iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VII **Month of publication:** July 2021

DOI: https://doi.org/10.22214/ijraset.2021.35770

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

Text Analyzer - An Approach for Hate Speech & Offensive Language Detection

Dr Sweeta Bansal[1], Sonali Agrahari[2], Siddhant Srivastava[3], Shubham Srivastava[4], Yash Chandel[5]
[1]*Assistant Professor,* [2, 3, 4, 5]*Student, Inderprastha Engineering College, Ghaziabad, India*

Abstract: ***As we know that the social crowd is increasing day by day, so is the hatred among them online. This hatred gives rise to hate speech/comments that are passed to one another online. Recently, the hate speech has increased so much that we need a way to stop them or at least contain it to minimum. Keeping this problem in mind, we have introduced a way in which we can detect the class of comments that are posted online and stop its spread if it belongs to hateful category. We have used Natural Language Processing methods and Logistic Regression algorithm to achieve our goal.***

I. INTRODUCTION

Hate speech, due to its explosion in the social world, needs to be contained. Hate speech can hurt people's feelings with respect to caste, color, religion, gender, etc. Hence, we need to stop it as it leads to inter-personal hatred, hatred among communities, societies, religions, nations, etc. Hate speech creates an environment of prejudice and intolerance, which then leads to discrimination among us and may also lead to any violent act, sarcasm, impoliteness, vulgarity, etc. Hence, it is our duty to avoid such circumstances by whatever means necessary/possible.

II. LITERATURE REVIEW

We all are aware of the great increase in the number of users on online social media platforms. According to the recent data Facebook is leading with over 2700 million users followed by YouTube, WhatsApp, Twitter and Instagram. Cumulative count of active users over these social media sites is over 10 billion implying the vast use of social media sites in daily lifestyle. Messages and information on through these social media platforms can be accessed and transferred quite quickly and easily. Though this may come as a great advantage for social media platforms, it also creates a great disadvantage for social media sites as the hateful and offensive messages can also be spread easily and quickly creating conflicts between individuals and society.

So, what exactly is Hate Comments or Speech? Hate Speech is any type of communication towards a person which may hurt his feelings. These communications can involve various traits of that person like physique, likeness, mental health, caste, color, religion, etc. Our project can be used in a system which would flag the texts or comments by a user under the categories: Hateful, Offensive and Clean. If the user attempts to comment any offensive text, then the system will prompt the user with a warning, otherwise the comment is good to be posted.

In present scenario the attacks on a person's color, race and religion through offensive comments are the most reported crime in hate speech on social media. Controlling cyber-crimes has always been a big issue and now with growing number of offensive and hateful comments on social media platforms, there is a great need for the development of autonomous hate speech detection system. These systems can be developed by using Machine Learning approach. Data required for development can be accessed through different social media platforms. In our project we have discussed the performance of Logistic Regression Algorithm.

However, in our work, we have used Logistic regression on two different types of Feature Extraction techniques namely, Count Vectorizer, which is also known as Bag of Words (BOG) and TFIDF Vectorizer (Term Frequency – Inverse Document Frequency).

III. RESEARCH METHODOLOGY

Now, it's time to discuss about the dataset and the procedure of conducting hate speech detection.

A. Getting the Dataset

It is the process by which we have collected our data and annotating it.

1) Data Collection

We have collected our dataset from the website mentioned below:

Forum-Dataset: https://github.com/t-davidson/hate-speech-and-offensive-language/blob/master/data/labeled_data.csv

2) Data Annotation: In this, each of the sample/observation/tweets will be labelled whether it is a hateful comment or not.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

B. Hate Speech Detection

The hate speech detection process comprises of following stages:

1) Pre-processing of the dataset.
2) Feature extraction of our data.
3) Classification and evaluation of training and testing data.
4) Testing the model on the input provided by the user.

a) Pre-processing

We adopted the pre-processing method used with little modification.
There are six steps in the pre-processing stage, i.e.

