5-2022

Detection of Hateful Comments on Social Media

Essa AlZarouni
ema9905@rit.edu

DETECTION OF HATEFUL COMMENTS ON SOCIAL MEDIA

by

Essa AlZarouni

A Capstone Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Professional Studies: Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

May 2022

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name**: Essa AlZarouni**

Graduate Capstone Title**: Detection of Hateful Comments on Social Media**

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of committee

Name: Dr. Ehsan Warriach **Date:**

Member of committee

Acknowledgments

I would like to show my gratitude and appreciation to my Mother, Father, brothers, and sisters for their unwavering and unconditional support during this whole journey. I would also like to thank a friend and a brother – Saeed AlMarri - who supported me from the first day, as well, all of my extended family and friends. I would like to thank Dr. Ehsan Warriach, who was a bit tough on me, but I know it was always for my own good. Dr. Ehsan's continuous support during the whole capstone project process will never be forgotten. I would like to thank Dr. Sanjay Modak, Chair of Graduate Programs and Research at RIT Dubai, for his advice, guidance, and feedback during the whole program.

Abstract

Social media usage has grown tremendously in the contemporary communication landscape. Along with its numerous benefits, some users abuse the channels by spreading hatred, far from the intended purpose of building connections on a personal level. To date, an empirical method for detecting, quantifying, and categorizing hateful comments on social networks comprehensively and proactively is still lacking. Besides, majority of the cases remain unreported due to social confounders such as fear of victimization and the psychological implications of hateful comments, leading to a situation whereby, the detrimental effect of the situation is underestimated. The ill-defined situation in the growing online space impedes progress towards developing mechanisms and policies to mitigate the harmful effects of hate on social media, ultimately reducing the effectiveness of the platforms as effective communication tools. This proposal suggests Naïve Bayes classifier as a novel approach for detecting and classifying hateful social media comments to bridge this gap. Data set was taken from set provided by Kaggle and consisted of 30,000 Tweets. From the results of the use of this method, it was calculated that Bayes method is 62.75% accurate, which is not satisfactory. However, to bridge accuracy gap, nural algorithm was used which gain an improved accuracy of 87%.

Key Words: Social Networking, Connecting individuals, Twitter, Hateful Speech, Twitter hateful comments

Table of Contents

List of Figures

Chapter 1: Introduction

1.1 Background of the Problem

Online social networking trend has grown tremendously in the recent years, leading to emergence of numerous platforms. Among the networks, Twitter has arguably become the most popular microblogging platform, enabling users to share life experiences in real-time. Trends on the channel have created immense impact on communities, among them creating awareness on issues of global importance. For example, the "BlackLivesMatter" hashtag that appeared on Twitter almost 11.8 million times brought to the surface the underlying issue of racism across the globe (Anderson, 2016). Trends initiated in the channel generate widespread attention globally, whereby, users are prompted to express their opinions on specified subjects. According to available statistics, there were more than 211 million active Twitter users towards the end of 2021, with the primary reason marketers prefer investing in Twitter spaces identified as the need to promote their products directly to prospective consumers (Everson et al., 2013; Van den Brakel et al., 2017; Vilenchik, 2019). The 2021 social media usage data further indicate that Twitter ranks favorably among the top 15 world's most popular and active social media platforms. Owing to its massive popularity, opinion leaders use Twitter to express their views and share information about oncoming, ongoing, and past events.

Nonetheless, Twitter is among the most misused social media platforms. People abuse the platform to spread spiteful content that often causes irreparable harm to the subjects. It is against Twitter policy to spread violent threats, victimize others, and express hate. The company exists to promote free expression as a fundamental human right and expressly prohibits against hateful conduct. It does not allow accounts that incite harm towards other on the basis of race, ethnicity, and gender orientation among other divisions (Twitter, 2022). Despite its commitment to combat abuse motivated by hatred, many users still overlook the rules and use their accounts to incite violence and make hurtful comments on others. Among the main reasons why purging hate speech remains a challenge to the company include the vagueness of metrics used to measure what amounts to hate in posted content.

There is no standard definition of hateful speech, but the term has been used to refer to content that suggests hatred against an individual or community. The problem not only affects Twitter, but has been experienced in the entire social media space. Hateful speech has grown into

a big problem for almost all online social media platforms where user-generated content occurs (Salminen et al., 2020). Such can be in the comments section beneath a post, real chat sessions, or forum discussions, among others. Hateful content can isolate users, back radicalization, and provoke violence among communities (Al-Hassan & Al-Dossari, 2019). Hateful comments are any form of expression that instigate hostility and violence, both verbal and physical, against a person or community in society. Notably, the definition of hateful comments implied here does not include all instances of aggressive language. This study recognizes that people, including those on Twitter, use terms that are highly belligerent to others but in qualitatively fitter discourse. However, tweets have a broad reach, and the effects can be extensively diverse and severe.

The focus of this task is to detect, identify, and classify hateful comments on Twitter through sentiment analysis of real-time Twitter comments data based on Naïve Bayes classifiers. Sentiment analysis is a natural language processing approach to examining opinions and attitudes toward an event or user expressed through writing (Freund & Shah, 1994). Sentiment analysis reveals the feeling and attitude behind one's thoughts or expression about an entity, such as an event, another person, or institution.

