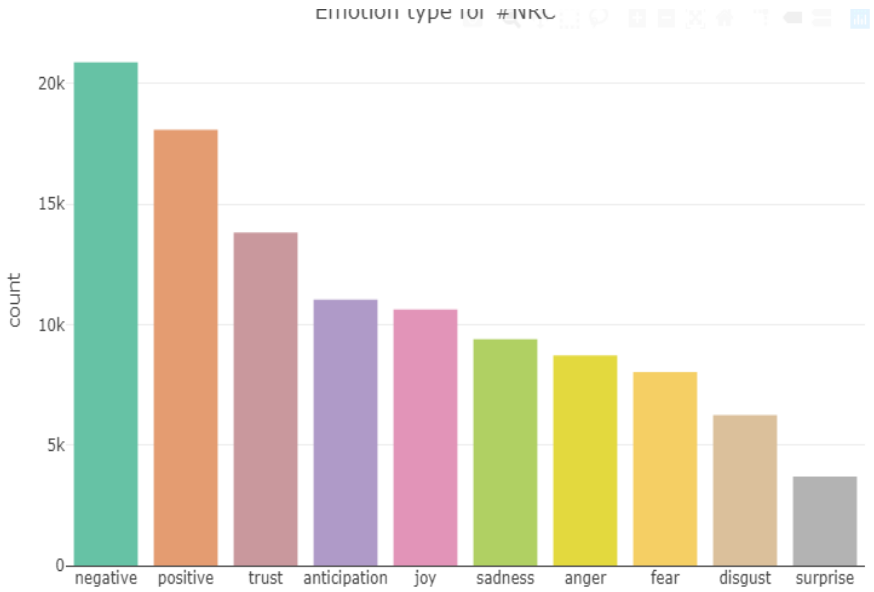

Figure 7: Politics Sentiment

In addition *Figure 9* explains better the content of our data set, as to why we have more positive sentiment than negative. The word cloud contains words like trump,vote, parti, people, media, social, support etc. This cloud represents the serious of activities that occurred during the time frame. In other to understand why positive sentiment, we dive in deeper to understand out data set by visualizing the most occurring topic of discuss in politics and the probabilities of occurrence.

The topic model (*figure 8*) shows the top 8 topic and words associated with them.

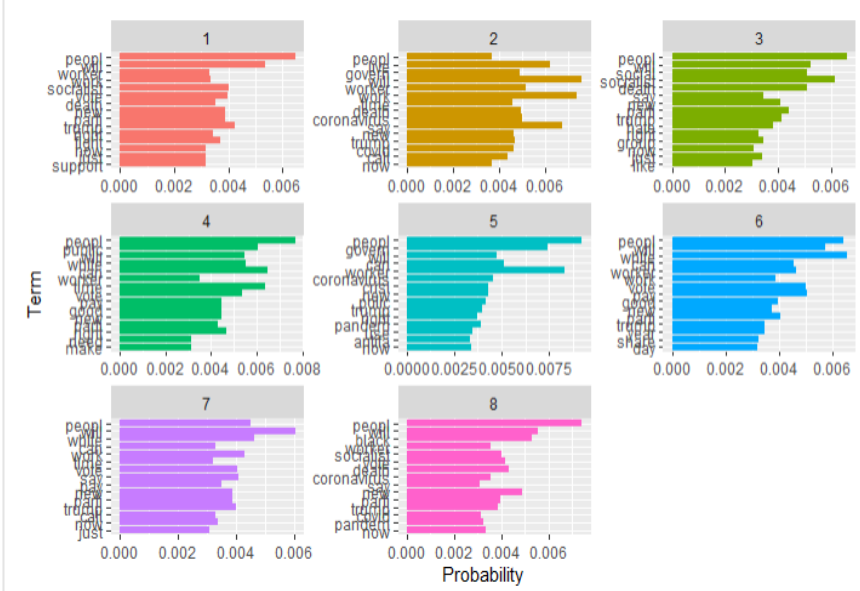

Figure 8: Politics Topic Model (X-axis probability, Y-axis Terms)

