|| Volume 5 || Special Issue SBPCOE || June 2020 || ISSN (Online) 2456-0774
INTERNATIONAL JOURNAL OF ADVANCE SCIENTIFIC RESEARCH
AND ENGINEERING TRENDS

ABUSIVE WORDS DETECTION USING MACHINE LEARNING FRAMEWORK

Akanksha Gajbahar[1], Suvarna Kolekar[2], Pranjal Chore[3] , Shubham Mandlik[4] , Anita Devkar[5], Sonali Deo[6]

Student, Dept. of Information Technology Engineering, P.E.S's Modern College of Engineering, Pune, Maharashtra, India. [1,2,3,4]
Asst. Professor, Dept. of Information Technology Engineering, P.E.S's Modern College of Engineering, Pune, Maharashtra, India. [5,6]

Abstract: Abusive language is an wording that accommodate abusive words which can be in the context of jokes, or to invoking someone. Nowadays almost every user make use of an abusive language in the social media platform such as Facebook, Linkedin, Instagram, Twitter, etc. It is one of the difficult task to identifying an abusive word in huge world of social media because this problem cannot be determined by simply word matching. With fast growing of social networks and communication between people from different countries and different state of mind has become more direct, which results into using more and more cursing words between these people. Therefore, it arises the need of detecting such speech automatically and modify any data that contains abusive language. In this project, we propose an approach to detect abusive words from reviews and classify the reviews as positive or negative. Our approach is based on unigrams and patterns that are automatically collected from the training set. We use Random Forest Decision Tree classifier to identify the review whether the review is abusive or not.

Keywords:- *Abusive word, Random Forest, Sentiment analysis, Decision tree, Data Pre-processing.*

I INTRODUCTION

Online social networks (OSN) and microblogging websites are attracting internet users more than any other kind of web-site. Services those offered by Twitter, Facebook and Instagram are more and more popular among people from different backgrounds, cultures and interests. Their contents are rapidly growing, constituting a very interesting example of the so-called big data. Big data have been attracting the attention of researcher, who have been interested in the automatic analysis of people's opinions and the structure/distribution of users in the networks, etc[1]. A large number of social media users often leads to uncontrolled communication and many people who communicate with an abusive language. Abusive language is an expression that contains abusive/dirty words or phrases, both oral and text. According to the causes of uncontrolled the use of abusive words in social media are the absence of effective tools to filter abusive language in social media, lack of empathy among citizens, and lack of parental guidance. Abusive language in social media needs to be filtered so that there are no children and adolescents who learn abusive language from the social media that they used. However, it is almost impossible to filter abusive language in social media manually because of a large number of people who write the abusive language. Thus, the abusive language in social media needs to be automatically detected [2]. Detecting an abusive language in social media is a problem difficult to resolve. Detecting an abusive language in social media cannot just use word matching. Moreover, the spelling and grammar from users when speech abusive language in social media is very informal. Especially in short text data, classifying short text data to detect an abusive language is more difficult to resolve. For example in Twitter data, there are a lot of users posting a tweet using abbreviations because of the word limit of a tweet, that some of non-formal words often used by users are: words that show feelings, character repetition to emphasize meaning, using slang words, and changing vowels to numbers[2]. So to overcome this problem we decided to look into this matter by examining which classifier is most suitable for detecting abusive content.

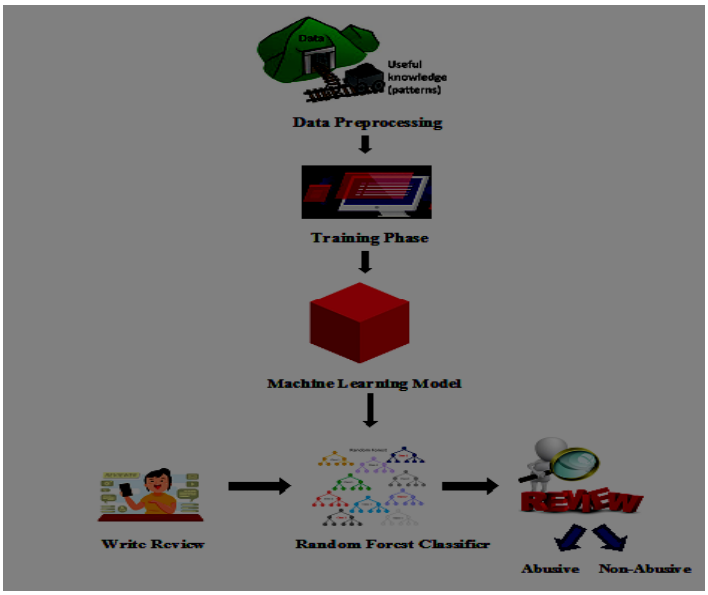
The domain of our project lies under Machine Learning which basically includes detection of abusive text using Random Forest Algorithm obtained from

customer reviews. While most of the online social networks and micro-blogging websites forbid the use of abusive speech, the size of these networks and websites makes it almost impossible to control all of their content. Therefore, arises the necessity to detect such speech automatically and alter any content that presents abusive language. In this project, we propose an approach to detect abusive words on blogging sites .The main purpose of this project is to use one of the machine learning approaches which is better for abusive word detection giving more accurate results for classification.