There are already developed tools for analyzing sentiments from textual data. The three most commonly used sentiment analysis methods are full-text machine learning, lexicon search methods, and linguistic analysis (Rasel et al., 2018). Full-text machine learning approaches to sentiment analysis rely on a human-pre-coded polarity library to train the algorithm to detect features that fall under positive, negative, negative sentiment categories. Lexicon searches begin with a set of words already pre-categorized into polarity groups and sometimes for the strength of the sentiment. A linguistic-based sentiment analysis approach examines the grammatical structure of the text for polarity classification.

This study proposes an approach to sentiment analysis of Twitter comments based on Naïve Bayes classification. The goal is to detect hateful comments in Twitter data.

1.2 Statement of the Problem

The digital space has grown immensely over the past few decades, and the most significant phenomenon that has characterized this growth is the expansion of the social media space. The available data suggests more than half of the world's population uses social media, making it a part of any brand's primary marketing platform. Beyond marketing and brand building, social media

has created a space for people to communicate, express opinions, and share content. Content uploaded on social media attracts criticism from different user quotas, some of which constitute hate speech. Hateful comments on social media are a growing challenge in the online space and a primary concern for social media developers, marketers, and law enforcement authorities. The consequences of hateful comments and discourse on social media include public tension and provoking violence (Iacus et al., 2020). Hateful comments have also resulted in psychological and emotional upheavals among users (Al-Hassan & Al-Dossari, 2019). Despite being a growing concern, an empirical method for comprehensively and proactively detecting, quantifying, and characterizing hateful comments on social networks is still lacking. Besides, social confounds such as fear of being a subject and the psychological implications of hateful comments lead to underestimating the consequences of social media hatefulness (Rasel et al., 2018). This technical gap in the growing online space has slowed down the mitigation efforts for hate on social media, reducing the effectiveness of the platforms in building good brands and use as communication tools. This proposal proposes a novel approach for detecting and classifying hateful social media comments using a Naïve Bayes classifier to bridge this gap. The proposed approach will acquire data from twitter using Tweepy API. User comments streamed through Tweepy API will undergo a rigorous preprocessing stage to remove unwanted speech dimensions using n-grams. The cleaned tweets and comments dataset will proceed for cataloging and evaluation using Naïve Bayes. The study recognizes that not all instances of offensive language are hateful because people express themselves in different ways. In some cases, people use highly offensive terms to others but in a qualitatively healthier discourse. With the consideration above, the streaming of tweets ensured that the dataset contained a sufficient diversity of words, phrasings, and emotional signals for the model to learn. The expected outcome is a sentiment analysis classifier that streams real-time tweets and comments and applies supervised learning based on Naive Bayes to detect and classify hateful speech.

1.3 Aims and Objectives

The focus of the study is to develop a novel approach for detecting hateful Twitter comments based on the above definition. The study has two specific objectives:

1. To detect hate speech on Twitter comments.
2. To get the identities of the users who make hateful comments on Twitter.

Chapter 2: Literature Review

2.1 Introduction

The literature on hate speech has expanded in recent decades, and so has the definition for the concept. Numerous studies have been dedicated to identify, understand and analyze hate speech in social media circles. More particularly, there is a growing body of literature that focuses on assessing how hate speech has been projected across various social media platforms. Hate speech is a major research phenomenon in social media-related literature. Kwok & Wang (2013) proposed an approach to detect hateful comments against blacks on Twitter. The study gathered a large corpus of tweets and made a keyword analysis of the sentiments suggestive of hatred against blacks. The study judged the severity of the hatred based on the opinions of students of different races gathered using a standard questionnaire. A dataset of tweets and retweets was split into two; one used for training the model and the other for testing. The researchers then applied the model on the test set to classify the tweets as racist or non-racist. Besides doing the classification, prominent features were identified from the tweets with an accuracy score of 76% and average error score of 24%. Whereas a system set to screen posts using specific words like negro may capture hate conduct, language may have some other dynamics that cannot be captured comprehensively. For example, some phrases may imply hatred towards an individual or group without necessarily containing a clear loathing word. Overall, the study established that a bag-of-words model is insufficient to accurately classify hate speech.

2.2 Defining Hate Speech

Hate speech is a contentious term owing to the varying definitions and use of the concept. The use of the term hate-speech changes across periods, places, and contexts (Mugambi, 2017). As a result, research on hate speech has produced different conclusions that cannot be generalized unless within the applied definition or application limit. Because Twitter is the chosen platform for the analysis, this proposal selects the definition of hateful speech as it relates to the platform.

There have been several European Commission initiatives in recent years to combat anti-Semitism. Efforts to combat anti-Semitic rhetoric are continuing. By legislation, the European Union has taken another approach to this problem. European Commission recently required Facebook, YouTube, Twitter, and Microsoft to sign the EU speech code to prevent hate speech by ensuring that most impactful posts are read 24 hours after publishing. According to the Council of

Europe's Committee of Ministers, the term hate speech covers all forms of expressions intended to spread, incite, promote, or justify hatred based on intolerance (Kiilu et al., 2018). Going with this definition, hateful Twitter comments are those directed to a person or a particular group to express intolerance for the individual or their ideology. Twitter policy for controlling hate speech states that users should not promote violence against or directly thrash or threaten other people based on their racial identities, gender, religious affiliation, age, disability, or health conditions.

