Thesis Report

Thesis Title: Novel Approach to Detect Hate Speech and Profanity on Online Platforms

Submitted By:

1. Samin Islam - 18101444
2. Barha Meherun Pritha - 18101232
3. Tabassum Alam - 18101235

Supervisor:

Md Rayhan Kabir, Lecturer
Department of Computer Science and Engineering
BRAC University

Co-supervisor:

Nazmus Sakeef, Lecturer
Department of Computer Science and Engineering
BRAC University

Novel Approach to Detect Hate Speech and Profanity on Online Platforms

Samin Islam, Barha Meherun Pritha, Tabassum Alam
Md Rayhan Kabir, Nazmus Sakeef

Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Abstract. Hate speech is becoming more prominent and dominant in the virtual world, with the popularity of social media increasing day by day. People nowadays have various online platforms where they can express their hatred and write offensive speech in the safety of their home. They could even spread false rumors and incite hatred out of nothing. Cyberbullies often verbally attack the sentiments of people with different race, nationality, gender, beliefs and political views. They could also target young children and teenagers. It is also important to note that profane language or some sensitive topic may be bothersome when reached in front of young children and teenagers. It has become necessary for modern technology to detect all those profane and hate speeches so that they can be filtered or removed automatically before they can appear in front of young children or hurt the sentiments of targeted people. However, even though it is easy to detect profanities, it could be difficult to detect all the hate speeches which do not have any offensive or sensitive keywords. It is possible to spot all sorts of hate speeches on social media through the application of machine learning, neural networks and natural language processing. In our study, to identify and recognize hate speeches we will use various models and algorithms. Then we will design and implement an algorithm which will be able to detect hate speech and profane language more efficiently.

1 Introduction

Hate speech and abusive language has been ever present and persistent to humankind for the longest time. It was there from the earliest times of slavery; it is here today with the advancement of modern technology, and it will be also be present in the undefined future. Profanity has also been present for a long time, but it is becoming more prominent. Profanities are now used increasingly in movies, pop songs and rap music.

The concept of hate speech differs. Although there is no standardized concept of hate speech, many people claim that they can identify "hate speech" when they see it [1]. The term hate speech is determined as per the UN as any sort of conversation, conductivity or writing that targets slur or excessive profanity with respect to an individual or a group based on who they are, or in other words, on the basis of their nationality, origin, faith, ancestry, sex or other identification features [2]. All have the right to opinion and speech, as stated in Article 19 of the United Nations, 1984 and The United General Assembly, 1966, which also includes the right to freedom of opinion without interruption and to search, obtain and impart knowledge and skills through any platform without being dependent on boundaries. It demonstrates that even though hate speech is seen as an act of discrimination, it is not considered illegal and is protected by the fundamental

human rights. Although it had been widely discussed in the legal field and with context to offensive speech on school campuses, the first amendment of the U.S. constitution also protects the right to free speech, religious belief and the news media. Fortunately, there are regulations banning hate speech against ethnic minorities in countries like the United Kingdom, Canada and France and those accused of using hate speech will also face substantial penalties and even incarceration [3].

The definition of profanity has changed through time. Before profanity referred to showing disrespect to religious values. Now, which is used in our research, profanity means offensive language, cursing or swearing.

2 Motivation

Usage of the internet and other technologies is increasing tremendously every day. Through that the practice of using social media is also accumulating. Nowadays, most of the people who have access to the internet use a number of online platforms including Twitter, Facebook, Instagram and YouTube. With the aid of such platforms, people can leave hateful or offensive speech through posts, comments, chats, etc. from the safety of their homes. It has become a challenge for networking sites to establish a platform where no hurtful, contempt or profane language can take place. Online media are establishing and updating their policies regarding what kind of content or comments can be shared, but even then, it seems that not all hate speech can be filtered out. Even though human beings can identify hate speech through words, tones and sentence structure, it is a bit difficult for artificial intelligence to evaluate if anybody encourages hatred or just explains what has happened to them [4]. It is important to note that the tone of sarcastic speech can be similar to the tone of hate speech. Both Facebook and Twitter have faced a lot of criticism for not doing sufficient to stop hate speech on their sites. The Chief Executive Officer and Co-founder of Facebook, Mark Zuckerberg once stated that hate speech has no space on Facebook [5].

The theme of offensive speech is becoming high in recent times yet further progress has been made to discover how to better recognise and discriminate between hate speech and other normal speech. We can trace profanity or receptive conversation via the use of clear and unambiguous keywords but not necessarily hate speech. It is pretty much guaranteed that racial and homophobic course will be categorised as hate speech, while sexist speech is typically categorised as profane [3]. Words like “fuck” can be implied profane but it may not be used to convey hate speech. Furthermore, certain words like “gay” can be considered sensitive, but then again, it can also be used in a positive manner. Besides, a hate speech may not contain any offensive keywords at all; in which case it will become difficult to correctly identify them.

There are a number of methods for sentiment analysis and due to the wide availability of diverse viewpoints on the data of social media, this is indeed a tough challenge, and

these methods mostly rely on lexicons. Lexical methods in detection tasks have been a common feature for explicitly identifying any predefined word, and this feature extraction alone can only be good for text classification with low accuracy.

Hateful contents on the internet can be diverse like the dataset in [13]. That might seem like a challenge, which is somewhat true but not impossible to do at all. Moreover, there may arise a question regarding the dataset, e.g. for high efficiency, what amount of training data is required? According to [14] the authors set a different bar to experiment the accuracy of the model and came up with the conclusion that the more recent data is preferable than larger dataset. Performance degrades over time if the model is not updated with new features; and based on three experiments on a temporal dataset it has been proven delicately in [14].

3 Literature Review

There have been numerous studies, experiments, and surveys conducted about hate speech and profane language over the past decade. Several models and algorithms have been implemented in order to detect abusive language including natural language processing, neural networks, machine learning, lexical analysis and sentiment analysis.

The research of [6] was one of the first to talk about the abusive language on the web. In their experiment, they used n-gram, sentiment analysis and contextual features to determine the offensiveness of previous sentences, to determine if the sentence is abusive or not.

In one of the researches [7], they explained why detecting hate speech is difficult to find. The reasons they stated are:

i. It is not possible to detect hate speech using keyword spotting.
ii. All ethnic and cultural slurs may be difficult to define, since any meaning that is offensive to one community might be perfectly fine for some other communities.
iii. Hate speech can have grammatical or spelling errors, and it can also have no grammatical or spelling errors.
iv. Abusiveness can be beyond the sentence boundaries.
v. Sarcastic speeches can have the same tone as hate speech.

According to the paper [8], they determined that a speech is considered abusive if it i) a sexist or racial insult is used ii) assaults, critiques or threats to censor a minority iii) endorses criminal offence or hate speech iv) deliberately pursue to misrepresent the truth on a minority; v) supporting controversial hashtags, e.g. "#BanIslam", "#whoriental", "#whitegenocide"; vi) protects misogyny or xenophobia vii) stereotyping a minority unfavourably.

In the research [9], they used profane words to identify hate speech. Here are some of the offensive keywords and the category of the targeted discrimination that they found in their research:

i. Sexual Orientation: gay, lesbian, faggot
ii. Physical dysfunction: douchebag, fucktard, dumbfuck, shithead
iii. Gender: cunt, bitch, pussy, dick, cock, bull
iv. Religious belief: jesus, islam, god king
v. Ethnic group: sandnigger, nigga, nigger
vi. Class: bastard, sucker, fucker, motherfucker.

There are numerous methods for text classification, and according to any specific task data need to be flagged based on the right context. A survey paper [10] provides a succinct description of the automated identification of hate speech that thoroughly discusses the latest methods, concentrating on extraction of features in particular. Different features have been taken under consideration while working on this task whether the approach should be predictive or not, for instance Simple Surface Feature, Word Generalization, Sentiment Analysis, Linguistic Features, Knowledge based features, Meta information, Multi-modal information (mentioned in this paper). Although the set of traits studied in the various work differs widely, the approaches of classification relies largely on supervised learning. Character-level methods perform different leading approaches than approaches at the token level. Classification can be supported by lexical tools such as a list of slurs, but normally only in conjunction with other types of functionalities. Different dynamic traits using too much linguistic skills, sarcastic statements, non textual contents (multi-modal features can be considered) are also cue suggesting the presence of hate speech. To perform experiments, a decent amount of dataset is needed from social medias like-Twitter, Facebook, Instagram, Yahoo, YouTube, ask.fm. They seem to have unique features, because these platforms have indeed been built for particular reasons and may also exhibit multiple subtypes of hate speech. When evaluating the usefulness of such functions or techniques added to them, the scale of a dataset must be taken into account.