The *Figure 8* display people as the having the highest probability of occurrence followed by people,govern,worker,black,white, black, coronavirus, worker etc. All of these text represent positive sentiment. These word despite having positive sentiment could be affected by the context of discuss.[12] argues that there exists the problem of offensive language on hate speech detection, such logic could be adapted in this case. words like coronavirus exist, people, but when the contexts is evaluated in the real life , you will agree with me that they represent a negative event. This supports his conclusion that the contexts should be looked into when evaluating sentiments of individuals. Further research should focus on the context of discussion , who has been referred to and the motive for such words.

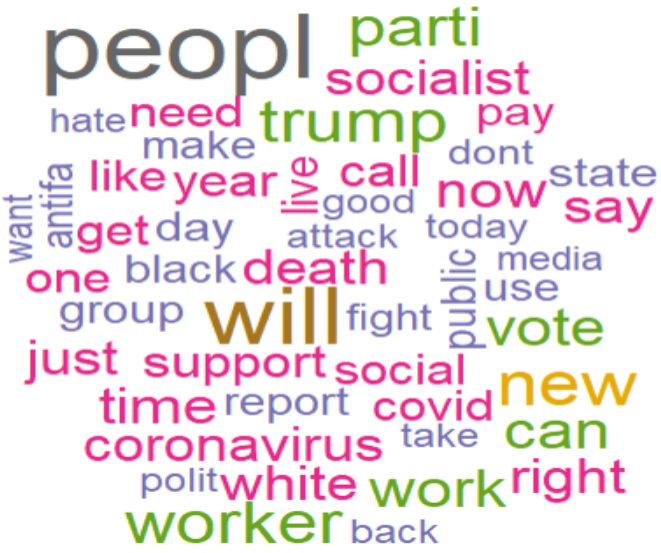

Figure 9: Politics Word cloud

5.3 Sexuality

Similarly an interesting discovery was made after analyzing the data about sexuality. *Figure 10* provides insight on the most frequent words used in the data set is people followed by gay, yet the sentiment shows a negative perception from people on twitter. This caught my attention during analysis. About 10.551 words portray negative sentiment while 8201 represent a positive sentiment amongst others.As discussed above every individual has the right choose his/her sexual orientation. But the presence of violent and aggressive words used against people can affect their lives in various ways see *figure 10*.

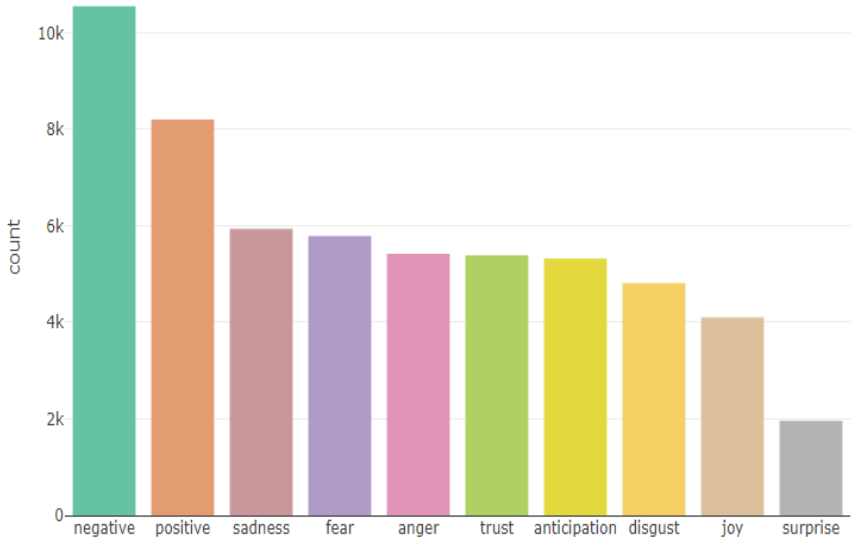

Figure 10: Sexuality Sentiment bar plot

Figure 11: Sexuality Word Cloud

Also in other to take a closer look at our data set we visualize the word cloud to see the composition of words that make up this sentiment.From the words cloud we discover a combination of words that represent discrimination, words like; homophobic, fargot, gay people etc. The major advantage of applying sentiment analysis on platforms data set is that , it exposes your mind to realities of discrimination and how it affects the society. In addition topic model was carried out to evaluate the probabilities of certain words occurring.*Figure 12* also provides valuable insight into the data set by showing words with the highest probability of occurrence.