2.3 Varied Forms of Hate Speech

Hate can be expressed in different forms: intended or unintended. While some words can expressly be used to disparage an individual or group, other people may use the same words innocently. Twitter users can express hatefulness in the form of aggressive antagonism with highly offensive and attributional terms in their tweets or comments (Vilenchik, 2019). On the flipside, a user may post something that is perceived differently by their followers and hence generate unexpected responses. A post that triggers hate sentiments, while in most cases follows the tone set by the owner, may in some instances take a different trajectory than the one intended. In his findings, Vilenchic (2019) finds that overall, the activity of the user is not correlated with the feedback that the user receives on that activity. Therefore, whereas simple statistics may be applied to track hate in social media data on instances where specific words are targeted, sometimes they cannot sufficiently be tracked.

2.4 Validity of Detection Methods for Official Statistics

Millions of user generated posts are generated in any single day. As the quantities of data accumulate, it increasingly becomes complex to classify and manipulate the data accurately according to a stated formula. Van den Brakel et al. (2017) applied a rule-based method for classifying antagonism on tweets and Twitter comments using associational terms as the classification features. Their study incorporated accusation and attributional phrases directed at a person or an entity following an interesting or socially disruptive event to empower their approach further. Consequently, their approach showed significant improvement in standard learning methods and procedures used in measuring the extent of hate. Thus, it is possible to evaluate the extent of hate in social media platforms with significant levels of precision.

Researchers have applied similar hate-speech detection efforts on social media platforms other than Twitter. In particular, Facebook has received significant research focus, targeting hate

speech discourse and user posts (Bianchini et al., 2018; Guo & Johnson, 2020; Haoxiang, 2020; Iacus et al., 2020; Kalsnes & Ihlebæk, 2021; Leonhard et al., 2018; Meza et al., 2018; Miškolci et al., 2020; Rodriguez et al., 2022, 2019; Vilenchik, 2019). These studies have deployed different mechanisms to categorize Facebook content as hate speech with different contextualization's. For example, Rodriguez et al. (2022) developed an architecture for detecting hate groups on Facebook using text mining analysis. The mechanism proposed in the study extracted features such as the frequently used keywords within a group of Facebook users.

The different forms of content for various social media platforms may present challenges for developing a foolproof method for detecting hatred sentiments. Whereas screening the text-based microblogging sites may have fairly straightforward criteria, others based on different forms of media, for example videos, may be challenging to detect. Döring & Mohseni (2020) developed a framework for detecting atrociousness and radicalism on YouTube videos. Their approach exploited philological, syntactic, and content-based features from the user-generated data through videos and reactions to posts and deployed various classification algorithms to categorize the contents. Their findings indicate that video-based posts are perceived more intensely than other forms of content. Thus, their effect should not only be based on their absolute values, but also on their practical implications.

2.5 Justification of the Current Model

The framework proposed in the present proposal is different from the above work. It is noticeable in the above literature review that most of the approaches proposed previously concerned a single domain, such as an oncoming, live, or immediate past event. Previous works are intensely thematic, for example, targeting religion, race, or ethnicity-based hatred. The proposed approach deviates from this monotony by focusing on generic issues that are generalizable to different forms of hate. The approach also differs from previous work by the use of real-time Twitter data and user account profiles. The advantage of real-time analytics is that they have a proactive advantage in preventing fateful comments from trending. Besides, an approach incorporating different features, such as unigrams and bigrams, is unique.

2.6 Summary of Literature Review

While many of the studies in the literature review look at the content posted by people on Twitter, other social media platforms such as Facebook are also very relevant. Each of the studies looks at varying models to analyze what type of content would be considered as hate speech and contextualization of these types of posts. In some cases, the content that is being analyzed is clearly expressed through very offensive language, while in other cases it is using associational terms that qualify as hate speech. In order to be able to gather such large volumes of information and to distinguish for analysis, there are different frameworks and mechanisms that have been developed depending on the purpose of the study and what is the scope of analysis. There are different ways that large volumes of data could be gathered from social media platforms to analyze the trends of what constitutes as hate speech and, more importantly, how prevalent they are. One method is done through text mining analysis, which could detect and extract key words that would be used by Facebook users. However, the large volume of data online is not only text-based, posing an even more challenging task of identifying and extracting such information from video platforms through the comments and reaction posts of the public within the algorithm that helps in the content identification and categorization. The unique approach taken by the author is not to address the obvious types of hate speech, which is mostly towards ethnic or religious topics, but rather a more generalized approach on the hate form through real-time data tracking.

Researchers have applied similar hate-speech detection efforts on social media platforms other than Twitter. The study gathered a large corpus of tweets and made a keyword analysis of the sentiments suggestive of hatred against blacks. Hate speech is a major research phenomenon in social media-related literature. The approach also differs from previous work by the use of real-time Twitter data and user account profiles. Besides, an approach incorporating different features, such as unigrams and bigrams, is unique. Proposed an approach to detect hateful comments against blacks on Twitter. The mechanism proposed in the study extracted features such as the frequently used keywords within a group of Facebook users.

- Hate speech takes different forms on social media platforms
- Extracting and analyzing meaningful data requires a robust framework
- Getting real-time data tracking allows for early detection of hate speech and plans of hate groups.

Chapter 3: Project Description

3.1 Research Methodology

The framework for detecting and classifying hateful comments and associating the inciters' profiles proposed in this study will follow a 3-stage process. The first stage will involve the Tweepy API in streaming and collecting Twitter comments to create a dataset. The dataset with Twitter comments will proceed for preprocessing in the second stage, ensuring readiness for mining and feature extraction. A variety of data cleansing procedures will be applied to the dataset containing Twitter comments. Formatting elements, such as punctuation, hyperlinks, and white spaces will be removed as part of the cleaning process. Using TF-IDF vectorizer, training features will be extracted and the data transformed into an array.