a) Cleaning the text
b) Converting all the text in lower case
c) Removal of punctuations
d) Removal of links
e) Removal of stop words based on the language
f) Removal of extra spaces
g) Tokenization of sentences
h) Stemming of sentences

b) Features Extraction

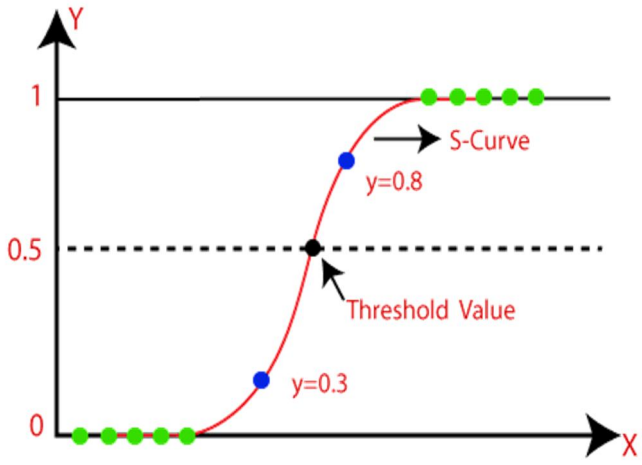
For the sake of simplicity, we have kept things as simple as we can as long as it doesn't cost much on our results.
Firstly, we will use Count Vectorizer for representing the text in the form of vectors.
We are keeping max features set to 10000.
Secondly, we will use TF-IDF vectorizer and here too we are keeping max features set to 10000 here too.

c) Classification and Evaluation

We will be using supervised learning for detection of hate speech. We have used Logistic Regression algorithm with maximum iterations set to 1000, and given vectorized inputs to it. We will use jupyter Notebook (Python3) to conduct the experiments.
Both, Bag of words and TFIDF vectors are given to the machine learning algorithm, which in our case is Logistic Regression.

d) Logistic Regression

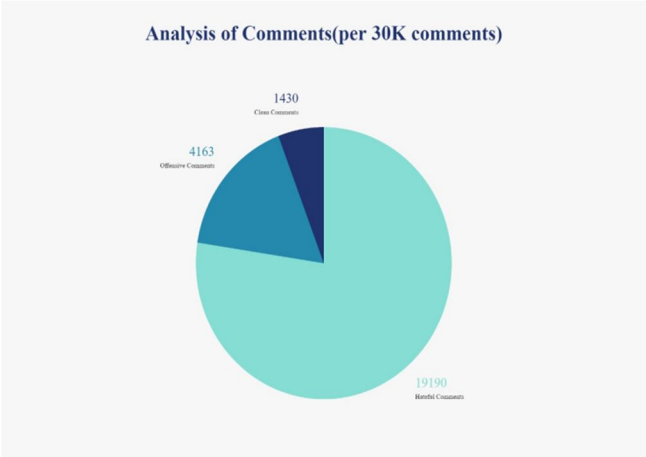
A simple overview:
Logistic Regression, as others, is also a Machine Learning algorithm, widely used for classification problems. It uses predictive techniques and tells the result in the form of probability. A threshold value is given and on the basis of that threshold value, the class of the prediction is defined.

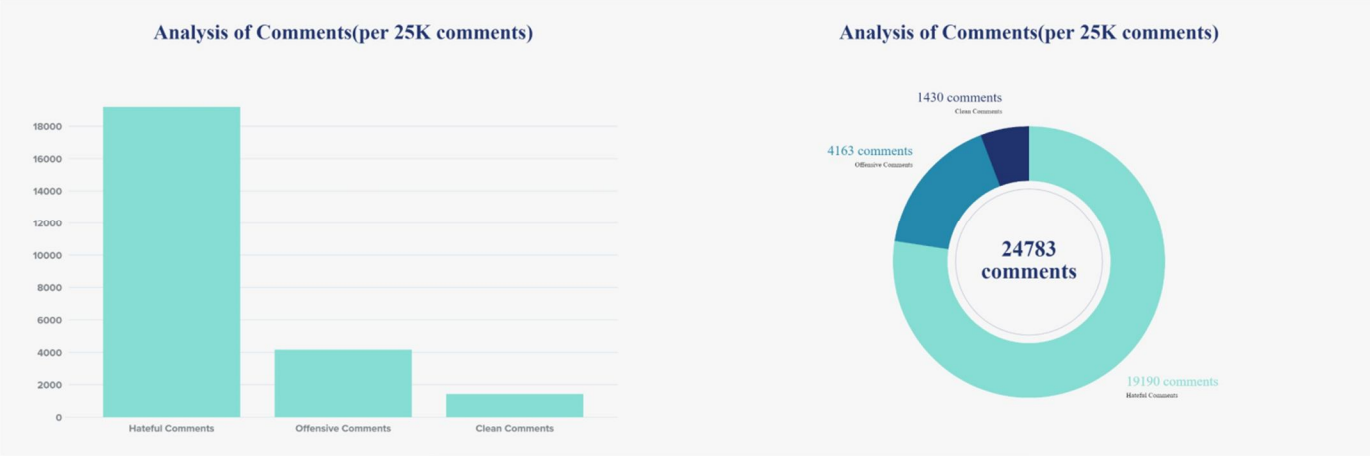


International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VII July 2021- Available at www.ijraset.com