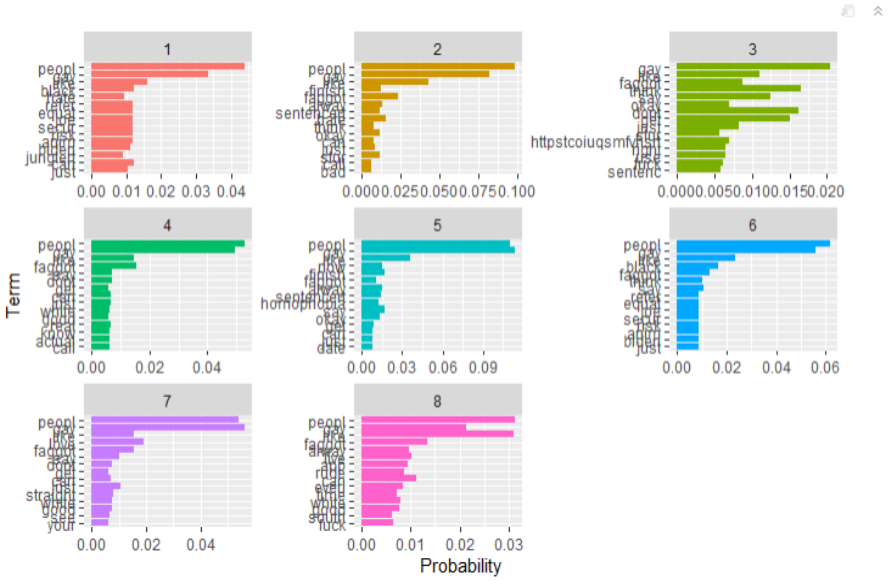

Figure 12: Sexuality Topic Model (X-axis probability, Y-axis Terms)

5.4 Race

Our data consist of 14,627 tweets, after applying sentiment analysis in other to understand the sentiment of individuals, we discovered that a negative sentiment dominates the tweets. This discovery in turn was not surprising due to the sensitive nature of the topic of discuss"race". *See figure 12*. This finding is supported by [53] studies which say that platform like Facebook and twitter should pay more attention to issues like gender stereotype, racism,digital inequality,immigration and sexual orientation.

Furthermore, we analyze the data deeper to visualize the words that make up these sentiments and we discovered certain racial slurs like, nigger,nigga,black white, dindu, nuffin etc.See word cloud below. Irrespective of the content or context in which it is used these words, do not represent a friendly conversation. The consistent use of words like this can also motivate violence [2] say that the use of words like this triggered the 2015 paris bombing. Such events end up having an after effect on people, directly or indirectly. Also we carried out topic modelling with the aim of understanding the most topic of discussion. it was discovered that words like ,

