e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

HATE AND OFFENSIVE TEXT DETECTION ON TWITTER

Greeshma KS*1, Ms. Francy TL*2

*1,2Department Of Computer ScienceVimala College Autonomous Thrissur, Kerala, India.

ABSTRACT

Twitter's main objective is to make it possible for everyone to create and share thoughts and data as well as communicate their assumptions and beliefs without restrictions. Hate speech is a text that contains threatening or bad words against a particular group of people in a community. With the growing use of social media, this type of content is growing every day. Hate speech can be harmful to an individual or a community.It is difficult, time-consuming, and error-prone to manually analyze this massive and constantly expanding material. Because of this, it is increasingly important for online social media platforms to screen out any messages that contain hate speech before they are posted to the network. I propose an automated classification system that detects whether a tweet is hateful, offensive, or neither. This approach involves extracting various features like TF-IDF with n gram features, sentiment features, doc2vec features, and enhanced features. These features are used to train machine-learning classifiers. Exhaustive experiments are conducted on the existing Twitter dataset and compared to their accuracy in detecting whether a particular tweet is hateful, offensive, or neither. Also find out the best feature set to performthe same.

Keywords: Hate Speech, Twitter, Tf-Idf, Sentiment Analysis, Random Forest Classifier, Logistic Regression, Support Vector Machine, Guassian Naive Bayes.

I. INTRODUCTION

In the last century, the number of people using social networks and online forums has grown exponentially. Twit-ter has more than 348 million users, and 350000 tweets are created on the platform every second, according to its official website. People from various cultural and intellec- tual backgrounds voice their opinions on daily issues likepolitical views, product reviews, and movie reviews. Suchinformation is valuable to many businesses and organizations because it allows them to learn what consumers think about political issues, products, and entertainment options . Due to disagreements over ideas, communication between individuals can occasionally become abusive. As a result, hate speech on social media is becoming more prevalent every day.In order to manage such a big number of users on social media, methods for automatic detection of hate speech are needed. In this study, we classify whether a tweet contains hate ,offensive or neither using different machine learning techniques.

Machine learning is a subset of Artificial intelligence. ML models are used for classification ,regressions and clustering. A ML model can be one of the following types: supervised, unsupervised, or reinforcement. A supervised ML model is used for classification as well as regression problems. A classifier classifies a data point into a specific group or class. It is done by using a labeled data set. A regression model tries to solve regression problems. In which the model produces a continuous value as output rather than discretevalues. A regression model tries to solve regression problems. In which the model produces a continuous value as output rather than discrete values. An unsupervised model produces clusters based on their similarities. Unsupervised model uses the unlabelled dataset for training and the model tries to study hidden patterns from the data set. Reinforcement learning doesn't use the dataset. It learns by itself through a trial-and- error mechanism in order to increase the reward.

In my work, We aim to determine whether a tweet is hateful, offensive, or neither. For my work, We used hate speech and offensive language detection on the Twitter dataset.The work starts with the preprocessing of text and extracting various features. Finally, We trained random forest, logistic regression, naive Bayes, and SVM models by using the variouscombinations of the above features and predicted which featureand model was best for determining whether a tweet is hateful,offensive, or neither. According to Paula Fortuna and Sergia Nunes [1] "Hate speech is language that attacks or diminishes, that incites violence or hate against groups, based on specific characteristics such as physical characteristics, sexual orientation, gender identity, religion, descent, national or ethnic origin, and other factors. It can also happen with various linguistic approaches, even subtly or when humor is utilized."

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

This paper begins, in Section II, from literature review, which produce an summary of existing models that helped my research. In the following section ie; section lll, we explain my proposed work, section 1V, we reveal our results. Finally in section v explain the conclusion of my research.

II. LITERATURE REVIEW

In terms of research, the field of hate speech detection is still quite new, yet it has made many important advancements already.In [2] Pranav Malik et al. proposed a model named Toxic Speech Detection Using Traditional Machine Learning Models and BERT and Fast Text Embedding with Deep Neural Networks. In which they used the HASOC and ALONE datasets. They tokenized with an ngram range of (1, 3) and performed basic preprocessing techniques like stopword re- moval, stemming, lemmatization, and part-Of-Speech tagging.