The third stage will involve training classifiers on a training subset of the preprocessed dataset and applying the trained model on the test data to classify the Twitter comments as positive or negative. Python's Scikit-learn module will be used to split the data into training, and testing sets based on the train-test-split approach. The algorithm will be called and applied to the training data. Models are trained using training data, and then predictions are made based on a variety of metrics, including accuracy (f1 measure), confusion matrix (precision), and recall (recall) on the testing set. We store the model using the pickle library at this point so that it may be used again in the future. The testing data stored in a separate file will be used to produce classifications and test model performance.

The objective of the classification stage is to categorize the comments into two classes; one contains the subject of the comment or tweet, and the other describes the sentiment contained in the tweet as hateful or normal. The Scikit-learn provides powerful support for multinomial Naïve Bayes classifiers; hence, it is the preferred classification approach. It also has a variety of tools, such as classification, clustering, regression, and visualization algorithms, that can be used to improve the model. Figure 1 shows the architecture of the proposed identification and classification approach. After completing the classification, the results will be retained in a text file. A model-specific tool will be developed and applied to determine the percentage of positive and negative comments in the file and visualize the results.

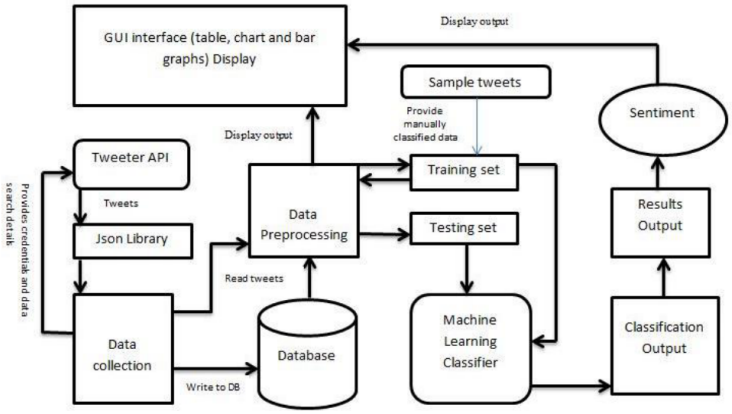


Figure 1: The architecture of the proposed identification and classification approach.

3.2 Procedure

Then write a deep learning model using Keras that has several tense layers. The first is the embedding layer which converts the text into a vector. Then those vector goes into the neural network translators, and the prediction comes. So this plot, the class balance plot, shows how many the number of instances for each positive, negative and neutral class. Then we generated a plot with Sea bourn that shows the frequency of words, and after that, we made a graph that shows the number of instances in each class. So what this function calculates is the similarity between words, the string of it as words plotted on a 2D XY plane. Then the distance is calculated between each word. Deep learning-based Embedding thinks of it as a predictive model that encodes your text into vectors. Then we've used the word cloud package to plot a word count graph attached to the word file given.

So, we imported all the dependencies, then the preprocessing stage to first load up the data and drop the column ID from it because it's an index column and it's not useful for us. A confusion matrix is just a matrix that shows the actual outputs and compares the predicted output. It calculates the relative frequency of each word in our text and makes a model out of it. Another technique we use in deep learning is a convolution technique used with Max pooling. This function is called Count Vector sentiment model optimized. Then we lemmatize the text (lemmatize means to reduce the words). Then we split our data into training input and outputs and testing input and outputs. So we predicted according to this, plotted confusion metrics, and saved different scores.

3.3 Data Sources

The proposed methodology will employ the Tweepy API to stream tweets and comments from twitter. Tweepy is a convenient Python-based way of accessing the Twitter API. To collect tweets and the corresponding comments will require a data collection script that implements Tweepy. The following are the steps needed to collect Twitter data using Tweepy.

The first step is to obtain a secure connection to the Twitter API. We achieve this by providing a consumer API key and a consumer API secret. The two are available while using a developer profile with Twitter. Application for access to a developer profile is through https://developer.twitter.com/en/apply-for-access and we supply it to Python using the following steps:

```
consumer_api_key=os.environ["TWITTER_CONSUMER_API_KEY"]
consumer_api_secret=os.environ["TWITTER_CONSUMER_API_SECRET"]
auth = tw.OAuthHandler(consumer_api_key, consumer_api_secret)
```

The authentication token gotten from the above procedure will help to initiate the API call using the procedure below. This will require specifying to wait on rate limit in order to stream larger volumes of tweets surpassing the limitation on rate.

```
api = tw.API(auth, wait_on_rate_limit=True)
```

We will download the tweets using the Tweepy cursor that we will have specified via a search string using Twitter API specification. This step will allows us to specify both the language and timeframe of the tweets and comments that we intend to analyze. The downloaded data will be saved locally as a CSV file.

```
tweets=tw.Cursor(api.search,q=search_words,lang="en",
since=date_since).items(RATE_LIMIT)
```

An alternative to the above procedure is to choose one of the Twitter datasets published at https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset and analyze it following the methodology suggested in the proposal. This is a less tiresome way and overcomes the challenges of being denied a Twitter developer profile and the required access to tweets to stream a dataset.