IV. ANALYSIS

Here is some analysis of the dataset we will work on:



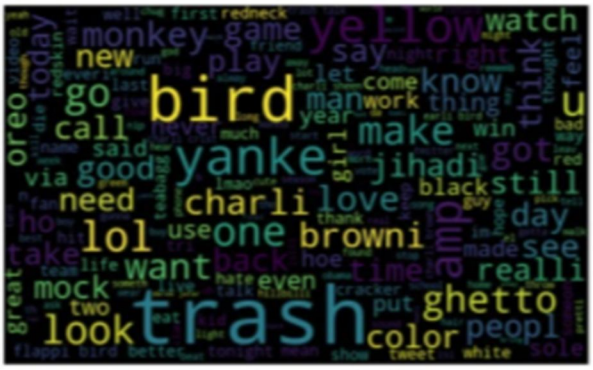



```
In [14]: #count of different classes:0--> hateful, 1--> offensive, 2--> clean
         dataset['class'].value_counts()

Out[14]: 1    19190
         2     4163
         0     1430
         Name: class, dtype: int64
```

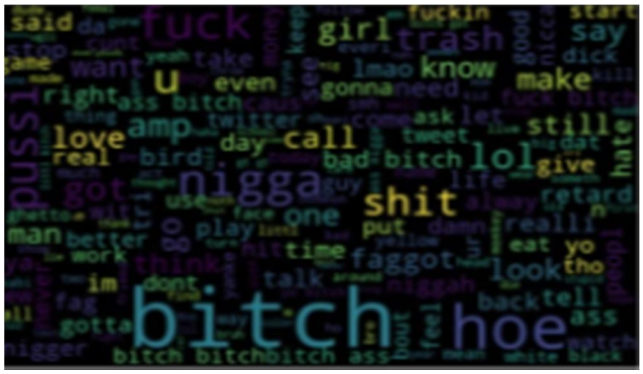
Let's check the most common words used in the dataset using Word Cloud.

1) Here are the most common words used in the dataset with class label as "2 (Clean Comment):



2) Here are the most common words used in the dataset with class label as "0 (Hateful Comment):

3) Here are the most common words used in the dataset with class label as "1 (Offensive Comment):



V. RESULTS

We have collected datasets from various sources, then we pre-processed the data. After pre-processing we applied the machine learning algorithm that is Logistic Regression. After applying the algorithm on BOW model, we got the desired result with accuracy of 82.17%, which then was improved with an increment of 2.7%, to 84.87% when we used TF-DIF.

This project also allows to take the user input, and predict its class immediately. Because of which, we can use it in a real-life web or mobile applications. Hence, this way we can detect user's mindset before posting the comments and prevent any future conflict as a result of hate or offensive speech.

Below is the result when we used Count Vectorizer (Bag of Words) with Logistic Regression.

	precision	recall	f1-score	support
0	0.31	0.65	0.42	164
1	0.98	0.81	0.89	1905
2	0.70	0.95	0.81	410
accuracy			0.82	2479
macro avg	0.66	0.80	0.71	2479
weighted avg	0.89	0.82	0.84	2479

Accuracy Score: 0.8229124647035094

Below is the result when we used TF-IDF (Term Frequency – Inverse Document Frequency).

	precision	recall	f1-score	support
0	0.39	0.61	0.48	164
1	0.97	0.85	0.91	1905
2	0.69	0.95	0.80	410
accuracy			0.85	2479
macro avg	0.68	0.80	0.73	2479
weighted avg	0.89	0.85	0.86	2479

Accuracy Score: 0.8487293263412666

A. User Input

Below are the results with user input:

VII. FUTURE SCOPE

A stand-alone keyboard app for smartphones we can be build which can detect and warn users for their hateful post just before they post it online.