Then they used the TF-IDF approach over the traditional BOW approach for creating the count vectorizer. They trained ML models such as LR, DT, and RF, ensemble models such as gradient boosting, stacking, and bagging, Word embedding techniques like Bert and Fastext, and deep learning models like MLP, LSTM, and CNN. Finally, compare the results. They found out that CNN is the best model for detecting whether a tweet is hateful, offensive, or neither.In [3] Ha- jime Watanabe et al. A Pragmatic Approach to Gather Hateful and Offensive Expressions and Conduct Hate Speech Detection on Twitter is a model that was put out. In which they combined three datasets named as: A first dataset publicly available on Crowdflower2, a second dataset publicly available also on Crowdflower, and a third data set that has been published on GitHub. In which they extract sentiment- based, semantic based features, Unigram features, and Pattern Features. Then they trained the model toolkit Weka using the J48craft algorithm. They found that unigram and pattern-based features produce high accuracy in predicting hate and offensive text on Twitter. In [4]Garima Koushik et al. proposed another model named Automated Hate Speech Detection on Twitter to detect hateful tweets. With the help of the publicly available Twitter dataset, they extract its bag of words and TF IDF (term frequency-inverse document frequency) features, these features are then provided to train logistic regression classifiers. In [5]Havannur Sahi et al. proposed a model for automated detection of hate speech towards women on Twitter. To find out about hate speech towards women, they used the hashtag kiyafetimekarisma. They collected 1288 tweets from Twitter by searching the hashtag kiyafetimekarisma and created a dataset that contains attributes named hate and non-hate. They extracted bigram, unigram, and trigram using a feature engineering technique called TF-IDF scores. Instead of that, they used the Flesh-Kincaid grade score and the Flesch reading ease score to capture the quality of tweets. Finally, they applied a variety of models supporting vector machines , Decision Tree, Random Forest to different feature sets and compared their accuracy. In [6]Sindhu Abro et al. proposed a model for classifying a tweet into three different classes namely, “hate speech, offensive but not hate speech, and neither hate speech nor offensive speech”. In which after preprocessing they performed tf-idf, word2vec, doc2vec feature engineering techniques. Then split the data set for training and testing . Finally they trained on different ML models uch as NB, SVM, KNN, DT, RF, AdaBoost, MLP and LR and compare the result.

III. PROPOSED MODEL

In fig 1 shows the basic idea of the methodology proposed in our research work for the purpose of classifying text into either of the three categories: hate, offensive, or neither. As an initial step, We collected the dataset named hate speech and offensive language detection on Twitter obtained from Kaggle. As a part of the first step, the data is pre-processed through the removal of white space, removal of urls , removal of punctuation, capitalization, tokenization, and stemming.

Then extract various features like n gram based TF-IDF, sentiment features, Doc2Vec features, and enhanced features. The combination of different features is used to train the ML models named Random Forest, Logistic Regression, Support Vector Machine, and Guassian Naive Bayes.

A. DATA SET

For my research we used publicly available hate speech and offensive language detection on twitter dataset from kaggle. Which consist of 25296 rows and 6 columns named id, count, hate, offensive, neither, class and tweet. class label for majority of CF users. 0 - hate speech 1 - offensive language 2 - neither Tweet : text tweet from twitter The class distribution of dataset is pictorised in fig2 . From the data set we extract features tweet and class.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

B. DATA PREPROCESSING

In order to produce better results Twitter® comments were pre-processed before analyzing it [7].

Space removal :The majority of the time, text data contains additional spaces, or more than one space is left between the text after using the aforementioned preprocessing approaches, thus we need to address this issue. The regular expression library works effectively to address this issue.

Removal of capitals : Because lowercase and uppercase are handled differently by the computer, if the text is in the same case, it is simple for a machine to read the words. Words like Bat and bat, for instance, receive distinct treatment from the computer. In order to avoid these issues, the text mustbe written in the same case, with lower case being the most desired.

Url removal : Url,https are unwanted words because it doesn't produce any important information. So it is best to remove it.

Removal of unwanted symbols : also it is like url removal. Unwanted symbols like @,! are not Such informative

Tokenizing : It is the process of splitting a sentence into small units like words.

Remove Stopwords : The most frequent words in a text thatoffer no useful information are called stopwords. Some exam- ples of stopwords include they, there, this, where, and others. A popular library called NLTK is used to eliminate stopwords,and it includes about 180 stopwords that it eliminates. The addmethod makes it simple to add any new word to a group of terms.

Stemming : The practice of distilling a word to its root stem is known as stemming. For example, the word run is the root of the phrases eat, eating, eats, and runed. Stemming is basically taking the prefix or suffix out of nouns like ing, s, es, etc. The NLTK package is used to stem the words. Figure 3 displays the preprocessed dataset's word cloud.