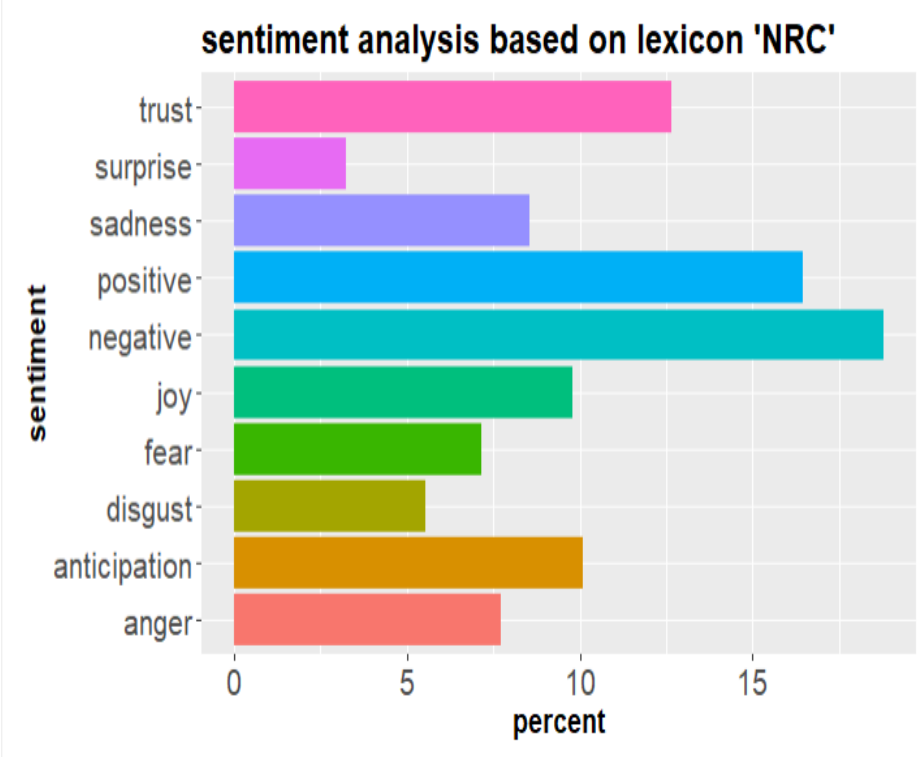

Figure 13: Race percentage, Sentiment Analysis

white,people,black,nigga,dindu had the highest probability of occurrence. *See figure 16*. The presence of discriminatory speech on social media can not be overemphasized, as well as the need for platform stakeholders to pay attention to the impact it has on human lives. Therefore the study reveals that there exist racial discriminatory narratives on digital platform like twitter.

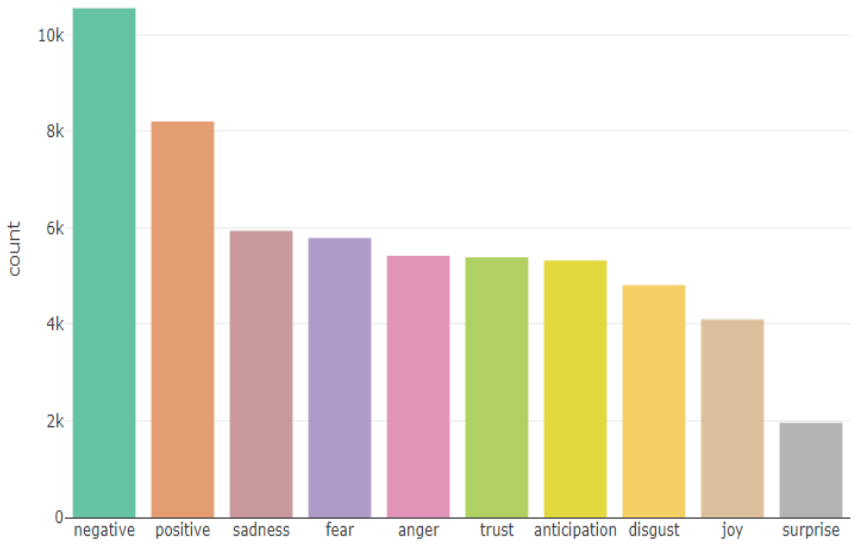

Figure 14: Race Sentiment Analysis

6 Conclusion

This research has uncovered the presence on platform discrimination, which is a new form of discrimination derived from the social media culture and activities. Their design, technical affordance, business models, policies and specific culture of use (like propagation of far right ideologies) has made them an arena for hate speech. Our findings support the results of [44] who analysed the targets of online hate with an attempt to unveil a set of important patterns, used during hate interaction, and found that hate speech is present on these digital platform Our effort applies sentiment analysis and topic modelling to identify hate narrative amongst these variables. We hope that our data set and methodology can help researchers, students, government bodies and private human right institutions make better platform policies. The research has displayed how digital discrimination affects race, politics, gender and sexuality and is directly linked to real life events [18].

