Hate Speech Detection on Twitter

Saurav Pathak
School of Engineering and Applied Sciences
University at Buffalo
Buffalo, New York
sauravra@buffalo.edu

Ishpreet Kaur
School of Engineering and Applied Sciences
University at Buffalo
Buffalo, New York
ikaur2@buffalo.edu

Sai Teja Mattam
School of Engineering and Applied Sciences
University at Buffalo
Buffalo, New York
saitejam@buffalo.edu

Sumedh Khodke
School of Engineering and Applied Sciences
University at Buffalo
Buffalo, New York
sumedhk@buffalo.edu

Omkar Rajguru
School of Engineering and Applied Sciences
University at Buffalo
Buffalo, New York
orajguru@buffalo.edu

***Abstract*—Social media has not only become an integral part of our lives but is also a medium to express our thoughts. Twitter, being most widely used, is host to massive and unfiltered feed of tweets which have a huge impact. The tweets that include hate speech are alarming, when they are directed towards an individual or a group, thus creating an urgency to develop effective countermeasures. This project aims at classifying such tweets into hate speech or non-hate speech by using various classification approaches such as Naïve Bayes, Linear Support Vector Machine, XGBoost, Long Short Term Memory and Logistic Regression and monitor hate trends by identifying its characteristics based on various parameters. This would not only help to reduce hate against certain ethnic group/individual but also perform predictive policing.**

Keywords—toxicity, hate, twitter, speech, social media, classification, tweets, sentiment analysis

I. INTRODUCTION

Microblogging sites are host to millions of users across the world. These sites have become an important medium for people with various educational backgrounds, different cultures to portray their thoughts in the form of words. These microblogs are people's experiences on topics ranging from reviewing products or properties at tourist places, opinions on current affairs or world politics, propagating ideas or as a mode of announcements for all leading organizations. The most widely used microblogging site is Twitter. Users on Twitter have a choice to choose what they want to see in their feed and respond to them. It is thus observed that users on Twitter post harsh opinions targeted towards an individual or a group. With this enormous rise in the data, the hate speech on twitter also increases. Such speech fuels hate which in turn escalates and may incite violence. Any speech that attacks an individual or a group on basis of any attribute related to him/her or the group can be classified as hate speech.

A. Problem Statement

Tweets including such hateful words are alarming, in a sense creating an urgency to develop effective countermeasures. The main objective of this paper is to classify such tweets into hate speech and non-hate speech using modern classification techniques.

B. Previous work

- A lot of research has been done in terms of toxicity detection on twitter. One such method is 'Ensemble Method for Twitter Hate Speech Detection' [1]. The ensemble method for detection of hate speech in Indonesian language by employing five stand-alone classification algorithms, including Naïve Bayes, K-Nearest Neighbours, Maximum Entropy, Random Forest, and Support Vector Machines, along with two ensemble methods, hard voting and soft voting, on Twitter hate speech dataset to classify the tweets to hate or non-hate speech
- Another such research conducted is the construction of a multi-view SVM approach that achieves near state-of-the-art performance while being simpler yet, at the same time, producing more interpretable decisions than neural methods to classify hate speech. [2]
- A model classifier was created in a way that it uses sentiment analysis techniques in particular, subjectivity detection to not only detect that a given sentence is subjective but also to identify and rate the polarity of sentiment expressions.[3]

C. Objective

The main objective of this project is to analyze the tweet sentiment based on the words and emoticons used in them. To develop such a model, it is imperative to understand type of users and the content posted through their account. For this purpose, specific words expressing strong emotions and hashtags are used as keywords, and finally classified into hate or non–hate speech.

D. Outline

In this report, we will summarize the work that has been done on the field of classifying hateful and non-hateful tweets by us by using our data cleaning approach and classifying the tweets into hateful and non-hateful speech by using various models. A brief outline of our approach is as given below:

- Clean the tweets by applying various data cleaning techniques.

- Add/remove irrelevant words/tweets.
- Apply various exploratory data analysis techniques to gather insights from the data.
- Get the words and word count which helps the most in classifying the tweets.
- Apply various algorithms on the data to get the result and filter the tweets as hate or non-hate speech

II. MOTIVATION

A. *Social Media*

"One positive trend we see is that there's significantly more healthy conversations going on than toxic ones, but the toxic conversations get more press. We want to make the positive conversations more effective."

- Professor Matthew Williams, HateLab's Principal investigator [4]

Social Media is a platform for sharing data with a huge and vast number of audiences. It can be addressed as a medium of propagating information through an interface. Social media in tandem with social networks helps individuals cater their opinions to a wider society and reach out to more people for sharing or promotion.

As in today's world, everyone shares their emotions and opinions online, through social media platforms and thus, the data generated by these platforms can be used for the analysis of the sentiments expressed by the users on various posts. But hate and toxic speech targeted at certain group of people or race, or community needs to be blocked before posting it on social media as it causes social disturbance among the various sectors of communities due to hate speech.

Hate crime strand	2016/17	2017/18	2018/19	2019/20	2020/21	% change 2019/20 to 2020/21
Race	58,294	64,829	72,051	76,158	85,268	12
Religion	5,184	7,103	7,202	6,856	5,627	-18
Sexual orientation	8,569	10,670	13,311	15,972	17,135	7
Disability	5,254	6,787	7,786	8,465	9,208	9
Transgender	1,195	1,615	2,185	2,542	2,630	3

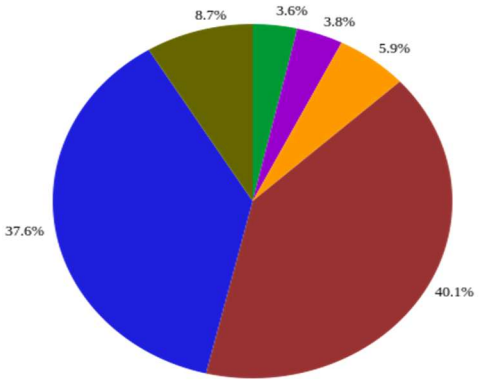
Fig 1: Hate Crime percent change by years.

B. *Importance of tweets on twitter*

Twitter is the most widely used microblogging website. The microblogs posted on this site can be referred to as tweets. A tweet contains images, videos, emoticons, or words with a limit of up to 280 characters. A user has the choice to control the content that is shown in the feed.

A tweet is a medium for a user of expressing thoughts, opinions, or experiences. The domain of users ranges from individuals representing themselves or a group, to organizations building their brand value through various advertisements and announcements.

The major challenge is to consider that not all users on twitter are active, for example, there are users that do not post any content but only view/follow other people organizations. Also, another issue while considering the tweets is the number of fake accounts. This number is considerably large since users try to influence others by imitating the activity of an authentic account. Twitter itself solves this issue to a major extent by verifying profiles that are likely to be imitated.



Content of tweets according to Pear Analytics:

- News (3.6%)
- Spam (3.8%)
- Self-promotion (5.9%)
- Pointless babble (40.1%)
- Conversational (37.6%)
- Pass-along value

Fig 2: Classification of content of tweets

III. APPROACH

A. *Dataset details*

The dataset was a collection of set of tweets with labels as toxic and nontoxic. The dataset has 3 variables – 'id', 'label' and 'tweet'. Each of the variables is defined as:

- id: This is the unique identity for each row in the dataset.
- label: This variable stores the value '0' for tweets not containing hate speech and value '1' for tweets containing hate speech.
- tweet: This variable stores the tweets in the text format.
- Since 'label' is numeric, we converted the numeric value of the column to a factor using a built-in R function.
- Further, a new dataset is created, that is basically a subset of the original data but includes only 2 variables: 'label' and 'tweet'.

1) *Data cleaning and preprocessing:*

- In this step, we performed preprocessing steps for data cleaning to make ready our dataset before applying machine learning algorithms.
- Using VectorSource() a corpus is formed using the available tweets. The required package 'tm' is installed and included in the R code.
- This vectorsource function places our text into a vector, and then load the data into a variable named corpus.
- The new corpus formed in the previous step is used to create a Term Document Matrix using the function available in the 'tm' package.
- The following the operation performed on corpus for cleaning the text in the corpus which are irrelevant for our classification:
    - toLower(): Using this function we converted all the data into lower case.
    - removeNumbers(): By using this function we removed the numbers if any, present in the text.
    - removePunctuation(): By this we removed all the punctuation marks if any, present in the text.
    - removeWords(): This function removed specific words which are irrelevant for the analysis and stop words in the English language.
    - stripWhitespace(): This removed all the whitespaces in the tweets.
- Stemming was applied to remove the last few characters of commonly occurring variations of words by stemming them to their root words.
- Lemmatizer was applied to variations of words to convert them to their base form by considering the context of the sentence in the data points.
- Punctuation correction and Spelling correction was applied to correct the incorrect grammar in the corpus of words.
- Stop words or the commonly occurring words which did not contribute to the label decision step were removed.

2) Exploratory Data Analysis

- We performed EDA to understand the distribution of the tweets as imbalanced distribution may skew the analysis.
- To avoid overfitting on any of the labels by understanding distribution, we plot bar chart of total tweets across tweet labels.
- The following features can be observed after creating the matrix using the 'inspect' function:
    - a) Binary Weighting
    - b) Term Frequency

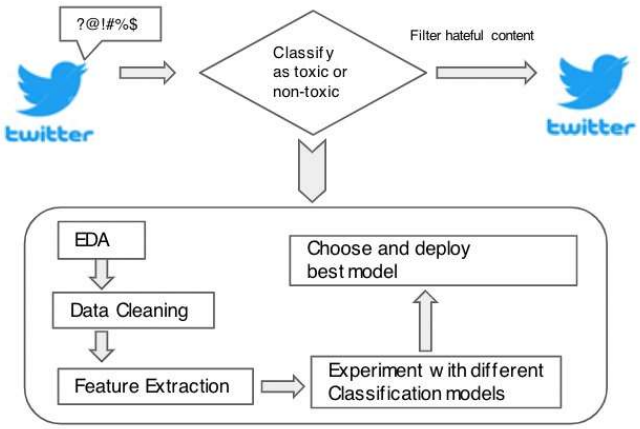


Fig 3: Initial Approach

3) Feature extraction

- We used TF-IDF for feature extraction.
- We used the 'weightTfIdf()' function to evaluate the term frequency of the relevant words in our tweets.
- TF-IDF defines importance of a term by taking into consideration the importance of that term in a single document and scaling it by its importance across all documents.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

- Term frequency answers the question of, how many times does this word appear in this document among the number of times all words appear in this document? In other words, how important is this word to this specific document?
- Inverse document frequency answers the question of, how common (or uncommon) is this word among all the documents?
- TF-IDF is a popular approach used to weigh terms for NLP tasks because it assigns a value to a term according to its importance in a document scaled by its importance across all documents in your corpus, which mathematically eliminates naturally occurring words in the English language and selects words that are more descriptive of your text.

- TF-IDF gave us the features that we would utilize for our machine learning binary classification problem.

4) Models

After collecting and extracting the required features from our dataset, we now have to feed our dataset to various classification models.

The various models which we have used to perform our analysis and the details of how they work is described next.

Naïve Bayes:

- This is algorithm based on the Bayes theorem of calculating conditional probability where P(A) & P(B) are the probabilities of events A & B and A|B is the occurrence of event A given the occurrence of event B [5].

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

- This algorithm is useful is classifying categories, but it violates the independence assumption of Bayes Theorem. We assume that none of the features are dependent and that each feature is considered with same priority. This assumption of features being independent of each other is naïve and not the case in practice for real world examples. Thus, the algorithm is called Naïve Bayes Theorem, because of this naïve assumption.

XGBoost:

- XGBoost is an open-source library providing a high-performance implementation of gradient boosted decision trees.
- It is an ensemble learning method. Ensemble learning provides an efficient solution to combine the predictive results of multiple learners, since it is not always enough to rely on one particular model.
- The underlying framework that is used by XGboost is Gradient Boosting framework. It also provides a parallel tree boosting that solves many data science problems with less time and more accuracy.
- Gradient Boosting [6]:

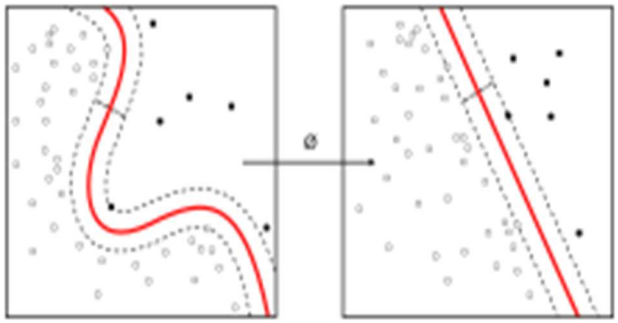


Fig 4: Scatterplot featuring a linear support vector machine's decision boundary (dashed line)

SVM:

- Support Vector Machine algorithm creates a decision boundary that can segregate the data points into different classes so that a new data point is correct classified in the correct class [7][6].
- SVM finds a linear separator based on the maximum margin using support vector. SVM chooses the extreme points that create the hyperplane.
- Support vectors are the data points that lie closest to the hyperplane or on the margin. They are the critical elements of the training set that would change the hyperplane dividing the datapoints.
- SVM constructs a set of hyperplanes in infinite dimensional spaces, which can be used for regression, classification, and other tasks like outlier detection.
- Additionally, a good separation is achieved by the hyperplane that has the largest distance to the nearest point.
- A larger margin indicates lower levels of generalization error of the classifier.

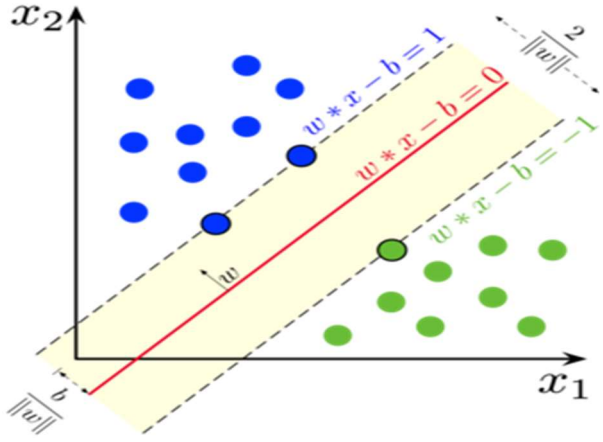


Fig 5: Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin

Word2vec:

- It is a two layer neural network that processes text by vectorizing words.

- The purpose of word2vec is to group the vectors of similar words together in vectorspace.
- Word associations are learnt from a large corpus of text using a neural network model.
- The cosine similarity between the vectors indicates the level of semantic similarity between the words represented by the vectors.

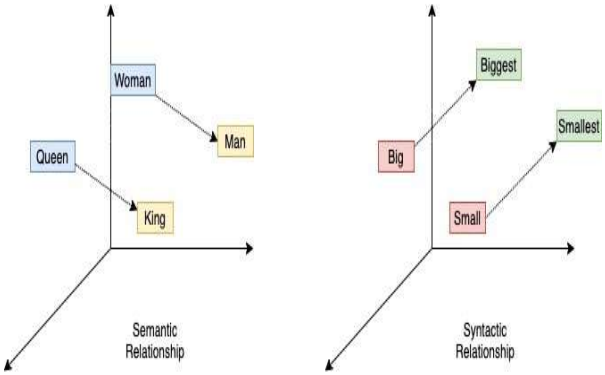


Fig 6: Trained Word2Vec Vectors with Semantic and Syntactic relationship [8]

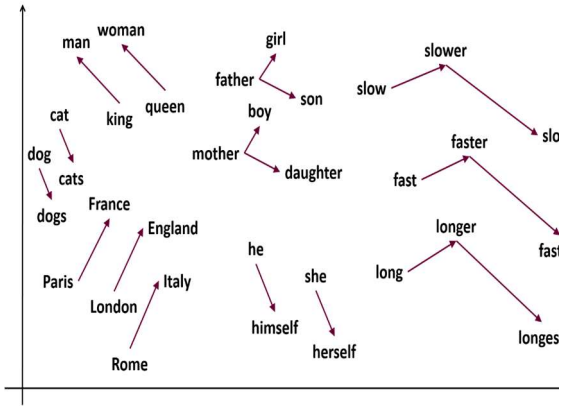


Fig 7: Word2Vec Vectors [9]

LSTM:

- Long Short-Term memory is an artificial neural network structure. It has feedback connections that are helpful in the field of deep learning. LSTM is able to process single data points (For e.g., Image) and also sequences of data (For e.g., Video). Applications of LSTM include anomaly detection in network traffic, speech recognition etc. LSTM cell can process data sequentially and keep its hidden state through time [10].

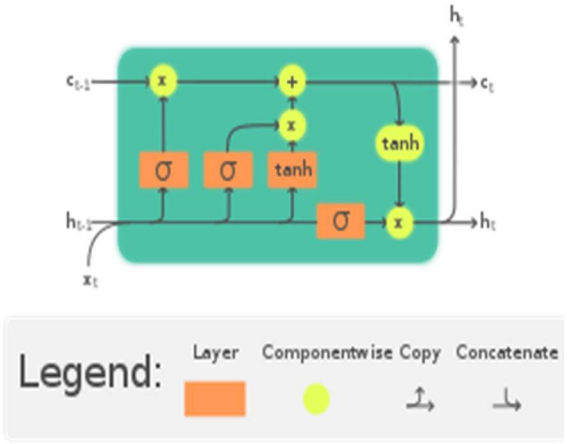


Fig 6: The Long Short-Term Memory (LSTM) cell can process data sequentially and keep its hidden state through time.

Logistic Regression:

- Logistic regression means modeling the probability of a discrete outcome given an input variable. There are different types of logistic regression. Multinomial logistic regression can model situations where there are more than two possible discrete outcomes. [11]
- It is a linear regression but for classification problems where the dependent variable (target variable) is categorical.
- Logistic regression uses a logistic function to model a binary output variable (yes or no).

IV. RESULTS

- Initially we started by visualizing our data. We plotted the graph of tweets which are actually hateful speech, and which are non-hateful.
- We found approximately 43000 out of our total 57000 tweets on the data set is hate speech.
- Only about 14000 tweets were non hateful speech. This meant that our data is heavily skewed on one side that is hate speech.
- The following graph will effectively demonstrate the extent to which our data was biased:

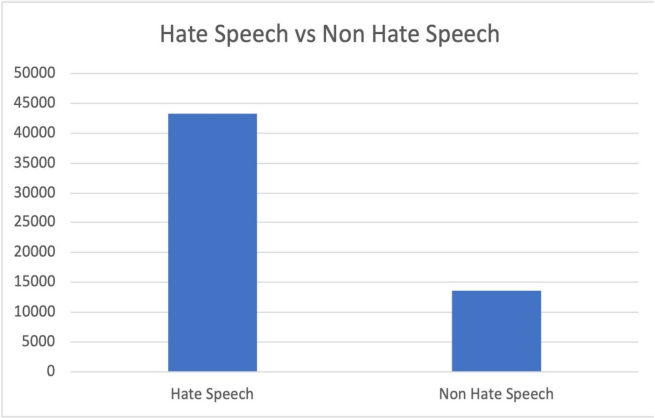


Fig 7: Distribution of tweets in the original dataset.

Key Metrics:

Sensitivity and specificity describe the accuracy of a test which reports the presence or absence of a condition, in our case hate or non-hate speech, in comparison to a 'Gold Standard' or definition [12].

In our case, Sensitivity (True Positive Rate) refers to the proportion of those tweets which is a toxic speech (when judged by the 'Gold Standard') that received a positive result on our analysis model. Whereas, Specificity (True Negative Rate) refers to the proportion of those tweets which are non-toxic (when judged by the 'Gold Standard') that received a negative result by our model.