C. FEATURE ENGINEERING

It is crucial to give a machine learning model a list of attributes that are more beneficial than others when performingprediction when training the model. This extraction of features

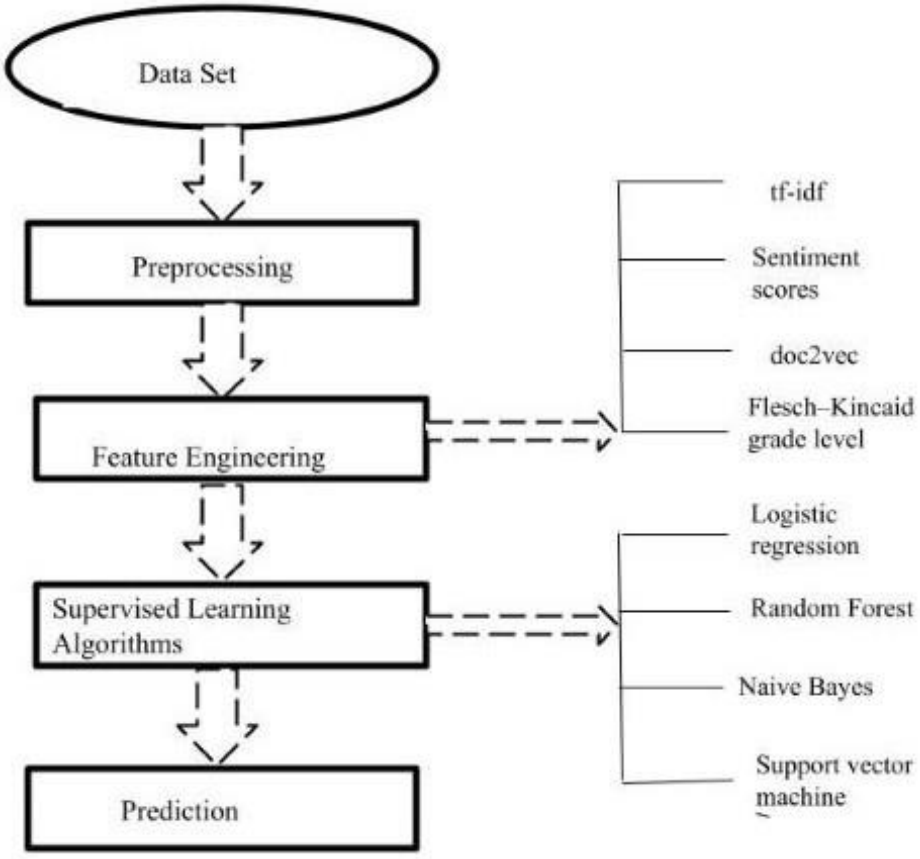

Fig. 1. System architecture

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

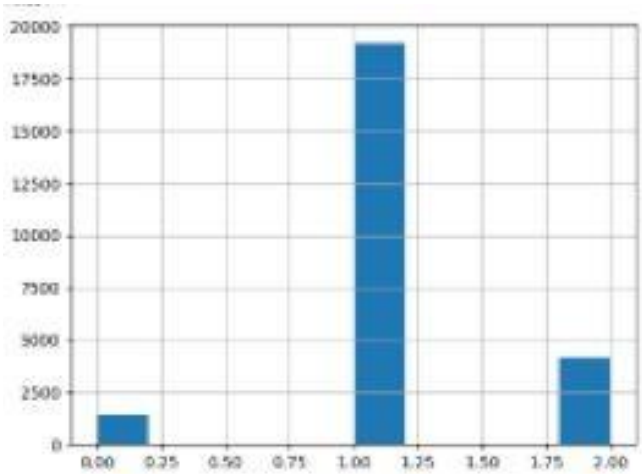

Fig. 2. Histogram of dataset class distribtion

It is referred to as feature extraction.Feature Extraction is a method to convert each tweet into a fixed set of attributes so that it can be easily interpreted by machine learning models

[8] . In this study, we extract several features from the tweets and feed them to a machine learning classifier to divide the tweets into two categories.

Fig. 3. Word cloud of preprocessed dataset

1) tf-idf vectorization: TF-IDF stands for Term Frequency Inverse Document Frequency of records. It can be summed up as determining how pertinent a word is to a corpus or seriesof words in a text. The frequency of a word in the corpus (data set) makes up for the meaning increasing according to how many times it appears in the text. It is implemented by tfidfvectorizer. TF-IDF can be calculated as follows

TF: Term Frequency, which measures how much a word appears in a text. Although each document is specific in duration, it is likely that the word will occur far more oftenin longer documents than in shorter ones.