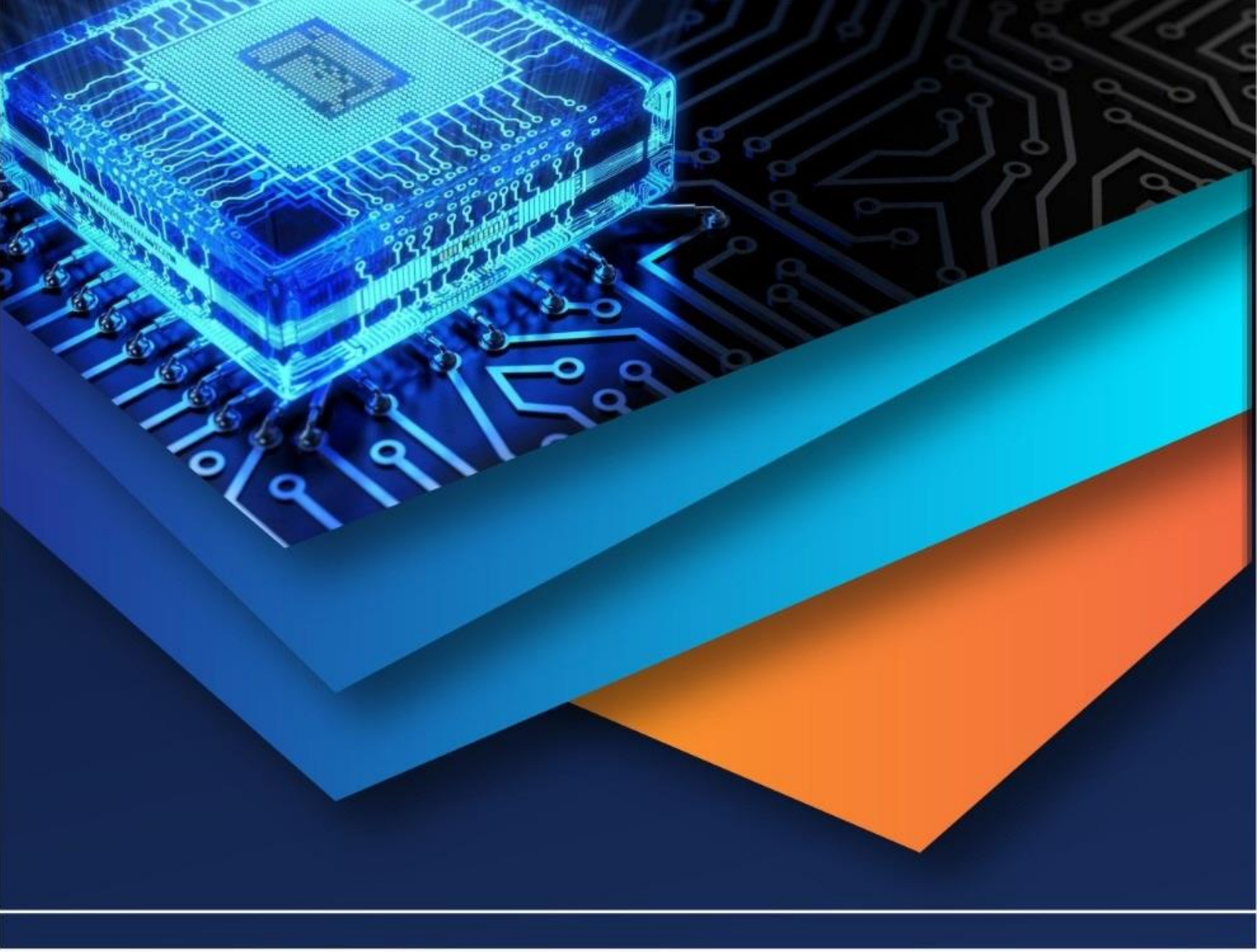
Integrating the model in any social app or a website which will detect any kind of harsh comments and warn user for their behaviour otherwise suffering a penalty.

VIII. ACKNOWLEDGEMENT

We would like to thank our project guide Prof. Sweeta Bansal for her enormous cooperation and guidance which helped us to develop a very good project idea. We have no words to express our gratitude towards the person who constantly and whole-heartedly supported us throughout the project. She has always given us her precious time and knowledge. The technical guidance provided by her was more than useful and made this project possible. She has always been an inspiration for us. We are also thankful to our HOD Dr Vijay Singh and all the members of the computer department for providing various facilities and support.

REFERENCES

[1] https://towardsdatascience.com/natural-language-processing-on-multiple-columns-in-python-554043e05308
[2] https://apoorva18.github.io/proj3.pdf



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY