IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Cyber-Bullying Detection In Social Media Using BERT Algorithm

[1]Govardhanarao Inampudi

Assistant Professor
Department of Computer Science & Engineering
Osmania University, Hyderabad

***Abstract*:** Social networking sites like Twitter, Tumbler, Face book etc. plays a very important role in present world. Twitter is a micro blogging site that offers millions of data that may be applied to a wide range of Sentiment Analysis applications, including forecasts, reviews, elections, marketing, etc. The technique of extracting information from massive amounts of data and classifying them into several categories known as sentiments is defined as sentiment analysis. The problem of detecting cyberbullying has been the attention of numerous researchers over the past decades because the majority of people nowadays utilize their social networking sites to share offensive content online.

Index Terms - Navie Bayes (NB), Bidirectional Encoder Representations from Transformers (BERT), Random Forest, Decision Tree.

I Introduction:

Twitter is a well-known platform for data exchange that is used frequently prior to, during, and following live events. Online forums are frequently used for abuse and the dissemination of materials that may be humiliating, violent, or otherwise detrimental to people. Users are not allowed to post threatening or harassing content on Twitter that involves violence. Many users continue to violate the regulations and using their social media accounts to disseminate negative sentiments and hate content.

STATEMENT OF THE PROBLEM:

Proposed method work aims to classify the given statements into HOF or NOT.

- **Profanity or offensive language identification:** Model is more concerned with finding abusive or profane posts and posts which include any kind of (untargeted) profanity. There are two categories or classes under which the tweet can be classified.
- **NOT:** -NOT message does not contain insults or profanity. Non-offensive messages do not contain insults or profanity**.**
- **HOF (Hate, Offensive, Profane):** HOF post contains offensive language or intended (veiled or direct) insults. In short, this category includes insults, threats and writings that contain malicious language.

SCOPE:

Proposed work can also extend and can be used on a military basis, for identifying the type of messages that have been sent for communication purposes. In a given data set we can find several abusive words as positive and negative. Based on the score of prediction the tweets will be assigned either 0 or 1. Method group some tweets into dataset and train them for feature extraction.

II. LITERATURE SURVEY:

In paper [1] they detected hate speech based on different aspects including religion. They defined hate speech in their work and then gathered data from Yahoo and American Jews Congress (AJC), where Yahoo provided its data from newsgroups and AJC gave URLs marked as offensive websites. They classified data at paragraph level in their first attempt and then used this data set for annotation by asking annotators to manually annotate the data set. They focused on stereotypes and thus decided to make a language model for stereotypes to mark hate speech.

Motivated by work done in paper [2] proposed a method for detecting hate speech against black over Twitter. They arranged hundreds of tweets to analyze keywords or sentiments indicating hate speeches. To judge the severity of arguments, a questionnaire was floated to students of different races. A training dataset of 24582 tweets was pre-processed to correct spelling variation, remove stop words and eliminate URLs, etc. To classify tweets, NB classifier highlighted racist and nonracist tweets, and prominent features were identified from those tweets. The classifier showed an accuracy of 86%.

Various machine learning approaches have been made to tackle the problem of toxic language. The majority of the approaches deal with feature extraction from the text. Lexical features such as dictionaries and bag-of-words used in some studies. These features were found not to understand the context of the sentences. N-gram-based approaches have also been used and have shown relatively better results.

III. METHODOLOGY:

Since the majority of the research on hate speech that is currently available focuses on a single subject or domain according to the developing or emerging phenomenon in a given era, the proposed technique differs from the prior existing works as it has attempted to concentrate on various topics, including religion, race, ethnicity, and other studies. I have provided a method to concentrate on general issues in this study. Second, unlike the previous works, this strategy makes use of both user profiles and the streaming of real-time tweets.

CHALLENGES AND DIFFICULTIES

- The first and critical step is to learn the numerical representation of text messages, informal language, and emojis, in different languages.
- Secondly, cyberbullying is difficult to explain and choose from a third read because of its intrinsic ambiguities.
- Thirdly, training the model is difficulty. Because of Internet user protection and privacy concerns, a minor portion of messages are only left in the Internet and most of the bullying posts are deleted.

IV. PROPOSED SYSTEM:

Classification is the process of identifying, understanding, and grouping ideas and objects into predetermined categories or subpopulations. Machine learning programs use pre-classified training data sets and various algorithms to classify incoming data sets into classes. In summary, classification is a form of "pattern recognition" where to find the same pattern (similar words or emotions, series of numbers, etc.) in future data sets, classification algorithms are applied to training data.

DESCRIPTION OF APPROACH:

The strategy includes the following steps for NON-BERT models:

Step-1

Creation of data frame.

Data Frame is a mutable and, bi dimensional size, potentially heterogeneous tabular data structure with labeled axes (rows and columns). First step is the creation of the data frames by reading csv file that contains tweet task_1.

Step-2

Then preprocessing these twitter tweets help them for fitting in mining process and extraction of features

Preprocessing involves:

- Removing special characters.
- Tokenization of tweets.
- Removing irrelevant words (stop words) from tweets.

Step-3

After pre-processing, tweets data passed for Sentimental analysis by using VADER sentimental analyzer, which will classify the tweets based on their polarity score values into either positive-1 or negative-0 considering the label name as predicted.

(-1 to 0.5 as negative: 0 and 0.5 to 1 as positive:1)

Step-4

The next sequence of steps involves counting the number of HOF (Hate, Offensive, Profane) words in preprocessed tweets. These counts are stored in the HOF-count column.

Values in predict and values of HOF-count are considered as attributes for further process.

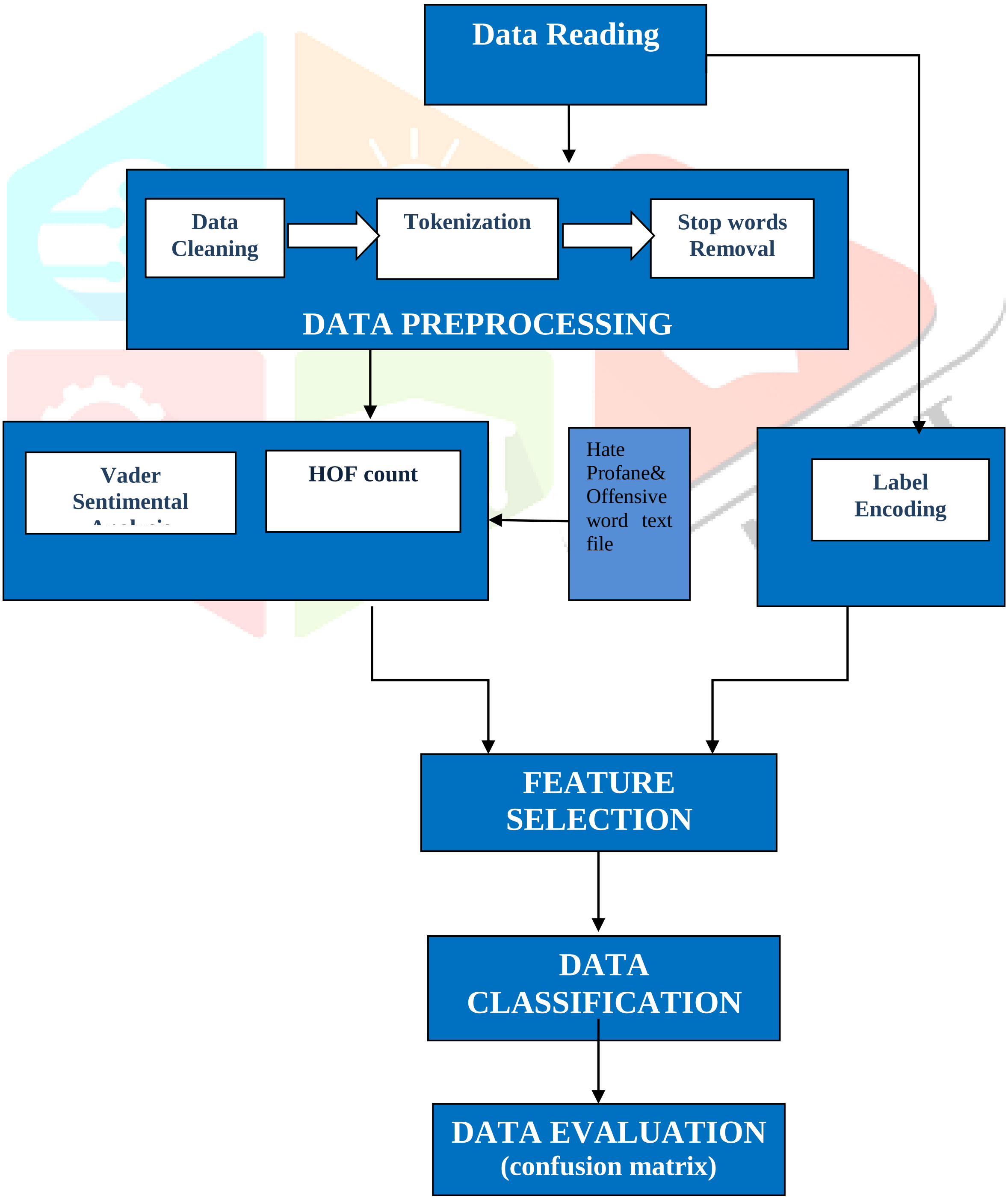
Step-5

Task-1 column in data frames is encoded into either positive or negative (1 or 0 respectively) using a label encoder. These values are considered as target or class labels for the classification process.

Step-6

Input and target are passed for classification in which 80 percent of data is considered the training set and 20 percent is the test set.

Various classification algorithms, for instance, Naïve Bayes, Random Forest, and Decision Tree are applied for calculating the accuracy score which best suits the problem statement.



The strategy includes the following steps for BERT model:

Step-1

Creation of data frame.

Data Frame is a mutable and, bi dimensional size, potentially heterogeneous tabular data structure with labeled axes (rows and columns). First step is the creation of the data frames by reading csv file that contains tweet task_1.