TF(t) = Count of term t in a document/Total count of terms in document

IDF : Inverse Document Frequency, which measures the importance of a term while calculating TF. Usually, all terms are considered to be equal in terms of importance. However, when certain terms like" is"," of", and" the", may occur a greater number of times but have little importance. Thus, it is necessary to give weight to each and every term.

IDF = log(Total number of documents / Number of docu- ments in which term t is present)

TF-IDF = TF*IDF

2) sentiment features: Sentiment features enable us to identify the underlying emotions of a piece of text. Sentiment analysis is one of the most commonly applied techniques in social media analysis across all domains. The crucial component of sentiment analysis is the examination of a body of text in order to comprehend the context and viewpoint it conveys. We can use Polarity to measure the sentiment towards the positive and negative sentiment because computers can only read numerical data. Through this feature extraction, we categorise the negative, positive, and neutral emotions of a tweet on the basis that hateful and

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

offensive text always liesin a negative tweet rather than a positive tweet.

3) Doc2Vec features: It is a paragraph vectorization model. Act as though a document had an additional floating word-like vector—which we will refer to as a "doc-vector"—that contributes to all training predictions and is updated similarly to other word-vectors. This algorithm is implemented by the Doc2Vec class in Gensim. Doc2Vec can recognise word relationships and comprehends the semantics of the text. An unsupervised approach called Doc2Vec learns fixed-lengthfeature vectors for sentences, documents, and texts. It is an unsupervised method for learning fixed-length vector representations of documents. It is the same as word2vec, but the only difference is that it is unique among all documents [9].Like word2vec, Doc2Vec architecture also contains two algorithms. These two algorithms are Continuous Bag of Words (CBOW) and Skip-Gram (SG), respectively. With the exception of the addition of a paragraph id vector, one of the algorithms in doc2vec is known as Paragraph Vector - Distributed Bag of Words (PV-DBOW), which is comparable to the SG model in word2vec.

4) Enhanced features: Through the use of Flesch-Kincaid readability tests, enhanced features are derived. It is done using readability tests, which show how challenging it is to comprehend an English piece. The Flesch Reading-Ease and the Flesch-Kincaid Grade Level are the two assessments. De- spite using the same basic metrics (word length and sentence length), they are weighted differently. These scores reveal that non-hate tweets are more sophisticated and complex whereas hate tweets are short in length and not complicated.

D. MODEL DESCRIPTION

We applied variety of models that have been used for the classification purpose in the literature including logistic regression, Random forest, Support Vector Machine, GaussianNaive Bayes.

1) Logistic regression: With the use of nominal, or- dinal, or continuous independent variables, the value of a categorical dependent/outcome variable is predicted using a binary supervised machine learning approach called a logistic regression classifier. With the aid of the sigmoid function, logistic regression maps the value of the dependent variableto the interval [0,1].

2) RandomForest Classifier: The supervised learning method includes the well-known machine learning algorithm Random Forest. Regression and ML classification problems can be addressed with it. Its core principle is ensemble learning, which is the process of combining different classifiers to improve the model's performance and solve a difficult task. Random Forest is a classifier that, as its name implies, consists of a number of decision trees on different subsets of the provided dataset and averages them in order to increase the datasets predictive accuracy.Instead of depending on a single decision tree, the random forest uses forecastsfrom all of the trees to anticipate the outcome based on which predictions received the most votes.

3) Support Vector Machine: SVM raises the original lower-dimensional data's dimension via either nonlinear or linear mapping. It searches for the linear optimal dividinghyper-plane within this new dimension to separate the tuples between the sets. Information from two array scans is always distinguished by a hyperplane for adequate nonlinear scaling to an acceptable high dimension. The SVM uses support vectors to locate this hyperplane. The examples that are closer to the edges are known as support vectors. There might be an infinite number of separating lines drawn here. With regard to previously unseen tuples, the goal is to categorize the "highest"one with the least amount of classification error . maintaining the specifications' integrity.

4) Guassian Naive Bayes: It is a machine learning (ML) classification method that is based on the probabilistic method and Gaussian distribution. Gaussian Naive Bayes makes the assumption that each parameter—also known as a feature or a predictor—has the ability to predict the output variable on its own. The final prediction, which provides a probability that the dependent variable will be classified into each category,is the culmination of all previous predictions. The final designation is given to the group that has the highest likelihood.