II IDENTIFY, RESEARCH AND COLLECT DATA

To prevent or dispel the use of abusive words, there are so many work done to look for the strategies or best methods. For example in Reference [2], they propose a pattern-based approach to detect hate speech on Twitter: patterns are extracted in pragmatic way from the training set. Reference[9] includes an automatic invective language detection method which extracts feature and applies classification which is done by neural network. Finally, in Reference [7] They proposed framework that detect cyberbully words from the short hand text and emotions on the comment sections using Latent semantic analysis (LSA). The Cyberbully words will be classified using a Random Decision Forest algorithm. On the basis of Literature Survey we understood that Random Forest Classifier is more suitable and best approach for providing more and correct accuracy to detect abusive word in our project.

III PROPOSED SYSTEM



The goal of our system is to detect abusive words from given reviews and classify the review with associated with the subject. The output shows classification which signifies whether review associated with subject is abusive or non abusive. For our work we employ a Supervised Machine Learning algorithm called "Random Forest algorithm" that works through bagging approach to create a bunch of decision trees with a random subset of the data. Also we choose Naive Bayes Classifier for sentiment analysis to figure out whether the given review is positive or negative. Our system works in multiple steps and uses Random Forest Classifier for detecting abusive content. The system takes reviews as input and pass it to the classifier for classification with the help of machine learning model. To generate model we take the dataset from Kaggle (Toxic comment classification dataset). The dataset consist of total 10,000 instance with comments out of which 50,726 unique words of which their abusive values and corresponding multiple binomial labels – Toxic, Severe toxic, Obscene, Threat, Insult and Identity hate. We perform all the preprocessing tasks (tokenization, stop word removal and stemming) on the dataset on which our training model of Random Forest works. After building model, user can give the input data to the system and as a result system provides classification result as abusive or not. Every review is divided into two classes positive and negative, on the basis of the sentiment they exhibit.

IV TECHNOLOGIES USED IN PROPOSED SYSTEM

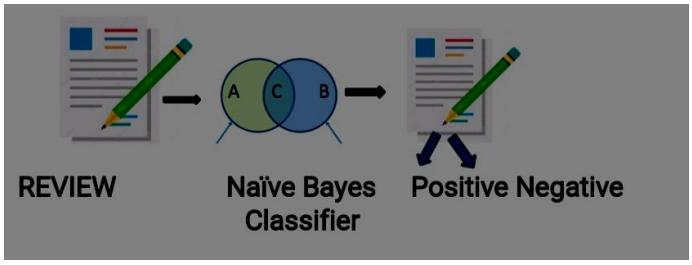
1.Random Forest-

Random forest is a supervised learning algorithm which is use for classification problems. Random forest algorithm creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by majority of voting. While working with random forest in first step, we create bootstrap dataset out of original dataset. Bootstrap dataset means shuffling of records, removal of duplicates and creating samples. In second step, we prepare decision tree from the bootstrap dataset. At every decision node, consider a random subset of elements (i.e. if we have 4 features consider only 2 features as candidates of decision). Best candidate of decision splits the dataset correctly. We use gini index and information gain method to select features from the n features of the gini index. Gini index is a

metric to measure how often a randomly chosen element will be incorrectly identified (means attribute with lower gini index should be prefer). Now, repeat above steps to generate huge number of decision trees. In next step, feed the test data to all the decision trees and record their outputs. The output given by maximum number of decision trees is the final output. We make use of Clf random classification object. Each word is search on the dataset and extract the label. After that we store attributes and object to send data for prediction which is inbuilt. Prediction which holds the output is return to the classify shows the final answer as 0 or 1.

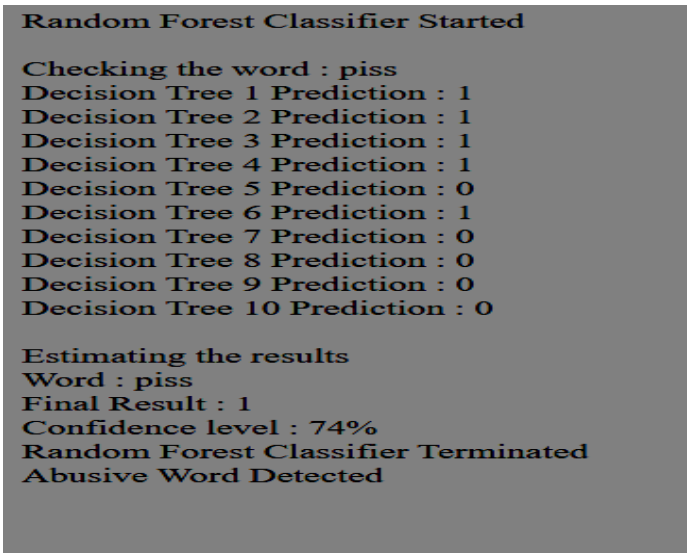
2. Naive Bayes -

Naïve Bayes is a probabilistic machine learning algorithm based on the Bayes theorem which we use in our system to classify the review as positive or negative. In our system ech word present in given review is an attribute itself. With the help of naive bayes algorithm we check the probability of each attribute and according to TRUE, FALSE count provide the final output. A Naive Bayesian model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. Use Python Package pickle to save the trained model in order to reduce the time taken for training it for every review.

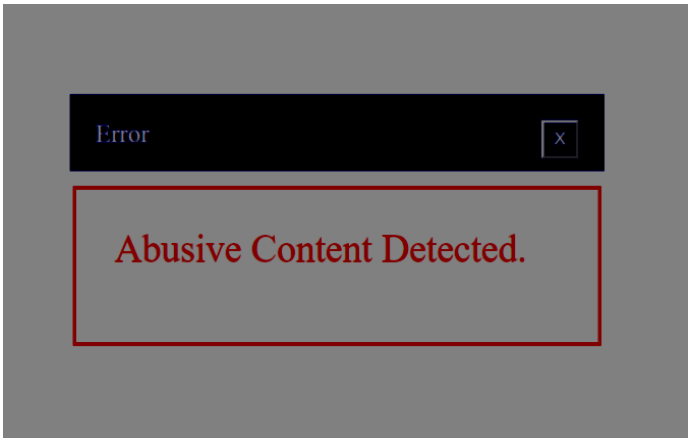


V RESULTS AND EVALUATION

Before finalizing the model, it is important to see how the model performs by using the validation set. This will give an indication whether the parameters needs to be tuned in for a better performance. For the classification problem it is essential to choose correct and appropriate datasets also the right choice of classifier to work the model in excellent way. Here after going to through the results and working of all the classifier such as support vector machine , NN classification, Natural language processing we found Random Forest Classifier of machine learning is work better among them with high rate of accuracy as we want. Hence we decided to use RF classifier for our project. we mentioned early, we selected 50726 unique words from original dataset with their abusive values were extracted. We used random forest for sentiment analysis before finalizing the naive Bayes classifier , due to some accuracy issues like 60% which so worst hence we decide to go for naive bays after doing lot of studies now with naive byes we are getting 73% accuracy performance. Below digram shows the final output of our project with prediction values 1(contain an abusive word) and 0 (not contain an abusive word) and confidence level.



If any single word get detected as an abusive word user will get following notification and due to containing an abusive word user unable to post their review.



VI CONCLUSION

In this Paper, We proposed a approach of combining two classifiers – one to classify an abusive language and another for sentiment analysis. Machine Learning representation and authentication make automatic revelation of abusive messages in online media

possible and ensures building a relevant and clear social media environment. This proposed system mainly concentrated on identifying the presence of abusive word in the blogging site platform using Random Forest Classifier. This system also uses the Naive Bayes classifier for sentiment analysis which helps to identify whether the comments are positive or negative. This study might help in recognizing abusive content on social media.

REFERENCES

[1]Hajime Watanabe, Mondher Bouazizi , and Tomoaki Ohtsuki "Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection".February 15, 2018

[2] Muhammad Okky Ibrohim, Indra Budi, A Dataset and Preliminaries Study for "Abusive Language Detection in Indonesian Social Media",Procedia Computer Science 135 (2018) 222–229.

[3] Jezabel Molina-Gil, José A. Concepción-Sánchez and Pino Caballero- Gil,"Harassment Detection Using Machine Learning and Fuzzy Logic Techniques" 20 November 2019.

[4] Navoneel Chakrabarty, "A Machine Learning Approach to Comment Toxicity Classification",2020.

[5] Rehanulla Khan,Nasir Ahmad,"Random Forests and Decision Trees",September,2012.

[6] Prakhyat Rai, Shamantha Rai, Sweekriti Shetty, "Sentiment Analysis Using Machine Learning Classifiers: Evaluation of Performance",pp.IEEE 2019.

[7] J.I.Sheeba, S. Pradeep Devaneyan, Revathy Cadiravane, "Identification and Classification of Cyberbully Incidents using Bystander Intervention Model" International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277- 3878, Volume-8 Issue-2S4, July 2019.

[8] Dinesh Kannan,Rajkumar Murukeshan ,"Online Abuse Detection",2018.

[9] Aishwarya Ganesan,"Offensive language detection using AI technique ".April 2018.

****

Shahajirao Patil Vikas Pratishthan's

S. B. PATIL COLLEGE OF ENGINEERING,

Indapur, Pune – 413106

Organized

An E-National Conference on

"SCIENCE AND TECHNOLOGY"

2K20

on **15th and 16th June 2020**

****