Future study should be done will larger data sets and longer time frames, and more variable to capture the activities of individuals on these platforms. Similarly, researcher should identify the difference between hate speech and offensive language because of the context and group of individuals have been addressed. Further research should be focused on understanding the behaviours of the person perpetration hate, and Investigating the reasons and social network behind hate perpetrator, this will in turn enable platform stakeholders make more informed decisions . We show here evidence that there exists hate speech on these platforms.

Figure 15: Race Word Cloud

6.1 Platform recommendations

Platform management should not only automatically take down messages or content with hate speech but also provide reasons why content is no longer available.This will in turn improve platform transparency and curb people who engage in defamation online. Secondly, platform owner or management should incorporate in their algorithms a notification that create awareness against hate speech. Thirdly, government and research institute should invest more in research that help bridge hate speech online. In addition, Further research should be carried out to examine platform racism around racial controversies and different social cultural contexts. Also institutions like GDPR, should create social media data protection laws against discriminatory speech, this will monitor platforms who violate such data policies. Finally platform owners or management could introduce discriminatory batches that warns individuals about the use of hate languages online. this will help identify prejudice and discourage the use of hate Narratives.

We have several suggestions for further research in this area. Taking into account the results of this thesis we recommend a few different perspectives on what specific aspects should be looked at. First of all, a very limited amount of academics have researched identifying and detecting hate speech online. More so, research should be carried out on the economic,social,and political effects of platformed racism or discrimination. Also its important to distinguish the context between hate speech and offensive language. Researchers can approach this problem from a legal standpoint, where the intensity of hate words are categorized in other to create better platform regulations. As could be seen on this thesis, the importance of detecting and moderating and predicting the probability of hate speech is obvious from the strong connection between hate speech and hate crimes, therefore platform stakeholder should create outreach

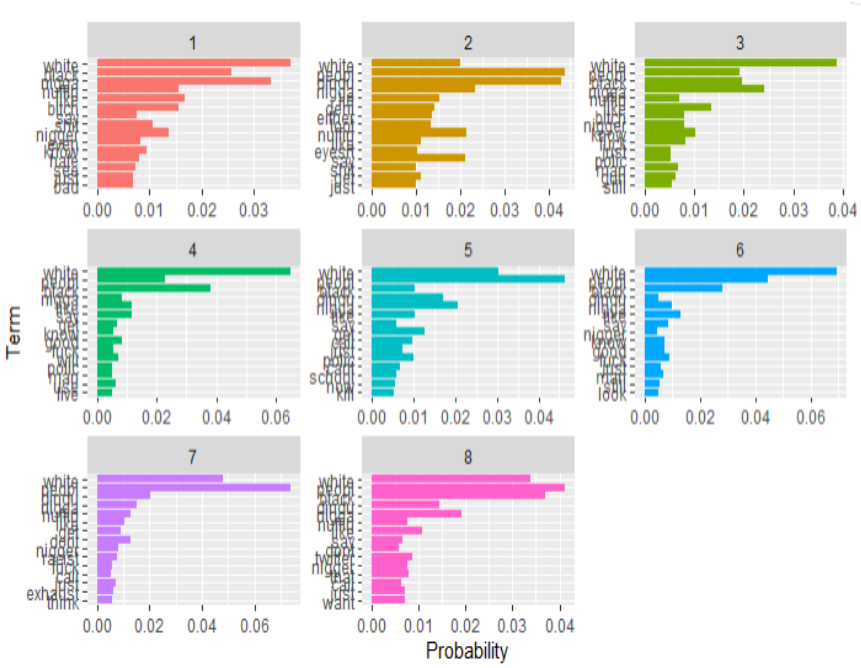

Figure 16: Race Topic Model

programs with an attempt to prevent escalation from speech to action.