IV. RESULTS

In Table 1 f1 denotes the tf-idf feature set,f2- tf- idf+sentiment-scores, f3- tf-idf+sentiment-scores+doc2vec feature, f4 -tf-idf+sentiment-scores+doc2vec feature+ enhanced features, f5- tf-idf+doc2vec feature+ enhanced features, f6- tf-idf+sentiment-scores+ enhanced features, f7- sentiment-scores+doc2vec feature+ enhanced features and itis clear that, Random Forest classifier works pretty well when it comes to tf-idf and also shows a

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:03/March-2024 Impact Factor- 7.868 www.irjmets.com

noteworthy performance in every other feature set; nevertheless, the exclusion of tf-idf scores from the feature set significantly reduces its performance.

Table 1: Accuracy Comparison Of Ml Models

Model	f1	f2	f3	f4	f5	f6	f7
Logistic Regression	89.7	89.89	89.79	81.21	80.81	88.35	66.24
Random Forest	90.35	89.36	89.06	87.95	88.35	88.40	82.3
Guassian Naive Bayes	64.91	65.01	65.01	66.24	66.24	66.2	74.4
Support Vector Machine	89.32	89.12	89.28	89.73	60.45	87.91	78.6

The logistic regression algorithm works consistently well with all feature sets but it's performance is degraded when tf-idf scores are removed and sentiment scores removed from the feature set.The overall performance of the Na¨ıve Bayes classifier is found to be less significant for the purpose of classifying tweets into hate, offensive or neither labels but it performs significantly better with the feature set of sentiment score+Doc2Vec+enhanced features compared to other feature sets . SVM classifier also seems to be consistent throughout all feature sets except for tf-idf+sentiment score+enhanced features and sentiment score+Doc2Vec+enhanced features.

Through this we analyze that the most important featurewas found to be the tf-idf scores which helps in better classification of hate speech The sentiment scores also show to be a significant characteristic in the differentiation of offensive language from hate speech. Since removing Doc2vec columns from the feature set barely affects anything, it is determined that these columns are not really significant for categorization purposes..On comparing all the models Random Forest is clearly the winner and tf-idf feature extraction is the best featur set for classifying whether a tweet is hateful, offensive or neither.

V. CONCLUSION

First, we gathered the data set. Using a text pre-processing technique, we remove punctuation, tokenize, remove stopwords, stem, remove urls, and remove mention names from the dataset in order to eliminate undesirable content. After processing, the text is sent on to feature extraction, where characteristics such as sentiment polarity scores, n-gram tf-idf weights, doc2vec vector columns, and other readability scores are retrieved and combined into various sets to match various classification models. The accuracy of these classification models are assessed in relation to various feature sets. According to my research, I found out that Random Forest is the best classifier for predicting the hate and offensive text on twitter and the best feature extractor is tf-idf vectorizer.

We would like to improve the accuracy of our model and perform experiments on more ML models as well as Deep Learning models, and we also wish to perform the same classification on Malayalam Language.

VI. REFERENCES

[1] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Computing Surveys, vol. 51, pp. 1–30, 2018.

[2] P. Malik, A. Aggrawal, and D. Vishwakarma in Toxic Speech Detection using Traditional Machine Learning Models and BERT and fastText Embedding with Deep Neural Networks, pp. 1254–1259, 2021.

[3] H. Watanabe, M. Bouazizi, and T. Ohtsuki, "Hate speech on twittera pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," IEEE Access, pp. 1–1, 2018.

[4] G. Koushik, K. Rajeswari, and S. K. Muthusamy, "Automated hate speech detection on twitter," in 2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), pp. 1–4, 2019.

[5] H. S¸ ahi, Y. Kılıc¸, and R. B. Sag˘lam, "Automated detection of hate speech towards woman on twitter," in 2018 3rd International Conference on Computer Science and Engineering (UBMK), pp. 533–536, 2018.

[6] S. Abro, S. Shaikh, Z. Hussain, Z. Ali, S. Khan, and G. Mujtaba, "Automatic hate speech detection using

machine learning: A compara- tive study," International Journal of Advanced Computer Science and Applications, 2020.

[7] S. Kokatnoor and B. Dr. K, "Twitter hate speech detection using stacked weighted ensemble (swe) model," pp. 87–92, 11 2020.

[8] B. Pariyani, K. Shah, M. Shah, T. Vyas, and S. Degadwala, "Hate speech detection in twitter using natural language processing," in 2021 Third International Conference on Intelligent Communication Technologies andVirtual Mobile Networks (ICICV), pp. 1146–1152, 2021.

[9] Q. Le and T. Mikolov, "Distributed representations of sentences anddocuments," in Proceedings of the 31st International Conference on Machine LearningLearning - Volume 37, pp. 1188–1196, PMLR, 2014.