Chapter 4: Data Analysis

4.1 Results and Analysis

The study applies secondary qualitative method for evaluating the data. The data has been collected through different social media platforms and evaluated for propagating hateful sentiments on social media. Social media users can express their hate, give the opinions a public dimension, receive applause from friends and followers, and feel somehow validated. Furthermore, the line between a virtual threat and a criminal action is minor. Thus, clear rules on what amounts to hate speech should be dissociated from free expression and the threshold for becoming a threat established. The algorithms used for analyzing and detecting hateful comments are as follows:

Accuracy

Accuracy is a metric used to determine the percentage of correct predictions. The metric is calculated by dividing the number of correct predictions by the total number of predictions

F1 Score

The F1-score combines the precision and recall of a classifier into a single metric by taking their harmonic mean

Algo 1 - TfIdf

Short for term frequency–inverse document frequency, Algo 1 is a numerical statistic that highlights how important a word is to a document in a collection or corpus.[1] It is often used as a weighting factor in searches of information retrieval, text mining, and user modeling. The term frequency–inverse document frequency (tf–idf) value increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word, which helps to adjust for the fact that some words appear more frequently in general

Algo 2 - Neural network

A neural network is a network or circuit of biological neurons, or, in a modern sense, an artificial neural network, composed of artificial neurons or nodes.[1] Thus, a neural network is either a biological neural network, made up of biological neurons, or an artificial neural network, used for solving artificial intelligence (AI) problems. The connections of the biological neuron are

modeled in artificial neural networks as weights between nodes. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed. This activity is referred to as a linear combination. Finally, an activation function controls the amplitude of the output. For example, an acceptable range of output is usually between 0 and 1, or it could be −1 and 1.

4.2 Algorithm used

```
def vectorize(doc):
    features = defaultdict(int)
    for token in tokenize(doc):
        features[token] += 1
    return features
vectors = map(vectorize, corpus)
```

To classify dataset, the following algorithm was proven useful:

```
def separate_by_class(dataset):
        separated = dict()
        for i in range(len(dataset)):
                vector = dataset[i]
                class_value = vector[-1]
                if (class_value not in separated):
                        separated[class_value] = list()
                separated[class_value].append(vector)
        return separated
```

4.3 Tweepy API in Streaming

Social networks have brought a paradigm shift in relation to the way people communicate. They allow users to express their opinions “freely”, without any kind of direct human contact. This opens up gaps for the emergence of hate speech on the internet. Hate speech

refers to any comment that attacks an individual/group with respect to their race, gender, ethnicity, national origin, religion, sexual orientation or other discriminatory aspect. On social networks that allow a certain degree of anonymity, such as Twitter, this problem can be exacerbated. It has been noted that few scientific contributions have been made to counter this problem in languages.

4.4 Wordcloud

This work aims to employ text processing and machine learning techniques to classify hate speech data. The objective was to explore different classification algorithms in a dataset formed by tweets (Rodriguez, Argueta & Chen, 2019). The results were evaluated through statistical metrics, and through them, comparisons were made between the results obtained and the main approaches considered as state-of-the-art. In relation to this, word cloud is prepared showing visual representation of text data, which is often used to depict keyword metadata on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color.

Figure 2: Wordcloud

In contrast to hate comments on social media, word cloud is made on opposite which specified that most common words includes love, happy new people and birthday along with other

joyful events. In the month of May, demonstrations took place in various parts of the world against racism and hate speech. There is currently a great deal of debate about the fine line between freedom of expression and hate speech. The first is fundamental for a democracy to exist; it represents intolerant and empathetic speech. Therefore, there is a need to understand what characterizes hate speech and how harmful it can be to a democratic society (Rodriguez, Argueta, & Chen, 2019). In this text researcher will see some views of scholars on this concept, examples of hate speech and its repercussions and, finally, ways to combat this practice, both on the internet and in the real world.

4.5 Corpus of Tweet Data

There is no single definition for hate speech; however, they are all similar. Hate speech is the manifestation of ideas that incite racial, social or religious discrimination in certain groups, most often minorities. However, from this aspect it addresses points of racial, social or religious discrimination, without considering, for example, gender, sexual orientation, weight, some type of disability, class, among others. Additionally, hate speech can be characterized by manifestations of hatred, contempt or intolerance against certain groups, motivated by prejudice (Mugambi, 2017). Therefore, based on these two concepts and on the common sense that exists about the term, we can conclude that hate speech is a set of actions with an intolerant content aimed at groups, most of the time, social minorities (such as women, LGBTs, fat people, people with disabilities, immigrants, among others).

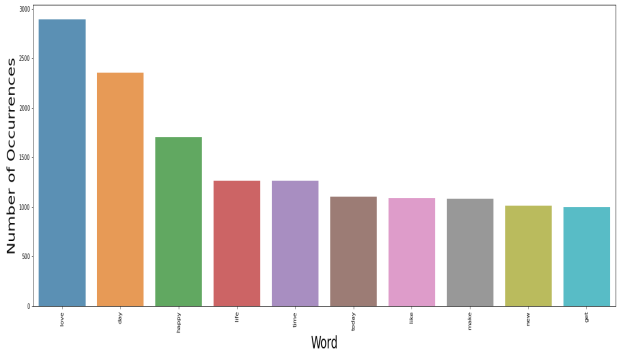


Figure 3: Corpus of Tweet Data

It is found that corpus of tweet data specified number of occurrences The internet has changed the way to communicate. Just having a cell phone is enough for individuals to be able to express opinion and comment on the infinite subjects and contents spread out there. But not

everything is perfect. Ease also served evil. Hate speech and prejudice gained strength, mainly spread by fake profiles. That said, Haters and hate speech: understanding violence on social networking sites (Miškolci, Kováčová & Rigová, 2020). The work seeks to understand the relationship between hate speech and the creation of the profile of haters (or haters, in the translation of the word) and their speeches on social networks, based on the theoretical concepts of violence and how hate is placed in the virtual. In order to end this, they analyzed the speech of a text from the blog and comments made in a Facebook publication about the journalist.