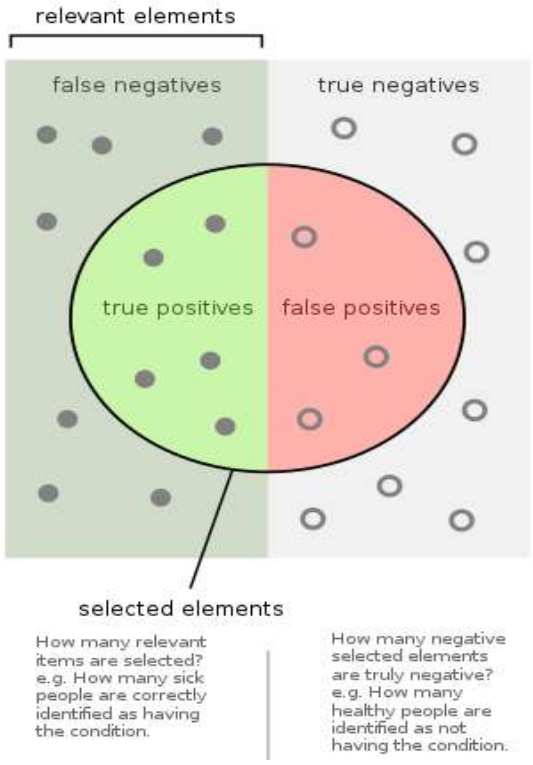


Fig 9: Sensitivity vs Specificity

Z-Score: This score helps to understand if a data value is greater or smaller than mean and how far away it is from the mean. More specifically, Z score tells how many standard deviations away a data point is from the mean.[13]

$$z = \frac{x - \mu}{\sigma}$$

Accuracy comparison:

- We trained our model after preprocessing and cleaning the data using various classification algorithms such as Naïve Bayes, SVM, XGBoost, LSTM and Logistic Regression
- Long Short Term Memory yielded the best of all the results in our analysis achieving an accuracy of almost about 89% on the test set.
- Following LSTM, XGBoost yielded the second-best accuracy of around 88% correct classifications on test set.
- When we used SVM on our test dataset, it yielded accuracy of about 87% which is the third highest accuracy of all the models we used.
- Naïve Bayes produced next to the worst accuracy of about 84% on our test dataset.
- Logistic Regression produced the worst accuracy among all the models we have used aggregating to about 75% of accuracy when deployed on the test dataset.

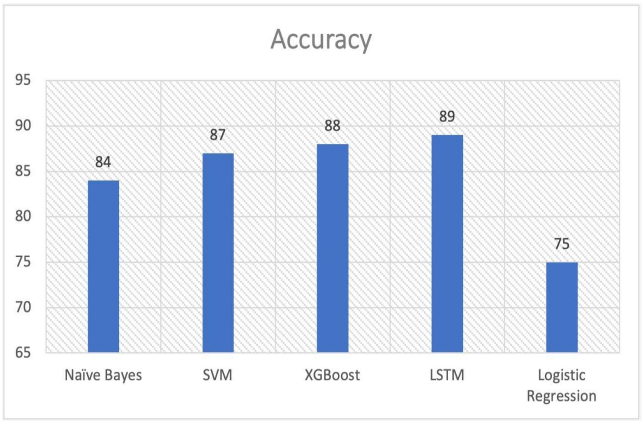


Fig 9: Accuracy comparison of various models

V. BROADER IMPACT OF HATE SPEECH DETECTION

- Monitoring twitter content trends to enforce predictive policing for hate speech.
- Filtering hate speech make humans more kind towards one another as one's mind is more used to only non-hate speech.
- Removal of toxic tweets will result in reduction of violence as one is less likely to get carried away by

hatred sentiments and in turn end up by doing some crime against certain specific individuals/groups.

- Results can help to curb the spread of hatred against certain groups/individuals that creates a negative social and political impact.
- In addition to existing techniques, automating such a model may save a lot of time with minimal human effort required for toxicity detection.