Finally, we plan to look more into prediction hate speech, because the best way to prevent hate speech is if it can be predicted ahead it can be totally prevented from occurring. Lastly, the secondary goal of this thesis was to find the presence of hate speech online, which was achieved but further studies could increase the data set, and approach this problem from other languages. This will help build the lexicon and create more awareness against platform discrimination.

References

[1] I. Awan and I. Zempi. Virtual and physical world anti-muslim hate crime. *The British Journal of Criminology*, 57(2):362–380, 2015.

[2] I. Awan and I. Zempi. The affinity between online and offline anti-muslim hate crime: Dynamics and impacts. *Aggression and violent behavior*, 27:1–8, 2016.

[3] P. Badjatiya, S. Gupta, M. Gupta, and V. Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760, 2017.

[4] P. Burnap and M. L. Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[5] I. Chaudhry. # hashtagging hate: Using twitter to track racism online. *First Monday*, 20 (2), 2015.

[6] Y. Chen, Y. Zhou, S. Zhu, and H. Xu. Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 71–80. IEEE, 2012.

[7] K. M. Christopherson. The positive and negative implications of anonymity in internet social interactions:"on the internet, nobody knows you're a dog". *Computers in Human Behavior*, 23(6):3038–3056, 2007.

[8] R. Cohen-Almagor. Fighting hate and bigotry on the internet. *Policy & Internet*, 3(3): 1–26, 2011.

[9] K. Crenshaw, N. Gotanda, G. Peller, and K. Thomas. Critical race theory. *The Key Writings that formed the Movement. New York*, 1995.

[10] Y. Dang, Y. Zhang, and H. Chen. A lexicon-enhanced method for sentiment classification: An experiment on online product reviews. *IEEE Intelligent Systems*, 25(4):46–53, 2009.

[11] J. Daniels. *Cyber racism: White supremacy online and the new attack on civil rights*. Rowman & Littlefield Publishers, 2009.

[12] T. Davidson, D. Warmsley, M. Macy, and I. Weber. Automated hate speech detection and the problem of offensive language. In *Eleventh international aaai conference on web and social media*, 2017.

[13] R. Delgado. *Understanding words that wound*. Routledge, 2019.

[14] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240, 2008.

[15] N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30, 2015.

[16] J. C. S. Dos Rieis, F. B. de Souza, P. O. S. V. de Melo, R. O. Prates, H. Kwak, and J. An. Breaking the news: First impressions matter on online news. In *Ninth International AAAI conference on web and social media*, 2015.

[17] K. M. Douglas, C. McGarty, A.-M. Bliuc, and G. Lala. Understanding cyberhate: Social competition and social creativity in online white supremacist groups. *Social Science Computer Review*, 23(1):68–76, 2005.

[18] L. Eadicicco. This female game developer was harassed so severely on twitter she had to leave her home. *Business Insider*, 12(10), 2014.

[19] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, volume 6, pages 417–422. Citeseer, 2006.

[20] I. Gagliardone, D. Gal, T. Alves, and G. Martinez. *Countering online hate speech*. Unesco Publishing, 2015.

[21] N. D. Gitari, Z. Zuping, H. Damien, and J. Long. A lexicon-based approach for hate speech detection. *International Journal of Multimedia and Ubiquitous Engineering*, 10 (4):215–230, 2015.

[22] E. Greevy and A. F. Smeaton. 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*, pages 468–469, 2004.

[23] N. Hall. Community responses to hate crime. *Community Justice*, page 198, 2013.

[24] C. Hardaker and M. McGlashan. "real men don't hate women": Twitter rape threats and group identity. *Journal of Pragmatics*, 91:80–93, 2016.

[25] E. Hargittai. 13 minding the digital gap: why understanding digital inequality matters. *Media Perspect. 21st century*, page 231, 2010.

[26] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. In *Proceedings of the 35th annual meeting of the association for computational linguistics and eighth conference of the european chapter of the association for computational linguistics*, pages 174–181. Association for Computational Linguistics, 1997.

[27] D. L. Hoffman and T. P. Novak. Bridging the racial divide on the internet, 1998.

[28] M. Hrdina. Identity, activism and hatred: Hate speech against migrants on facebook in the czech republic in 2015. *Naše společnost*, 14(1):38–47, 2016.

[29] P. Iganski. Hate crime: taking stock: programmes for offenders of hate, 2012.

[30] V. A. Imran and B. Brian. Policing cyber hate, cyber threats and cyber terrorism, 2012.

[31] L. Kendall. Meaning and identity in "cyberspace": The performance of gender, class, and race online. *Symbolic interaction*, 21(2):129–153, 1998.

[32] Y. Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[33] D. Lowe. Christchurch terrorist attack, the far-right and social media: What can we learn? *The New Jurist*, 2019.

[34] T. M. Massaro. Equality and freedom of expression: The hate speech dilemma. *Wm. & Mary L. Rev.*, 32:211, 1990.

[35] L. Nakamura. 'i will do everything that am asked': Scambaiting, digital show-space, and the racial violence of social media. *Journal of Visual Culture*, 13(3):257–274, 2014.

[36] B. Pang and L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[37] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[38] B. Perry. Anti-muslim retaliatory violence following the 9/11 terrorist attacks. *Hate and bias crime: A reader*, pages 183–201, 2003.

[39] E. Riloff and J. Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 105–112, 2003.

[40] E. Riloff, R. Jones, et al. Learning dictionaries for information extraction by multi-level bootstrapping. In *AAAI/IAAI*, pages 474–479, 1999.

[41] S. Schelter and J. Kunegis. 'dark germany' temporal characteristics and connectivity patterns in online far-right protests against refugee housing. In *Proceedings of the 2017 ACM on Web Science Conference*, pages 415–416, 2017.

[42] S. Schelter, F. Biessmann, M. Zobel, and N. Teneva. Structural patterns in the rise of germany's new right on facebook. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 440–445. IEEE, 2016.

[43] T. Shepherd, A. Harvey, T. Jordan, S. Srauy, and K. Miltner. Histories of hating. *Social Media+ Society*, 1(2):2056305115603997, 2015.

[44] L. Silva, M. Mondal, D. Correa, F. Benevenuto, and I. Weber. Analyzing the targets of hate in online social media. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[45] A. A. Stiftung and P. Asyl. Chronik flüchtlingsfeindlicher vorfälle. *Antonio Amadeu Stiftung und Pro Asyl. url: http://mut-gegen-rechte-gewalt. de/service/chronikvorfaelle (besucht am 05. 01. 2016)*, 2015.

[46] M. Taboada, C. Anthony, and K. D. Voll. Methods for creating semantic orientation dictionaries. In *LREC*, pages 427–432, 2006.

[47] M. Thelen and E. Riloff. A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 214–221. Association for Computational Linguistics, 2002.

[48] I.-H. Ting, S.-L. Wang, H.-M. Chi, and J.-S. Wu. Content matters: A study of hate groups detection based on social networks analysis and web mining. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 1196–1201, 2013.

[49] J. Van Dijck. *The culture of connectivity: A critical history of social media*. Oxford University Press, 2013.

[50] M. Walker and A. Troianovski. Behind angela merkel's open door for migrants. *Wall Street Journal*, 9, 2015.

[51] Z. Waseem and D. Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[52] H. S. Watch. Hate crimes: Consequences of hate speech, 2016.

[53] M. L. Williams and P. Burnap. Cyberhate on social media in the aftermath of woolwich: A case study in computational criminology and big data. *British Journal of Criminology*, 56(2):211–238, 2016.

[54] T. Wilson, J. Wiebe, and R. Hwa. Just how mad are you? finding strong and weak opinion clauses. In *aaai*, volume 4, pages 761–769, 2004.

[55] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354, 2005.

Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Festus Fortune Ikechukwu**,

1. Herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Discriminatory Speech on digital platform a case study of Twitter (Race, Gender, Politics, Sexuality),

   Supervised by Rajesh Sharma and Christian Simon Ritter.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Fortune Ikechukwu Festus
12/11/2020