4.6 Labels of Tweet

The outcomes showed the environment of social networks as a facilitating tool and capable of potentiating the violence of haters, with the dissemination of hatred and the propagation of ideologies that constitute the discourse. In this way, the design of profiles remains active and results in a group of biased and prejudiced followers. An internet application monitor posts on social networks that reproduce messages of hate, racism, intolerance and that promote violence (Meza, Vincze, & Mogos, 2018). The instrument will be launched this month and will allow users to be identified and reported. According to the professor responsible for the project, human rights are viewed in a pejorative way on the internet and hate speech has gained momentum. It is necessary to dismantle this process. By making the data available, he believes it is possible to create public policies that support and empower the victims. Commissioned by the Ministry of Women, Racial Equality and Human Rights, the Human Rights Monitor, as the application was named, will seek -keys in conversations that encourage sexual violence against women, racism and discrimination against blacks, Indians, immigrants, gays, lesbians, transvestites and transsexuals. In the figure below which is showing the labels of tweets, 0 denotes tweets that are not hateful, 1 denotes tweets that are hateful.

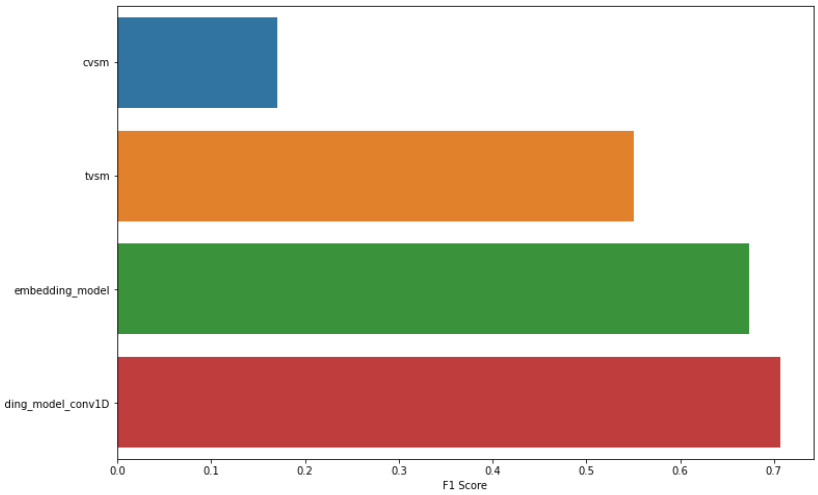


Figure 4: Labels of Tweet

In recent cases, one is different from the other. One thing is the beating of a street vendor who tried to help transvestites, from populations particularly exposed to collective violence - here it really is a hate crime, of hate for difference. They are almost always crimes inspired by the horror and fear of being able to identify with the victim - the feeling that "I kill the homeless person or the transvestite that could become and in such a way this will never become that same person. It is the fundamental basis of many racist behaviors, of extermination of different people. This is one type of mechanism of violence, but another is the case with the nightclub in Istanbul, for example, which is the desire to destroy the place where Westerners gather for their infidel parties because I don't want to be tempted by that and kill my own temptation to go crazy.

On the other hand, someone who didn't recognize herself in her body, a trans woman who lived in the interior and thought she was a monster, unique of her kind and destined for a hidden life, suddenly discovers that there are people like her at around the world, and groups, and people willing to listen, to give advice. This is the other positive effect of networks (Kwok & Wang, 2013). Now it is true that social networks are fundamentally built on the model of contemporary society, that is, individuals are worth the appreciation which they produce. Or in this case, the number of likes to posts is able to receive. This would happen even if social networks did not exist. That is, in contemporary society, individual does not worth their diplomas or even what their history is - what matters is who and how many like. This is how contemporary society works,

whether we like it or not. Now, the problem is, when person live, they feed off of the appreciation of others, it's very easy to get tangled up in absolutely amazing group formations.

4.7 Tweet Label Counts

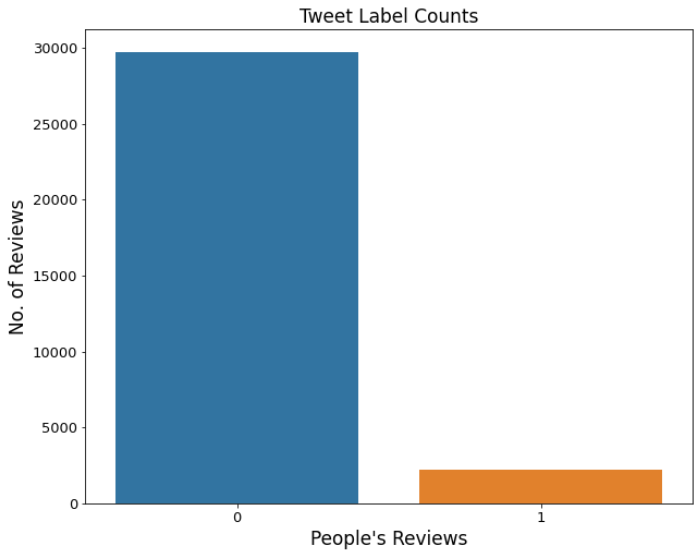


Figure 5: Reviews

The Internet, and especially social networks, have enabled greater interaction between people from different parts of the world. After all, we currently have numerous communication platforms at our disposal, such as Facebook, Instagram, Twitter and instant messaging applications such as WhatsApp, which allow us to make new friends and even have a romantic relationship. However, these services have been constantly used for the propagation of violence and hate speech that, in general, occurs anonymously (Kwok & Wang, 2013). The Internet has provided a transformation in humanity that, if on the one hand it is positive, has also been the scene of intense battles regarding prejudice and discourses of discrimination.