VI. Challenges and Future Work

- Major challenge is faced due to the presence of different linguistics on Twitter.
- Even if a user is writing a tweet in English, one cannot guarantee that it is the English language in its entirety.
- Another issue is with detecting the sentiment of the writer (of tweet).
- Past work in sentiment analysis has focused extensively on detecting polarity, and to a smaller extent on detecting the target of the sentiment i.e., feeling of the tweet.
- It is sometimes confusing for even humans to detect what the writer (of tweet) is semantically trying to convey.
- So, it would be extremely difficult for computers to do it automatically with a few thousand labeled tweets as input for training.
- The presence of sarcasm and ridicule in tweets makes it even harder to detect sentiment correctly.
- This is because they can often indicate a positive emotional state of the speaker (pleasure from mocking someone or something) even though they have a negative attitude towards someone or something.
- Usually, rhetorical questions can be regarded as a neutral sentiment as these are mere questions with no positive/negative annotation. For example, consider: Do you want to be a failure for the rest of your life?
- On the one hand, this tweet can be treated as a question one asks, but on the other hand, it can be seen as a negative annotation because of the presence of words like “failure”.
- Sometimes it is difficult to precisely identify the target of opinion in that tweet. Therefore, precisely determining the opinion on that subject becomes a non-trivial task.
- Similarly, quoting somebody else or re-tweeting is difficult to classify because it is often unclear and not explicitly evident whether the user who re-tweets has the same opinions as that expressed by the original user of that tweet.
- The challenges listed above need to be addressed on various levels and using just one model is not enough. We need to pre-process data carefully and also look for the quality of data that we’re inputting.
- Another remaining challenge is that automatic tweet toxicity detection is a closed-loop system; people are aware that a filtering is happening, and try to evade detection while spreading hatred.
- For instance, knowing twitter removes hateful posts from the platform. Users who wish to spread the hatred have also found profound ways to bypass these filtering measures by, for instance, posting the content as images containing the text, instead of the text itself.
- Although image recognition can be used to solve this particular problem, this further illustrates the complexity of hate speech detection moving ahead.
- It will be a constant back and forth battle amongst those who are trying to spread hatred and toxicity and those trying to block it.

VII. Summary

- As hate speech continues to be a major problem, the need for filtering out the toxic tweets is also rising. We have provided with an approach with our classifying method to filter the hate speech/toxicity detection of the twitter tweets along with achieving a decent accuracy. We hope our method will help reduce toxicity on twitter and help filter the hate speech on twitter. Additionally, we hope our approach will help to find deeper trends in tweet toxicity.

References

[1] Fauzi, Muhammad & Yuniarti, Anny. (2018). Ensemble Method for Indonesian Twitter Hate Speech Detection. Indonesian Journal of Electrical Engineering and Computer Science. 11. 294-299. 10.11591/ijeecs.v11.i1.pp294-299.

[2] MacAvaney S, Yao H-R, Yang E, Russell K, Goharian N, Frieder O (2019) Hate speech detection: Challenges and solutions. PLoS ONE 14(8): e0221152. https://doi.org/10.1371/journal.pone.0221152 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6701757/

[3] https://www.researchgate.net/publication/283125668_A_Lexicon-based_Approach_for_Hate_Speech_Detection

[4] https://developer.twitter.com/en/community/success-stories/hatelab

[5] https://en.wikipedia.org/wiki/Naive_Bayes_classifier

[6] https://en.wikipedia.org/wiki/Gradient_boosting

[7] https://en.wikipedia.org/wiki/Support-vector_machine

[8] https://towardsdatascience.com/word2vec-research-paper-explained-205cb7eecc30

[9] https://medium.com/analytics-vidhya/implementing-word2vec-in-tensorflow-44f93cf2665f

[10] https://en.wikipedia.org/wiki/Long_short-term_memory

[11] https://en.wikipedia.org/wiki/Logistic_regression

[12] https://en.wikipedia.org/wiki/Sensitivity_and_specificity

[13] https://en.wikipedia.org/wiki/Standard_score

[14] https://ieeexplore.ieee.org/document/9121057

[15] R. Martins, M. Gomes, J. J. Almeida, P. Novais and P. Henriques, "Hate Speech Classification in Social Media Using Emotional Analysis," 2018

7th Brazilian Conference on Intelligent Systems (BRACIS), 2018, pp. 61-66, doi: 10.1109/BRACIS.2018.00019.

[16]

[17] A. Sarlan, C. Nadam and S. Basri, "Twitter sentiment analysis," Proceedings of the 6th International Conference on Information Technology and Multimedia, 2014, pp. 212-216, doi: 10.1109/ICIMU.2014.7066632.

[18] https://help.twitter.com/en/resources/glossary

[19] https://media.twitter.com/en/twitter-basics

[20] https://en.wikipedia.org/wiki/Twitter

[21] https://hatelab.net/data/