The classic methods of ML alone cannot track down precisely all kinds of offensive speech for ambiguity. Feeling the necessity to establish accurate and automated models to identify abusive language online; the authors of [14] came forward with new features of some powerful NPL, text classification and task specified embeddings. The methodologies they have followed includes dependency parser, semantic distributional features, linguistic features and N-gram traits. They handled tactfully the noisy data; 80% of the dataset was trained and 20% tested, combining with all features and individual features in different models. The model with all features outperforms the model with a single feature. Character n-grams have the greatest contribution in terms of human characteristics and thus their future studies include extraction of comment thread and use it as a background to evaluate each comment.

Similarly, in [7] another fresh strategy to deep learning architecture to detect hate speech has been created in tweets using word embedding. The dataset of 16K tweets have been labeled as sexist, racist and neither sexist or racist. The authors analyzed both classic and deep learning methods; different semantic tweet embeddings and three neural network architectures - FastText, CNNs and LSTMs. The results were categorized into three parts as per the combination of the methods. The results of Part A are baseline methods while Part B involves approaches that use neural networks only and Part C incorporates Deep Neural Networks learned average word embeddings as features for Gradient Boosted Decision Trees. On average Part C has the best performance among all three of them, then Part C and lastly the baseline approaches.

Researchers on [14] targeted internet protection of adolescents and came up with the idea of filtering out the contents by parents or teachers before appearing on a web browser. Using lexical and parser features they proposed an approach which is noxious expression recognition on YouTube comments. Utilizing Support Vector Machines which includes features like automatically generated blacklists, n-grams, manually created regular expression and dependency parsing features, the analysis creates a supervised classification approach and through this they gain 98.24% precision success on the role of inflammatory sentence identification and 94.34% recall.

The distinction seen between [7] work and this one is that before extracting the feature they try to spell correct and stabilize noisy text. But authors in [7] found noise a theoretically strong harassment detection signal and therefore has functionality to catch multiple formed up noise. Both of the works have dependency features, [7] consisting of a far larger collection of tuples than [14].

Another paper [8] identifies the difference between racist and sexist slurs using a character n-gram based approach. The model has been trained on some extra linguistic features; such as gender, religion, location and length which is the highlight of this research. The authors searched all probable feature set groupings. They found that the word n-grams performed better than n-grams by 5 F1 scores in minimum. The problem faced during this research is when location, gender and length are trained altogether, the performance diminishes and due to lack of coverage demographic information, apart from gender, brings little improvement. But overall the authors could successfully present a model to identify racist and sexist slurs.

The conceptual framework proposed in this paper characterize conditions between sentiment analysis and other NLP tasks, and express the dependencies in first order logic rules which aims at exploiting information outside the document to improve sentiment analysis. The authors have focused on two types of knowledge which includes intra document knowledge and extra document knowledge. The external knowledge defined in this paper exploits knowledge outside the sentence and outside the document. Not only that,

the framework of this paper allows two types of evidence against the rules. The first case is when the event is involuntarily conducted and the second case is when an event is accidental. In spite of the fact that this paper is a conceptual framework, it bridges together various jobs of sentiment analysis and numerous jobs in natural language processing to deliver a complete tactic to sentiment analysis and others [15].

The paper [13] proposes an unused challenge on detecting hate speech in multimodal memes with hateful memes as a dataset. It did not train models from scrape. It adjusted and verified large scale multimodal models that were previously trained. The authors reconstructed basis memes from scrape by means of a customized tool. They had third-party annotators, who consumed about 27 minutes for each subsequent meme in the dataset. The memes are reconstructed using Getty images which allows several benefits like avoiding potential noise from optical character recognition (OCR) and reducing all errors that could be present in the graphic modality. The paper got some potential downsides too. Better multi-modal systems, for example, could lead to job automation in the future and be exploited for censorship or other undesirable ends. These dangers can be minimized in part by building AI systems to counteract them [13].

4 Dataset Preparation

We have collected the data using the Python library tweepy. We have used the OAuthHandler submodule to get authentication from Twitter. Later, we used the Stream submodule to filter the tweets using keywords and then store them in a file. We have fetched tweets from random people in the feed using the keywords. We had collected about 45k tweets. The process of data collection from Twitter is shown in **Fig. 1**.

The keywords we used to filter the tweets are: 'food', 'election', 'media', 'competition', 'vlog', 'travel', 'USA', 'US', 'economy', 'politics', 'programming', 'social', 'climate', 'game', 'tournament', 'movie', 'culture', 'torture', 'trump', 'biden', 'show', 'finance', 'stories', 'marketing', 'media', 'twitter', 'facebook', 'research', 'disaster', 'weather', 'life', 'motivation', 'fitness', 'science', 'goals', 'technology', 'festival', 'concert', 'song', 'review', 'hate', 'love', 'romance', 'beautiful', 'scenario', 'place', 'football', 'cricket', 'computer', 'religion', 'feminism', 'job', 'study', 'worst', 'shut', 'racism', 'kill', 'slang', 'gun', 'murder', 'suicide', 'racist', 'shit'.

Tweets are usually unstructured types of data. To work with the dataset efficiently aiming for a better accuracy, we needed to clean the dataset first as shown in **Fig. 2**. We have removed non-English tweets, urls, emojis, mentions, punctuations, numbers and all other special characters. Then we have removed all the retweets and have tokenized each word of the tweets. Then we have tagged them separately on the basis of - i) if they contain hate speech or not, and ii) if they contain any profanity or not. After cleaning and removing all the duplicate tweets, we remained with about 1.5k dataset.

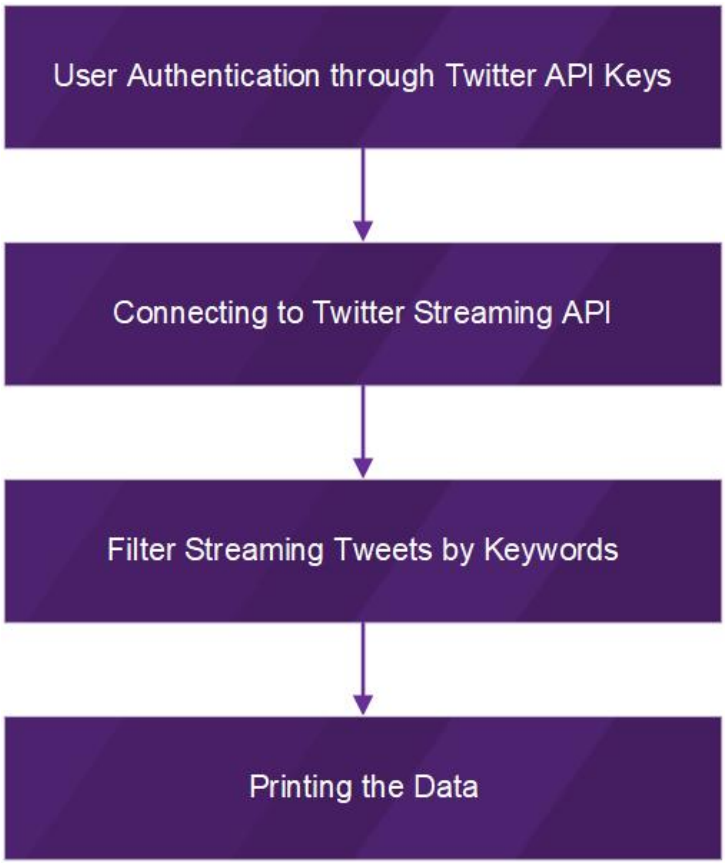


Fig. 1. The Process of our Data Collection

5 Data Description

Categories	**Labels**
Both (Hatespeech + Profanity)	Bully, Racism, Belief, Politics, Sexual, Self Hatred and Criticism
Hate Speech	Bully, Racism, Belief, Politics, Sexual, Self Hatred and Criticism
Profanity	Bully, Racism, Belief, Politics, Sexual, Self Hatred and Criticism
None	Thoughts, Grateful, Inspiring, Praising and Wishes

Table 1. Main Categorizations and their Types

We have divided our dataset into 4 different categories: Both (Hatespeech + Profanity), Hatespeech, Profanity and None. The tweets which had no profanities and cannot be considered a hatespeech is categorized as "None". The tweets that contained profane language, but cannot be considered as a hatespeech, are categorized into "Profanity". The tweets that can be considered a hatespeech but did not contain any profanity are categorized as "Hatespeech". Lastly, the tweets which had both hatespeech and profanity are categorized as "Both".

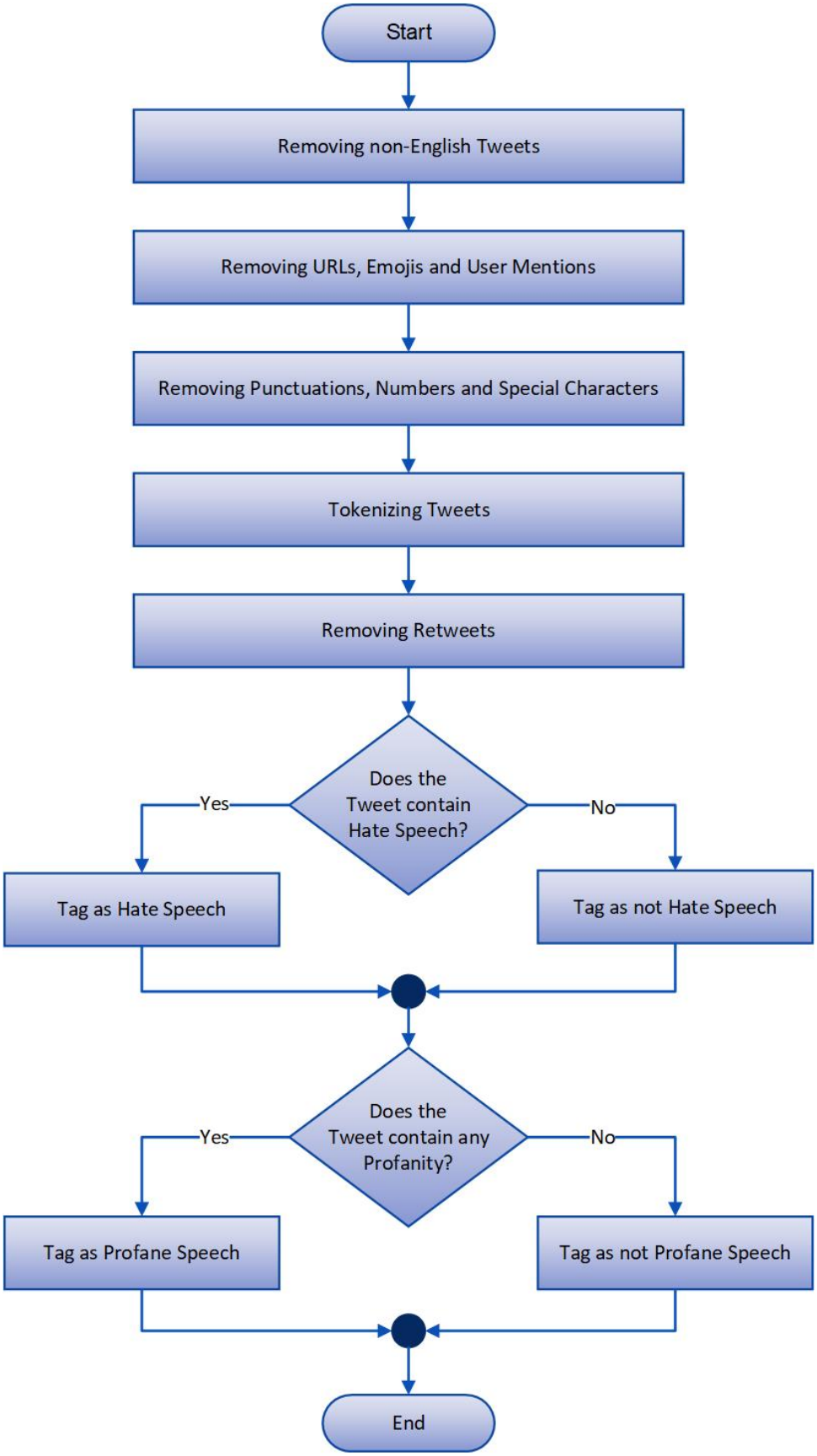


Fig. 2. Dataset Cleaning and Tagging

Both (Hatespeech + Profanity), Hate Speech and Profanity are further divided and labelled according to their types of discrimination: Bully, Racism, Belief, Politics, Sexual, Self Hatred and Criticism. None (tweets without any hatespeech or profanity) is divided into Thoughts, Grateful, Inspiring, Praising and Wishes. This is shown in the **Table 1**.

Fig. 3 illustrates the different Word Clouds generated from Tweets of different categories. They were coded with the help of python libraries in kaggle.

Word Cloud 1. Both

Word Cloud 2: Hate Speech

Word Cloud 3: Profanity

Word Cloud 4: None

Fig. 3. Word Clouds of different Categories

In **Fig. 4**, the bar chart depicts the proportion of tweets that contain hate speech depending on the label manually assorted to determine if a tweet depicts "self hatred", "beliefs", "sexual", "racist", "politics", "bully" or "criticism". Almost half of the total tweets i.e. 42.20% of them portrays criticism. Negativity towards politics and bullying also consists of 49.54% of the total tweets. A few proportion of around 8% of the tweets depicts self-hatred, hatred towards beliefs, sexual and racism.

In **Fig. 5**, the bar chart shows the proportion of tweets that contain profanity words based on the mark that was manually sorted to decide whether a tweet depicted "self hatred", "beliefs", "sexual", "racist", "politics", "bully" or "criticism". More than half of the total tweets i.e. 57.58% of them portrays criticism towards people. Sexual and bully tweets covered 22.10% with just 1.9% difference between the two. Racist tweets are just 1% more than that of self-hatred consuming tweets. And minorities of the tweets attack the beliefs and politics by 2.02% and 1.01% respectively.

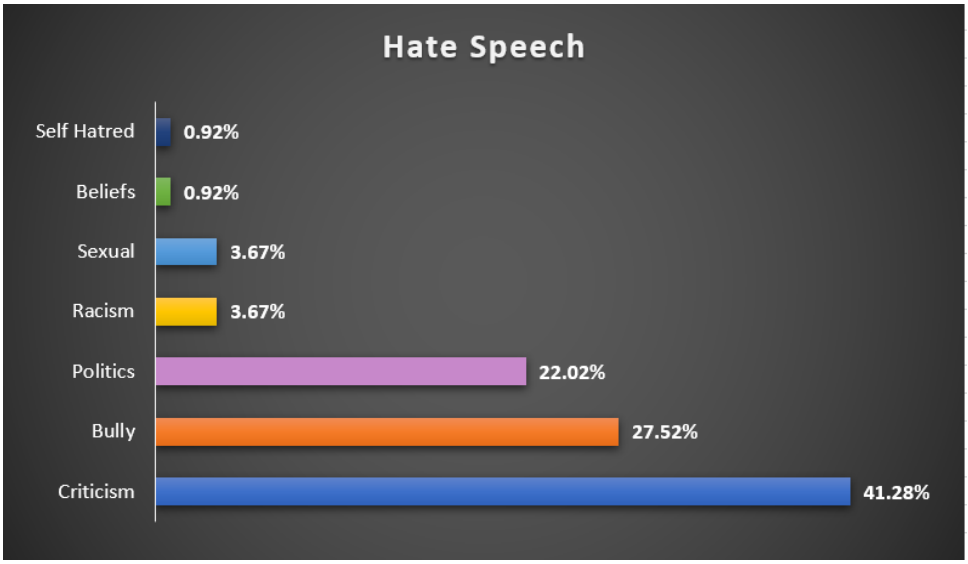


Fig. 4. Total Percentage of Tweets containing Hate Speech based on the Label

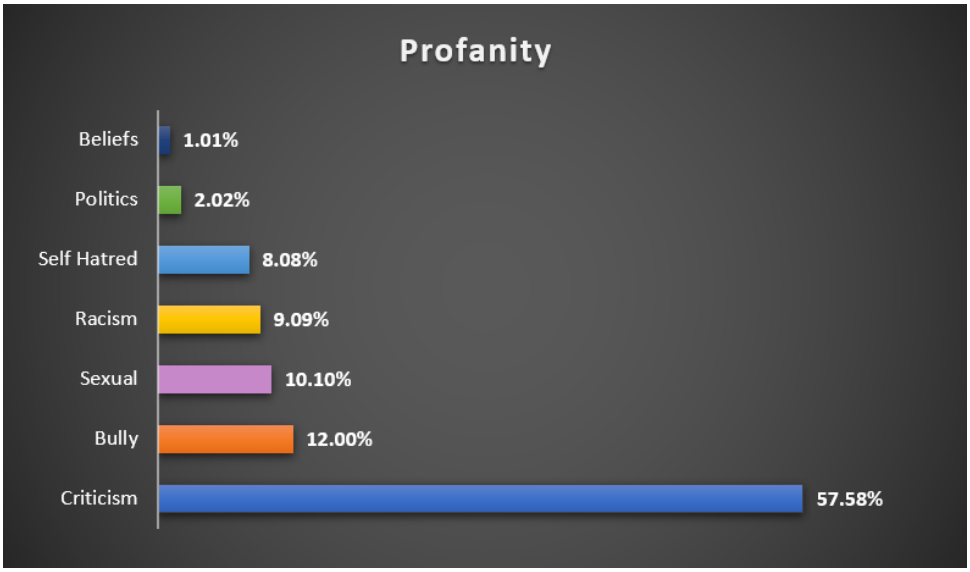


Fig. 5. Total Percentage of Tweets containing Profanity based on the Label

In **Fig. 6**, the bar chart depicts the proportion of tweets that contain both hate speech and profanity based on the label manually assorted to determine if a tweet depicts "self hatred", "beliefs", "sexual", "racist", "politics", "bully" or "criticism". Majority of the tweets containing slang and hatred are bullies i.e. 35.62% of the total. After that, most of the tweets are sexual covering 27.04% of the total tweets. One quarter of the tweets are criticism and attacking towards politics with 12.33% and 13.70% respectively. The percentage of racist tweets comes just after which is 8.22%. A small proportion of the tweets are portraying self-hatred and negativity towards beliefs in 1.37% each.

In **Fig. 7**, the bar chart illustrates the proportion of friendly tweets which do not contain any negative speeches based on the label manually assorted to determine if a tweet depicts "wishes", "grateful", "inspiring", "praising", and "thoughts". More than

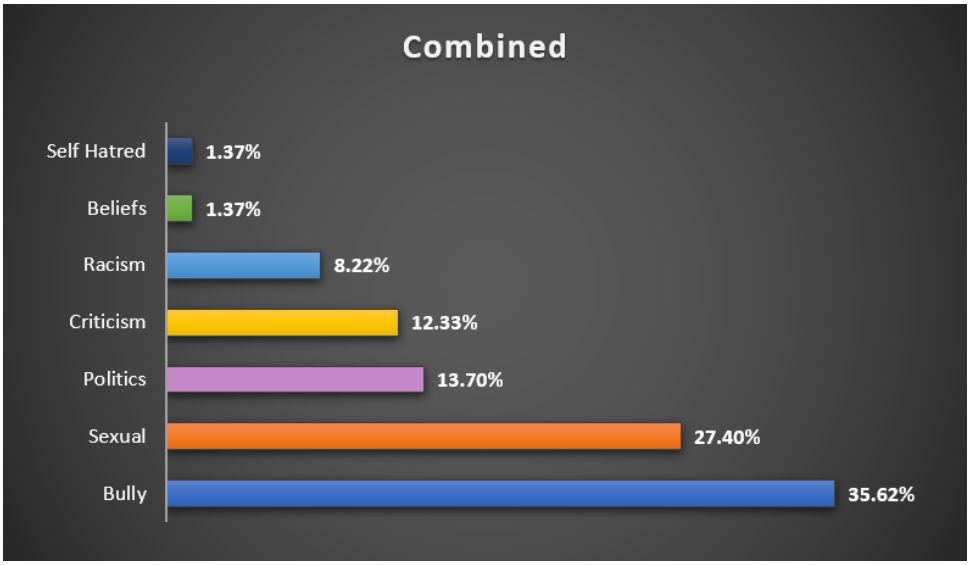


Fig. 6. Total Percentage of Tweets containing both Hate Speech and Profanity based on the Label

50% of the tweets i.e. 56.17% are the genuine thoughts of people. After that, gradually comes the percentage of praising, wishes, gratefulness and inspiration containing tweets. Almost 30% of the tweets contains praises and wishes and the rest of the 15% tweets consist of inspiration and gratefulness.

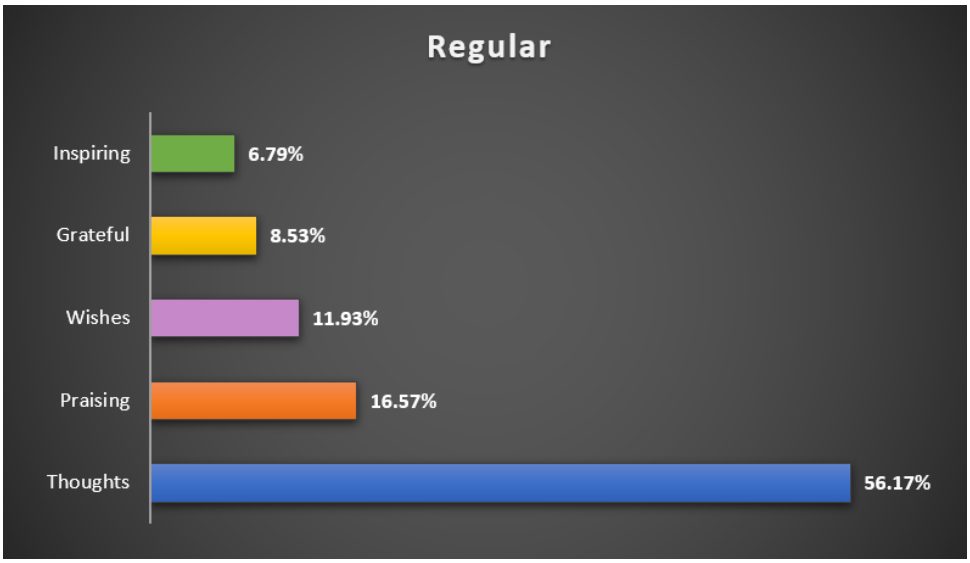


Fig. 7. Total Percentage of Friendly Tweets based on the Label

In **Fig. 8**, the bar chart depicts the proportion of the total tweets depending on the label manually assorted to determine if a tweet depicts "wishes", "grateful", "inspiring", "praising", "thoughts", "self hatred", "beliefs", "sexual", "racist", "politics", "bully" or "criticism". The green colored bars depicts the friendly tweets and the rest are either hate speech or profane tweets. Almost half of the total tweets i.e. 45.56% of them portrays

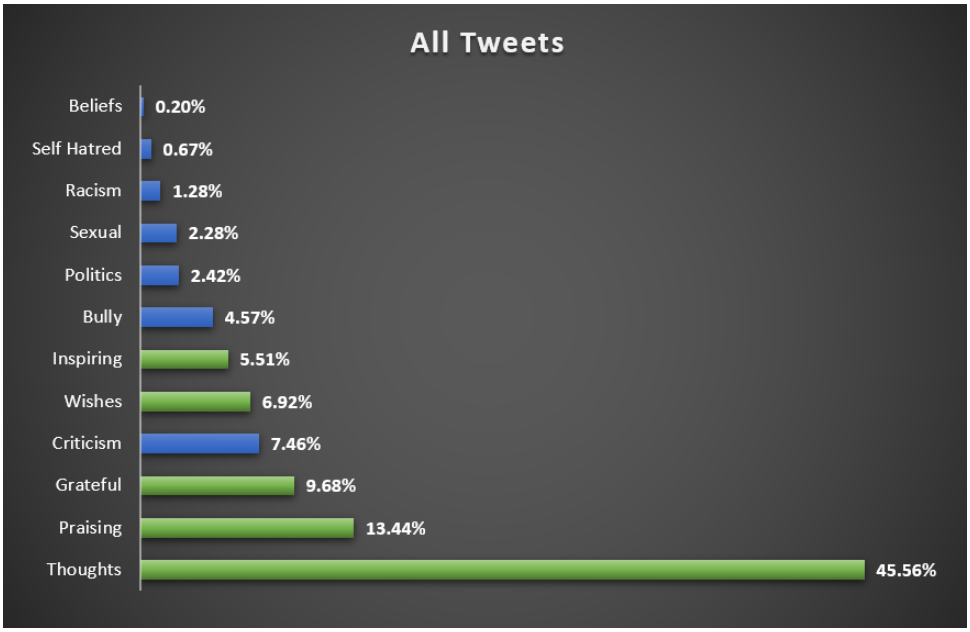


Fig. 8. Total percentage of Tweets based on the Labels

general thoughts of people. It is clearly illustrated that most of the tweets do not contain hate speech or profanity, a total of 81.15% tweets are friendly and non violating. On the other hand 18.85% tweets are hateful or profane tweets.

6 Data Inspection

Fig. 9 is the list of our last 10 tweets in our dataset. We printed the last 10 tweets (tweets no. 1488-1497) to check if our dataset has been read properly or not.

Our dataset is extracted from Twitter, thus almost every sentence contains hashtags. People use hashtags because it helps to link the tweet to other related tweets. Hashtag also increases the longevity of a conversation. **Fig. 11** represents the top hashtag counts in our dataset. We used a count value starting from 0 to 10. The most used hashtag is "pdx911" which is used to report an accident or any suspicious incident to Portland Police. The second highest tag is "SnackDown" which refers to a global programming competition that challenges the best programmers from across the world against each other. Then the 3rd most tagged hashtag is "WeirdoTrump". On 6th of December in the year 2020, President Donald Trump held his first post election rally where he mentioned about liking blueberries, cucumber, squash and other green plants. The users became irritated and some ashamed at the same time; that is from where the tag "WeirdoTrump" started. And the interesting fact is, this hashtag became the 2nd most used hashtag on Twitter within just 10 minutes! Then comes sequentially the hashtags "NFL", "Selena", "SelenaNetflix", "Selenatheseries" and "Toonami". The least tagged hashtags in top 10

	No	Tweets	Hatespeech	Profanity
1487	1488	i really put 2 & 2 together and be right. we not the same 🤣	0	0
1488	1489	Unprecedented aircraft movement in the last 10 days, 4 X the average, especially tonight, Sat night.🛩️\n\nMonkey Werks talks about lots of aircraft moving a lot of troops just outside of Las Vegas...	0	0
1489	1490	Proving again that drunk or sober Matt Gaetz is a flaming asshole.\n'You are not welcome in New Jersey': Governor slams Rep. Gaetz for attending maskless Republican gala https://t.co/9Fc8BYJ2uj vi...	1	1
1490	1491	Mount Washington,NH (MWN) ASOS reports gust of 73 knots (84.0 mph) from NW @ 0756Z -- KMWN 060756Z 32058G73KT 1/16SM SN BLSN FZFG VV001 M15/M15 RMK VRY LGT ICG	0	0
1491	1492	literally, I say this all the time. pickle all my shit https://t.co/zhgRZ06n1H	0	1
1492	1493	Finished my Cowboy Bebop re-watch. Last time was 10 years ago when I was living in Japan. Of course, still a masterpiece. And that final showdown... I want to play that. Such a beautiful use of mu...	0	0
1493	1494	Just posted a photo @ Stand Up Live - Phoenix https://t.co/9xHw7AMhbT	0	0
1494	1495	@RealCapnCrunch I love the crunch crunch crunch sound 😊 #CapnCrunchSweater #Sweepstakes	0	0
1495	1496	@1965_superfly @CodeMonkeyZ Audited by who? I assume you mean a financial audit but that wouldn't hold true. Goods in transit are included. Also, why did y'all have truckloads of goods that wer...	0	0
1496	1497	And outdoor #7footApart activities, like #7footApartHikes available to people, especially active Sr.s &\n50+ more at risk pop. who\nkeep health UP by being safe distance APART outside...low ri...	0	0

Fig. 9. Last 10 Tweets of our Dataset

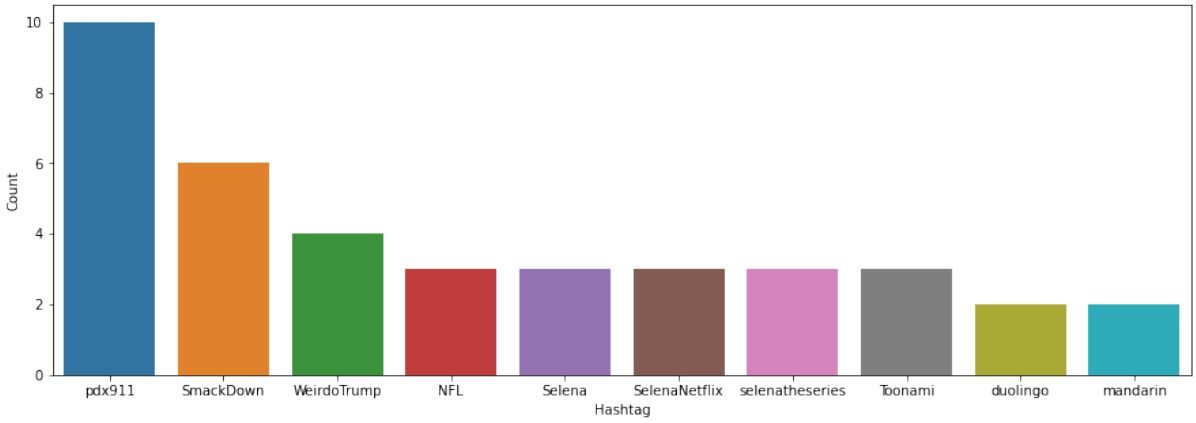


Fig. 10. Most frequent Hashtags used in our Dataset

of our dataset are "duolingo" and "Mandarin".

Fig. 11 represents the top 15 targets which are referring to either hate speech or profanity in our dataset. The sentence count is the number of sentences in our dataset containing the target words. The highest target is "trump" which is included in 23 sentences which represents either hate speech or profanity. The other target words are: "ass",

	sentence_count	target
0	23	trump
1	16	ass
2	13	man
3	13	people
4	11	bitch
5	11	thing
6	9	@realdonaldtrump
7	7	nigga
8	6	dem
9	5	biden
10	5	joe
11	5	stupid
12	4	short
13	4	living
14	4	hell

Fig. 11. Top Targets of Hatespeech and Profanity in our Dataset

“man”, “people”, “bitch”, “thing”, “nigga”, “@realdonaldtrump”, “dem”, “biden”, “joe”, “stupid”, “short”, ”living” and “hell”.

In **Fig. 12** we have shown the comparison between polarity and subjectivity with respect to sentence length and word count of our dataset. We have labeled them with 4 different colored dots assigning values ”Regular”, ”Profanity”, ”Hate Speech” and ”Combined”. The most important component of sentiment analysis is analyzing what kind of opinion a sentence expresses. There are two key properties of sentiment analysis: polarity and subjectivity. The sign of the polarity score is frequently used to infer whether the overall sentiment is positive, neutral, or negative. So, the value of polarity is a float value with range [-1, +1]; where -1 refers to negative sentiment and +1 refers to positive sentiment. Subjectivity refers to sentences which express personal feelings, emotions or judgments. Subjectivity is also a float value within the range of [0, 1].

7 Feature Engineering and Vectorization of Data

Presence of some words for rhetorical purposes are frequently used in texts. Words like ‘nobody’, ’everybody’ , ‘never’, ‘always’, and pronouns are used instead of the literal meanings or original subject. Thus it can be a bit difficult for the models to identify the target and meaning of those general statements. Hate based tweets can be targeted to a group of people where the sender could use these words for generalizing them.While training the models one needs to consider this problem a critical one and take necessary

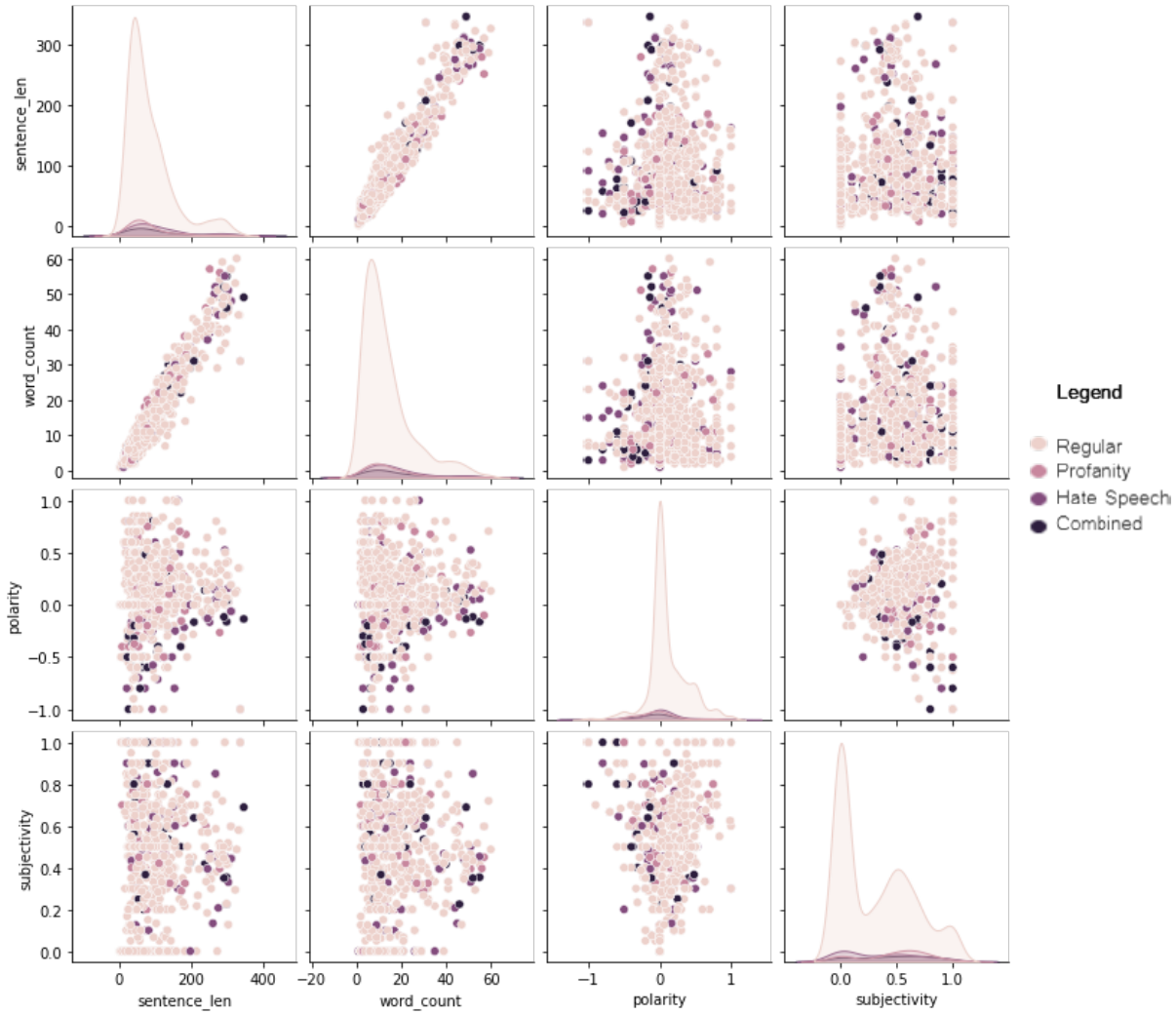


Fig. 12. Comparison of sentence length, word count, polarity and subjectivity of our Dataset

measures to avoid it.

Tweets are usually unstructured types of data. To work with the dataset efficiently aiming for a better accuracy, we needed to clean the collected dataset at first.

Vectorization of texts is an important part of feature extraction. Every unique word has a distinct meaning. To detect hate speech it is necessary to be able to identify unique words in the corpus and label them in the correct context.

We used some pre-deep learning approaches for this task as well as a word-embedding (Doc2Vec) method prior to the real task. After this feature-extraction phase, we trained different models of machine learning with our dataset.

- **Bag of words (BoW) -** This is the simplest vectorization method to convert texts into fixed length vectors depending on word frequency. A text is represented as a

bag of the word containing itself, overlooking the grammar. This feature generation method is highly considered for its simplicity, it lacks a bit for determining contexts of a sentence.

- **TF-IDF -** The term is an abbreviation for Term Frequency Inverse Document Frequency; another simple vectorization method. This is based on the previous method where word frequency is considered and also focusing on the relevance of words. TF (term frequency) is the ratio of a word's occurrence to the total number of words in a document (shown in equation 1).

$$\mathbf{TF(w, d)} = \frac{\textbf{Occurrences of w in document d}}{\textbf{Total number of words in document d}} \tag{1}$$

And IDF (inverse document frequency) measures the significance of a word. The prepositions or pronouns have little importance in a sentence but are used most frequently for grammatical purposes. IDF provides a solution of this particular problem using the following equation, where a word is w in N documents (shown in equation 2).

$$\mathbf{IDF(w, D)} = \frac{\textbf{Total Number of documents N in corpus D}}{\textbf{Total number of documents containing w}} \tag{2}$$

- **Sentiment Analysis -** Sentiment analysis is a technique to identify the emotional tone expressed in a text. For our research purpose, gauging the sentiment of the users in twitter is the prime task and for this sentiment analysis score has been regarded to detect hate speech. For calculating sentiment score, a polarity analysis is needed where words have been assigned to some scores; +1, -1 and 0 as positive, negative and neutral respectively. Then the sum of these scores of a sentence is the sentiment score. Example:

  i. "I didn't (-1) study for the course and got poor(-1) marks" : **score = -2**

  ii. "The day was so bright (+1) in the morning, I went for a walk and felt great (+1)" : **score= +2**

  iii. "I didn't (-1) get an A in the course but passed (+1) anyway" : **score = 0**

  Word embedding is the representation of texts in vector forms where words having similar meaning have close vector representations. Some popular word-embedding techniques namely; Word2Vec, Doc2Vec, GloVe etc.

- **Doc2Vec -** We used Doc2Vec as a word-embedding technique which is an unsupervised algorithm. The length of the vectors are always fixed for paragraphs, documents or texts. This is a generalized form of Word2Vec.

- **Combined Features -** This portion is the combination of the prior models to transform the words into numerical form.

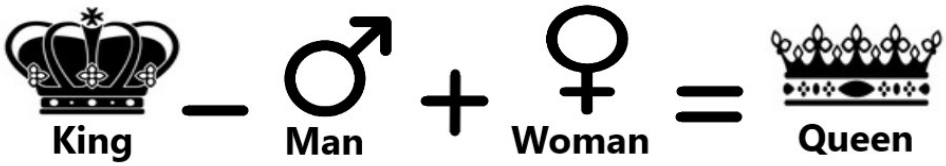


Fig. 13. Word Vector Representation of King - Man + Woman = Queen

8 Models

After vectorization of the corpus, the pre-processed tweets are ready to train different machine learning models for the detection. These models take inputs of embedding vectors and compress them into a lower dimensional representation. This representation effectively captures the information in the sequence of words from the numerical forms. We used some state of the art models; Logistic Regression, Random Forest, Naive Bayes and SVM.