Chapter 5: Conclusion

5.1 Accuracy of Naïve Bayes Classifier

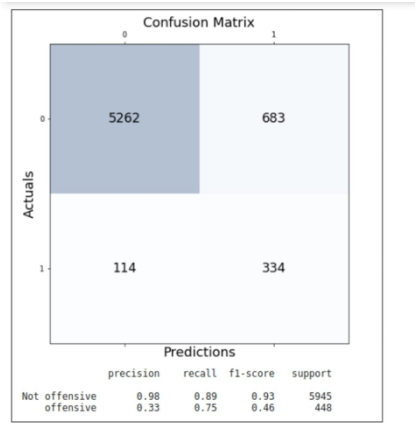
Naïve Bayes Classifier	
Accuracy	0.6275
95% CI	0.613, 0.657
Positive Class	1

5.2 Interpretation of the Naïve Bayes Classifier Indicator

Accuracy was calculated based on a 95% confidence level; accuracy is determined as the number of words identified in the positive class. Positive class is the review that falls within hateful comments. Based on this level of confidence there is a 95% confidence that 61.3% to 65.7% of hateful speech will be correctly identified. However, it can be argued that the 62.75% accuracy is not satisfactory. Therefore, it is not an accurate method.

5.3 Accuracy of Neural Algorithm & Confusion Matrix

In the aim of bridging the gap in the accuracy; a neural algorithm was employed and from the results a confusion matrix was constructed.



$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{5262 + 334}{5262 + 683 + 334 + 114} = \frac{5596}{6393} = 0.87 = 87\%$$

Figure 6: Confusion Matrix

From the matrix above, the level of accuracy in the predicted hate comments category is 87%. Therefore, this methodology could be consistent enough for users such as administrators or government officials to reasonably determine hate speech online.

5.4 Conclusion

Social Media, like any other space or tool, can be used to express good and bad sentiments. Because it is a huge space, many people believe that the internet is a lawless space where they are allowed to act in the way they want, without being faced with the consequences. That's why it's still common to see intolerant comments on social media. People are careful not to express prejudice and aggressive opinions in real life, away from the screen of computers and smartphones, mainly out of fear of consequences, in the virtual world these hateful behaviors seem to be released. For Cyber Crime Specialists, the Internet and, especially, social networks facilitated bullying and hate speech, which can be practiced at any time, with the aggressor having a false sense of impunity (Kwok & Wang, 2013). Victims of hate speech delivered over the Internet can prove their accusation in court with screenshots that contain the offensive content. However, for these prints to serve as evidence, it is important that they contain the offensive content, in addition to the respective posting dates and the link. Furthermore, it is ideal to list all prints in a single document that contains essential information. If the screenshots referred to messages from collective chats, it is important to take screenshots of the list of members of the group, as well as the contact of those who made the offensive publications. However, for both cases, it may be necessary to request the Notarial Minutes at a notary's office so that one can prove the veracity and non-tampering of the content that is being used for probative purposes. After all, in a judicial process, there is no possibility of contesting the notarial act.

As previously mentioned, there is a need for a specific law that defines the crime of hate speech and the applicable penalties. In addition, social networks, online games, forums and the internet as a whole also need to be active in the fight against this crime. For this, be sure to report posts and profiles with this type of speech. However, there is still much to be done in terms of awareness. This is because many people do not even know what hate speech is, which can cause this practice to be reproduced without even knowing how serious it really is. Therefore, awareness actions are essential (Meza, Vincze, & Mogos, 2018). From lectures to dynamics in schools and work environments it is the dissemination of this type of content on the internet. In addition, the social networks have regulatory mechanisms for sensitive content, which must be activated by users when they come across any publication of intolerant and disrespectful content. Thus, even though there is no specific law, it does not mean that a person who commits a hate crime on the internet can go unpunished.

By defining that no one will be subjected to inhuman or degrading treatment, as well as that the law will punish any discrimination that violates fundamental rights and freedoms, the Constitution defends human rights and punishes those who violate them, that is, those who practice hate speech. Hate speech occurs because of these singularities (origin and gender identity/sexual orientation), as if they demean the individual and make him less a human being than someone who is not in one of these classifications. This, in turn, goes against hate speech, which preaches prejudice against human beings who are part of some social minority. In other words, hate speech violates the guarantees and fundamental rights of each and every citizen .As stated earlier, hate speech is configured as a crime and attentive to the guarantees and fundamental rights of every citizen. Hate speech is considered a type of verbal violence, and its basis is the non-acceptance of differences and intolerance. However, the main debate that arises when we talk about this practice is the difference between hate speech and freedom of expression. This is because, many claim that freedom of expression gives them the right to express themselves in the way that best suits them on any and all topics.

5.5 Recommendations for Future Work

The study sought to establish an empirical method for comprehensively detecting, quantifying, and categorizing hateful comments on social networks. Whereas the study proposes an objective criterion that can be used to detect hateful social media posts and comments, some gaps in the area limit effective application of strategies that can contain the vice. The weaknesses and limitations of each of the tools and techniques developed in the research study have indicated a need to build the body of knowledge in the area as recommendations for further work. Better tools with healthier validity and accuracy levels should be developed. As earlier noted, scientific contributions made to specifically counter the problem of hate-speech in social media circles are generally few. In most cases, scientists will brush off the problem by citing the need to grant users freedom of speech. However, the problem presents a window of opportunity for researchers to build models that can have practical social impact on the global scale.

References

1 Al-Hassan, A., & Al-Dossari, H. (2019). *Detection of Hate Speech in Social Networks: a Survey on Multilingual Corpus*. *February*, 83–100. https://doi.org/10.5121/csit.2019.90208

2 Anderson, M. (2016, Aug. 15). The hashtag #BlackLivesMatter emerges: Social activism on Twitter. *Pew Research Center.* https://www.pewresearch.org/internet/2016/08/15/the-hashtag-blacklivesmatter-emerges-social-activism-on-twitter/

3 Bianchini, G., nzo Ferri, L., & Giorni, T. (2018). Text analysis for hate speech detection in Italian messages on Twitter and Facebook. *EVALITA Evaluation of NLP and Speech Tools for Italian*, *12*, 250.

4 Döring, N., & Mohseni, M. R. (2020). Gendered hate speech in YouTube and YouNow comments: Results of two content analyses. *SCM Studies in Communication and Media*, *9*(1), 62–88.

5 Everson, M., Gundlach, E., & Miller, J. (2013). Social media and the introductory statistics course. *Computers in Human Behavior*, *29*(5), A69–A81.

6 Freund, L. E., & Shah, K. M. (1994). A Survey of Simulation Courses in Industrial Engineering Programmes. *European Journal of Engineering Education*, *19*(1), 93–103. https://doi.org/10.1080/03043799408923274

7 Guo, L., & Johnson, B. G. (2020). Third-person effect and hate speech censorship on Facebook. *Social Media+ Society*, *6*(2), 2056305120923003.

8 Haoxiang, W. (2020). Emotional analysis of bogus statistics in social media. *Journal of Ubiquitous Computing and Communication Technologies (UCCT)*, *2*(03), 178–186.

9 Iacus, S. M., Porro, G., Salini, S., & Siletti, E. (2020). Controlling for selection bias in social media indicators through official statistics: A proposal. *Journal of Official Statistics*,

36(2), 315–338.

10 Kalsnes, B., & Ihlebæk, K. A. (2021). Hiding hate speech: political moderation on Facebook. *Media, Culture & Society*, *43*(2), 326–342.

11 Kiilu, K. K., Okeyo, G., Rimiru, R., & Ogada, K. (2018). Using Naïve Bayes Algorithm in the detection of Hate Tweets. *International Journal of Scientific and Research Publications (IJSRP)*, *8*(3). https://doi.org/10.29322/ijsrp.8.3.2018.p7517

12 Kwok, I., & Wang, Y. (2013). Locate the Hate: Detecting Tweets against Blacks. *AAAI*.

13 Leonhard, L., Rueß, C., Obermaier, M., & Reinemann, C. (2018). Perceiving threat and feeling responsible. How severity of hate speech, number of bystanders, and prior reactions of others affect bystanders' intention to counterargue against hate speech on Facebook. *SCM Studies in Communication and Media*, *7*(4), 555–579.

14 Meza, R. M., Vincze, H.-O., & Mogos, A. (2018). Targets of Online Hate Speech in Context: A Comparative Digital Social Science Analysis of Comments on Public Facebook Pages from Romania and Hungary. *Intersections. East European Journal of Society and Politics*, *4*(4).

15 Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering hate speech on Facebook: The case of the Roma minority in Slovakia. *Social Science Computer Review*, *38*(2), 128–146.

16 Mugambi, S. K. (2017). *Sentiment analysis for hate speech detection on social media : TF-IDF weighted N-Grams based approach*.

17 Rasel, R. I., Sultana, N., Akhter, S., & Meesad, P. (2018). Detection of Cyber-Aggressive Comments on Social Media Networks: A Machine Learning and Text Mining Approach. *Proceedings of the 2nd International Conference on Natural Language Processing and Information Retrieval*, 37–41. https://doi.org/10.1145/3278293.3278303

18 Rodriguez, A., Argueta, C., & Chen, Y.-L. (2019). Automatic detection of hate speech on Facebook using sentiment and emotion analysis. *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 169–174.

19 Rodriguez, A., Chen, Y.-L., & Argueta, C. (2022). FADOHS: Framework for Detection and Integration of Unstructured Data of Hate Speech on Facebook Using Sentiment and Emotion Analysis. *IEEE Access*.

20 Salminen, J., Hopf, M., Chowdhury, S. A., Jung, S. gyo, Almerekhi, H., & Jansen, B. J. (2020). Developing an online hate classifier for multiple social media platforms. *Human-Centric Computing and Information Sciences*, *10*(1), 1–34. https://doi.org/10.1186/s13673-019-0205-6

21 Twitter. (2022). Hateful conduct policy. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy\

22 Van den Brakel, J., Söhler, E., Daas, P., & Buelens, B. (2017). Social media as a data source for official statistics; the Dutch Consumer Confidence Index. *Survey Methodology*, *43*(2), 183–210.

23 Vilenchik, D. (2019). Simple statistics are sometimes too simple: A case study in social media data. *IEEE Transactions on Knowledge and Data Engineering*, *32*(2), 402–408.