Step-2

Then preprocessing these twitter tweets help them for fitting in mining process and extraction of features

Preprocessing involves:

- Removing special characters.
- Tokenization of tweets.
- Removing irrelevant words (stop words) from tweets.

Step-3

After, preprocessing is done, the next step is to download the BERT preprocessor and encoder for generating the model. Proposed model consists of one dense layer with 1 output unit that will give the probability of tweet being HOF or NOT as the sigmoid function is being used.

Step-4

Task-1 column in data frames is encoded into either positive or negative (1 or 0 respectively) using a label encoder. These values are considered as target or class labels for the classification process.

Step-5

Input and target are passed for classification in which 70 percent of data is considered as training set and 30 percent as testing set.

After running the code above for 10 epochs, an accuracy of 65.07% is achieved from the training dataset. The accuracy that we get can slightly differ due to the randomness of the training process

V. EXPERIMENTATION AND ANALYSIS:

DATA EXPLORATION

Exploring data sets and developing a deep understanding of the data is one of the most important skills every data scientist should possess. These are strong Python libraries for data mining. The goal is to produce a handy reference for certain frequently needed information.

DATA PREPROCESSING

Then preprocessing these twitter tweets help them for fitting in mining process and extraction of features

Preprocessing involves:

- Removing special characters.
- Tokenization of tweets.
- Removing irrelevant words (stop words) from tweets.

VADER SENTIMENT ANALYSIS

After pre-processing I have passed this data for Sentimental analysis by using VADER sentimental analyzer, which will classify the tweets based on their polarity score values into either positive-1 or negative-0 considering the label name as predict.

(-1 to 0.5 as negative:0 and 0.5 to 1 as positive:1).

COUNTING HOF WORDS

The next sequence of step involves counting the number of HOF (Hate, Offensive, Profane) words in preprocessed tweets. These counts are stored in HOF-count column.

Values in predict and values of HOF-count are considered as attributes for further process.

FEATURE EXTRACTION

Input and target are passed for classification in which the training dataset represents 80% of the dataset, and the testing dataset represents 20% of the dataset.

APPLYING CLASSIFICATION ALGORITHMS

Various classification algorithms for instance Naïve Bayes, Random Forest Decision Tree, Logistic Regression and BERT are applied for calculating accuracy score which best suites the problem statement.

VI. RESULT ANALYSIS

Proposed work first gone through the textual preprocessing on dataset. Because it might consist of irrelevant characters or notations which aren't suitable for the detection of the offensive language. In pre-processing several functionalities are to be used such as tokenization, removal of stop words.

After preprocessing I have performed sentimental analysis using Vader sentimental analyzer and we count HOF words from the text file containing hate, offensive, profane words. The next sequence of steps involves feature extraction and application of classification algorithms.

After completion of training and testing on dataset I have calculated the accuracy.

Classifier	Decision tree	Logistic regression	Random forest	Naïve bayes	BERT
Accuracy	0.940334	0.95226	0.940334	0.859188	0.64047

Table 1: Accuracy Comparison Table

From the above comparison table, I can say that Logistic Regression algorithm is best suited classifier with highest accuracy.

VII. CONCLUSION:

Proposed method aim is the evaluation of performance of Cyber Bullying detection algorithms in terms of algorithm accuracy. In proposed method, various machine learning algorithms like Random Forest, Decision Tree, Naïve Bayes, Logistic Regression, and BERT are compared for detection of the cyber bullying tweets in twitter. The result of the experiment shows that the classification yielded best results for the offensive tweets review with the Logistic Regression's method giving 95% accuracy and outperforming other algorithms.

Future research will focus on improving the accuracy through the use of various algorithms and expanding the sample data to also include tweets from various regions that represent various accents and cultures.

REFERENCES:

[1] W. Warner and J. Hirschberg. (2012). Detecting hate speech on the world wide web.Proceeding LSM '12 Proc. Second Work. Lang. Soc.Media, no. Lsm, pp. 19–26.

[2] Rapid Cyber-bullying detection method usingCompact BERT Models by Mitra Behzadi, Ian G. Harris and Ali Derakhshan

[3] I. Kwok and Y. Wang.(2013).Locate the hate: detecting tweets against blacks. Twenty-Seventh AAAI Conf. Artif. Intell., pp. 1621–1622.

[4] Burnap et al. (2013) a rule-based approach to classifying antagonistic content on Twitter.

[5] Ting et al. (2013) proposed architecture for discovering hate groups over Facebook with the help of social network and text mining analysis.

[6] I. Alfina, D. Sigmawaty, F. Nurhidayati, and A. N. Hidayanto.(2017).Utilizing hashtags for sentiment analysis of tweets in the political domain. In Proceedings of the 9th International Conference on Machine Learning and Computing, pp. 43–47.

[7] Freund, Y; Schapire, R.E. (1999) .Large margin classification using the perceptron algorithm. Machine Learning.

[8] Kim, Y.H. et al. (2000) .Text filtering by boosting naive Bayes classifiers. ACM SIGIR Conference: p 168-175.