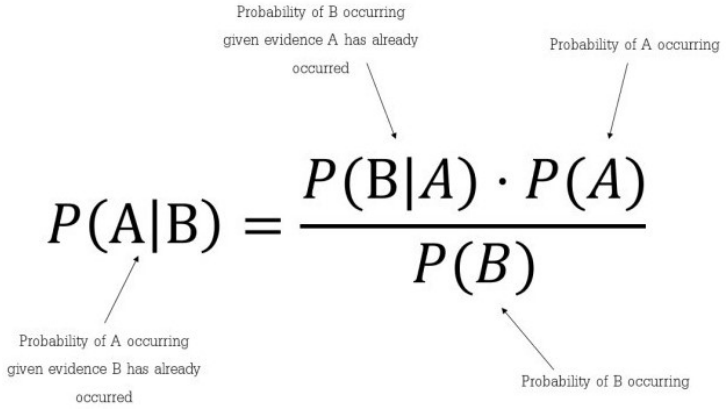


Fig. 14. Naive Bayes formula

- **Naive Bayes classifier -** As the name suggests, these classifiers are based on the principle of Naive Bayes' theorem. The theorem in the classifier states that the presence of one feature in a class is totally independent from the presence of any other feature. This approach is convenient for a large corpus and has always been outperformed by other sophisticated classification models.

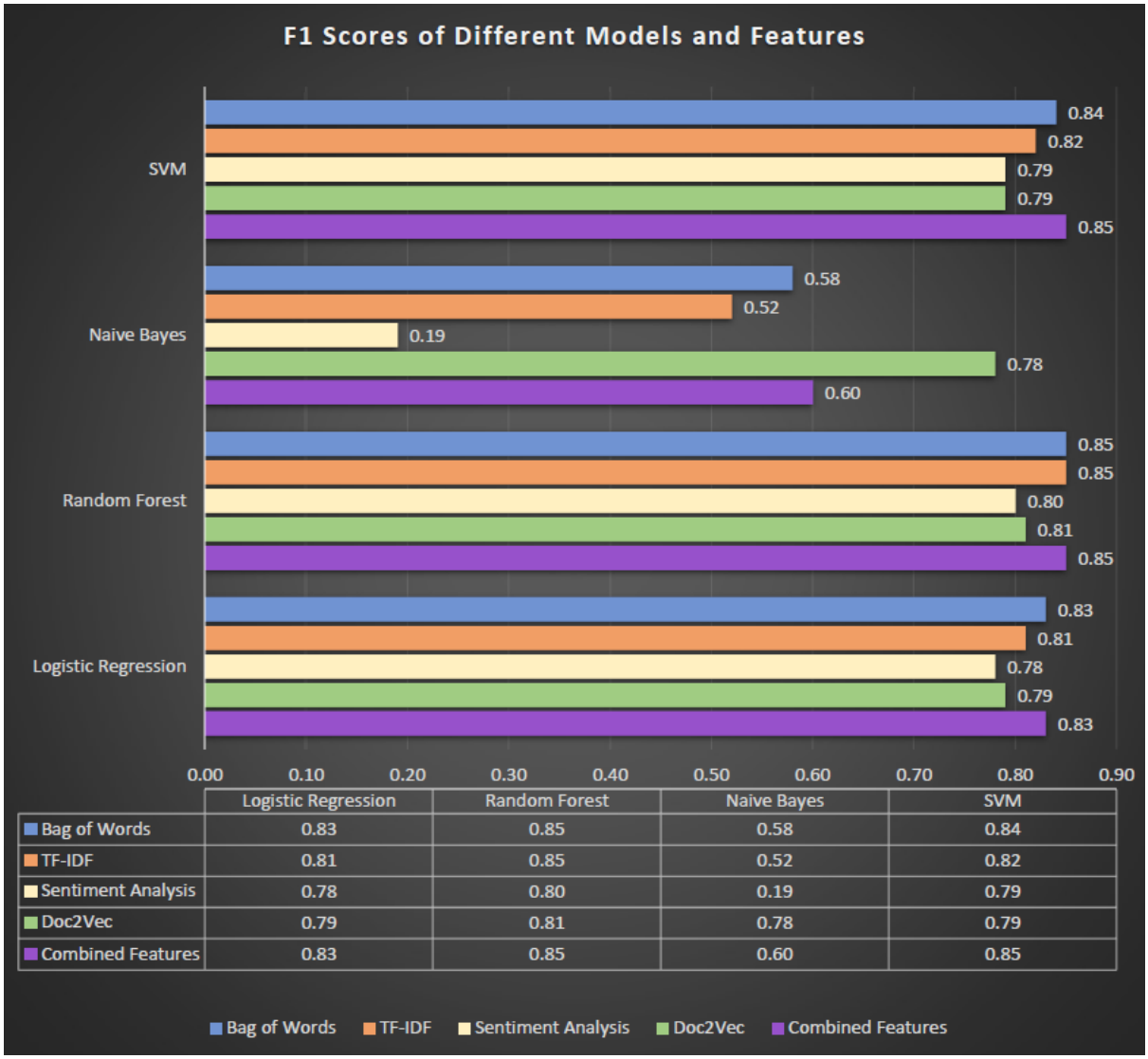
- **Logistic regression classifier -** This machine learning algorithm is a predictive model, mostly used for classification problems. The observations are assigned to a discrete set of classes based on the concept of probability. A logistic or linear regression model uses the sigmoid function as loss function and solves binary and multi-classification problems. The sigmoid functions map the vector inputs which are real values in between 0 and 1. Thus, this classifier provided us with the expected output based on possible probability scores.

- **Random Forest classifier -** Choosing an optimal set of hyper-parameters for machine learning algorithms can be a difficult task. Hence, random forest classifiers can be easier and flexible for classification problems. This classifier without hyper-parameter tuning provides better results than most of the ML algorithms all the time. The supervised learning algorithm is an ensembler of decision trees, tained with the combined method of different learning models with optimal results. The decision trees can measure the importance of different features; and later these can be a great aid to decide which features should be taken into consideration for a better result.

- **SVM -** A support vector machine (SVM) is a supervised learning method, popular for classification problems. This approach is best suited for a limited number of data. The principle idea of this method is that a hyper-plane has to fit at best between two categories of data from the input corpus. Support vectors are those points of the dataset that are nearest to that hyper-plane, and removal of either of them will change the positions of all.

9 Algorithm Testing and Analysis

We have presented our result after running the models in **Fig. 16**. After vectorization of the corpus, the pre-processed tweets are run in 4 different models : "Logistic Regression", "Random Forest", "Naive Bayes" and "SVM". We conducted "Bag of Words", "TF-IDF", "Sentiment Analysis", "Doc2Vec" and "Combined Features" using the 4 models and noted their accuracy level. It is seen that the Random Forest model could conduct Bag of Words, TF-IDF and Combined Features with an accuracy of 85%, and this model also gives the overall highest accuracy for Doc2Vec and Sentiment Analysis which is 81% and 80% respectively. On the other hand, Naive Bayes model conducted every feature with their lowest accuracy and it failed miserably to conduct Sentiment Analysis i.e. with an accuracy of only 19%. SVM model has shown a better performance, with 85% accuracy for combined features, 84% accuracy for bag of words, 82% accuracy for TF-IDF and 79% accuracy for both Sentiment Analysis and Doc2Vec. Similarly, Logistic Regression conducted Bag of Words, Combined Features and TF-IDF with an accuracy of more than 80%; Sentiment Analysis and Doc2Vec with 78% and 79% accuracy respectively. So after the comparison, we can sum up that Random Forest model is showing the best



	Logistic Regression	Random Forest	Naive Bayes	SVM
Bag of Words	0.83	0.85	0.58	0.84
TF-IDF	0.81	0.85	0.52	0.82
Sentiment Analysis	0.78	0.80	0.19	0.79
Doc2Vec	0.79	0.81	0.78	0.79
Combined Features	0.83	0.85	0.60	0.85

Fig. 15. F1 scores of the Models using different Features

performance for all the features. Both SVM and Logistic Regression models show almost similar results with high accuracy. And finally, we do not prefer to use Naive Bayes model as it has shown poor performance for all the features except for Doc2Vec.

10 Work Plan

As shown in **Fig. 16**, for our Pre-Thesis I, we have read some past research papers and summarized them. We have also collected the dataset from twitter, cleaned the dataset, and manually tagged them as whether they are considered hate speech or not, and whether they contain profanity or not. Finally, we have written the Pre-Thesis I report.

In our Pre-Thesis II, we reviewed existing algorithms and models. We developed different models in different word vector space. After that, we tested the algorithms, analyzed

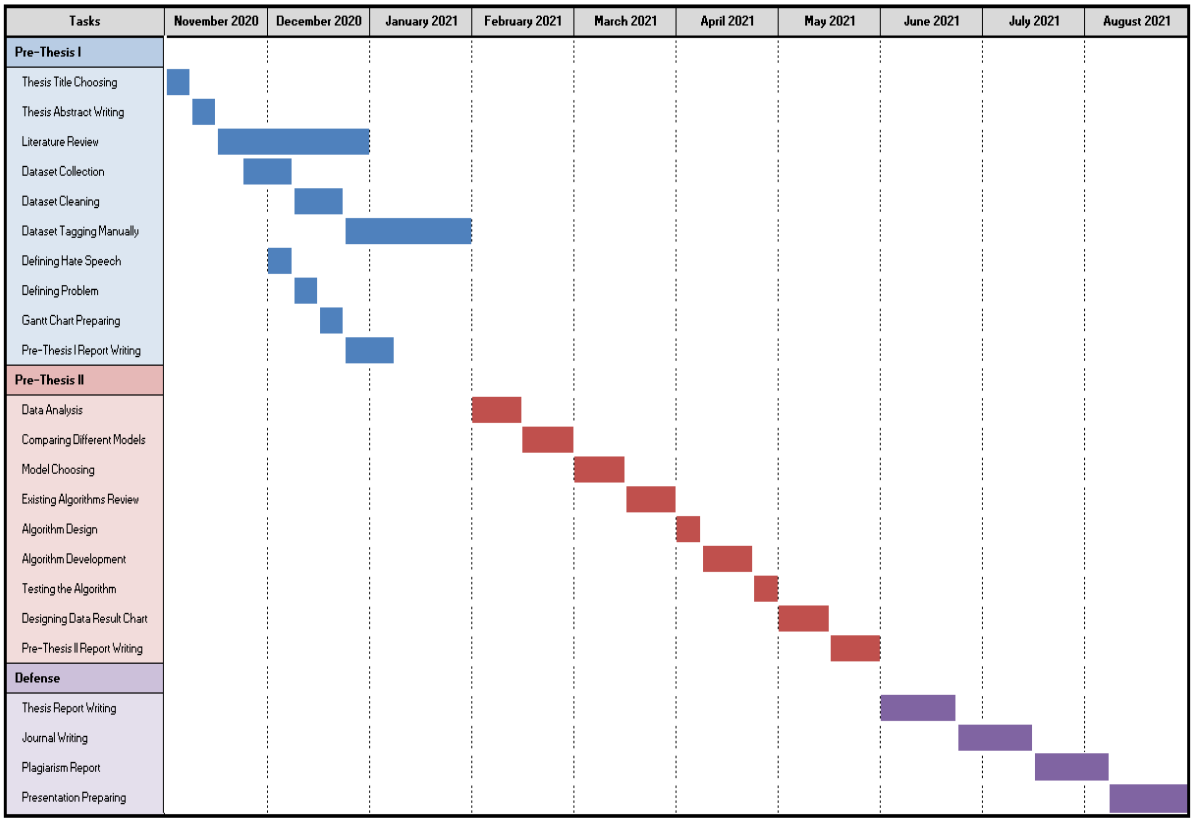


Fig. 16. The Gantt Chart of our Work Plan

the data, compared the accuracy of each model in each vector space, and designed the data comparison chart. Lastly, we have written this Pre-Thesis II report.

As we can see the models are simple and have moderate accuracy, we are mapping out to add more works for our final thesis work. It is known to all, polarity analysis that is classifying negative, positive and neutral tweets is so much easier to do. But emotional analysis is difficult, as we can see some ambiguity in expression. People always do not intend to express the literal meaning of a particular word. We got some good scores training the models. But the models could not identify all the actual meaning of the users and might have somewhat misunderstood the concept. Considering such a scenario, we aim to add some more features in those models for better accuracy along with correct contextual detection. We plan to do clustering on our data-set before training the models and will inspect if it helps to provide better accuracy. However, adding optimizers can also reduce time and give better accuracy.

Finally for our Defense, we will write the final Thesis report and then the journal in IEEE format. We will then prepare the plagiarism report. Lastly, we will prepare for our presentation.

11 Conclusion

As the influence of social media in daily life is deep-seated, moderating online contents should be taken into serious consideration. It has become an absolute necessity to establish accurate, efficient and automated methods to tag abusive language in these platforms. As natural language understanding is quite complex, many researchers have been working in this field for years coming forward with different approaches from NLP, deep learning, neural networking, classic ML methods and features. In this paper, we tried to understand the difference between hate speech, profane speech and regular speech, and the conflict that may arise while differentiating them through artificial intelligence. We have also designed algorithms of different models and vector space, compared their f1 scores. For our future work, we will build a more efficient algorithm to detect hate speech and profanities.

References

[1] Martins, Ricardo, et al. "Hate speech classification in social media using emotional analysis." *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*. IEEE, 2018.

[2] Guterres, A. "United Nations Strategy and Plan of Action on Hate Speech." *Taken from: https://www. un. org/en/genocideprevention/documents/U* 20Strategy (2019).

[3] Davidson, Thomas, et al. "Automated hate speech detection and the problem of offensive language." *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 11. No. 1. 2017.

[4] M. Castillo. "Facebook's artificial intelligence still has trouble finding hate speech." *Taken from: https://www.cnbc.com/2018/05/15/facebook-artificial-intelligence-still-finds-it-hard-to-identify-hate-speech.html.* 2018.

[5] "Zuckerberg on refugee crisis: 'Hate speech has no place on Facebook'." *Taken from:https://www.theguardian.com/technology/2016/feb/26/-mark-zuckerberg-hate-speech-germany-facebook-refugee-crisis.* 2016.

[6] Yin, Dawei, et al. "Detection of harassment on web 2.0." *Proceedings of the Content Analysis in the WEB* 2 (2009): 1-7.

[7] Nobata, Chikashi, et al. "Abusive language detection in online user content." *Proceedings of the 25th international conference on world wide web.* 2016.

[8] Waseem, Zeerak, and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech detection on twitter." *Proceedings of the NAACL student research workshop.* 2016.

[9] Teh, Phoey Lee, Chi-Bin Cheng, and Weng Mun Chee. "Identifying and categorising profane words in hate speech." *Proceedings of the 2nd International Conference on Compute and Data Analysis.* 2018.

[10] Schmidt, Anna, and Michael Wiegand. "A survey on hate speech detection using natural language processing." *Proceedings of the Fifth International workshop on natural language processing for social media.* 2017.

[11] Ahluwalia, Resham, et al. "Detecting hate speech against women in english tweets." *EVALITA Evaluation of NLP and Speech Tools for Italian* 12 (2018): 194.

[12] Badjatiya, Pinkesh, et al. "Deep learning for hate speech detection in tweets." *Proceedings of the 26th International Conference on World Wide Web Companion.* 2017.

[13] Kiela, Douwe, et al. "The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes." *arXiv preprint arXiv:2005.04790* (2020).

[14] Chen, Ying, et al. "Detecting offensive language in social media to protect adolescent online safety." *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing.* IEEE, 2012.

[15] Deng, Lingjia, and Janyce Wiebe. "How can NLP tasks mutually benefit sentiment analysis? A holistic approach to sentiment analysis." *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis.* 2016.