IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

Offensive Text Detection Using Classical Ai/Ml Techniques

Nandini Tidke[1], Komal Chopade[2], Gaurav Kalsait[3], Jai Singh[4], Prof. Santosh Biradar[5]
Students, Department of Computer Engineering[1, 2,3,4]
Assistant Professor, Department of Computer Engineering[5]
D. Y. Patil College of Engineering, Pune, Maharashtra, India

Abstract: *Due to the substantial growth of internet users and its spontaneous access via electronic devices, the amount of electronic contents has been growing enormously in recent years through instant messaging, social networking posts, blogs, online portals and other digital platforms. Unfortunately, the misapplication of technologies has increased with this rapid growth of online content, which leads to the rise in suspicious activities. People misuse the web media to disseminate malicious activity, perform the illegal movement, abuse other people, and publicize suspicious contents on the web. The suspicious contents usually available in the form of text, audio, or video, whereas text contents have been used in most of the cases to perform suspicious activities. Thus, one of the most challenging issues for NLP researchers is to develop a system that can identify suspicious text efficiently from the specific contents. We define this task as being able to classify a tweet as offensive or not. A set of ML classifiers with various features has been used on our developed corpus, consisting of 7000 English text documents where 5600 documents used for training and 1400 documents used for testing. The performance of the proposed system is compared with the human baseline and existing ML techniques.*

Keywords: Test detection

I. Introduction

The implemented system focuses on creating a system for classifying the tweets and creating awareness among the general public about the negativity on social media platforms. Allowing the general public to check their sentiments by using publicly available system and machine learning models, thereby achieving the goal at a cheaper cost.

1.1 Motivation

Increase in the usage of social media sites likes Facebook and Twitter have given the crowd a great platform to express their opinions/feelings for the individual, groups or events happening around them or in society. This digital media has become a great resource to share the information and also gives the full freedom of speech to every one on the platform.

With the gaining popularity of these platforms; there also comes the negative part along with its benefits. This feature of the social media to express something openly to the world have created the major problems for these online businesses and negatively impacted the well being of the societal decorum. There are increasing cases of the abuse or offense on the social media like Hate speech, Cyberbullying, Aggression or general Profanity. It is very much important to understand that this behaviour can not only immensely affect the life of an individual or a group.

So in the last, considering the importance and the sensitivity of this particular topic in the today's digital world there is, still a lot of emerging further scope in tackling and improvise on the previous work done in the field with the help of these new age AI and NLP techniques. Therefore we aim to contribute in this particular direction.

II. Description Of The Problem

Now a days with the increasing usage of social media platforms like Facebook and Twitter, we often see people to misuse this freedom of speech. Some of them try to use this platform to post the offensive or abusive things about a person or a group. This in turn can negatively impact the mental health of a community or an individual.

Considering the sensitivity of the topic, we aim to tackle this problem of detecting. the offensive language in social media through cutting edge techniques in Machine Learning, Deep Learning Natural Language Processing. To proceed further

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

with our topic we choose the Offensive Language Identification Dataset which was released in 2021. There are three sub-tasks for this new dataset which involved the detection, predicting the type the target of the offensive tweet respectively. We particularly focus on predicting whether the post is offensive or not which is the first and the most crucial sub-task for this dataset making it a binary classification problem.

2.1 Project Scope

Sentiment analysis, also known as opinion mining, is used to understand the emotions of the individual, more often on social media. In this project we aim to determine the feelings and emotions of a person by analyzing the information posted by him on social media platforms. By using machine learning techniques and natural language processing we are going to extract required features from the data and classify the emotions of the individual according to its polarity as positive, negative or neutral. This information will be very useful to determine whether an individual is in depression or not.

- **Ambiguity of natural language:** Ambiguity is a basic characteristic of human conversations and one that is particularly challenging in natural language understanding (NLU) situations. By ambiguity, we are in actual fact referring to sentences that have multiple alternative interpretations. This type of ambiguity denotes words that can have numerous assertions.
- **Sarcasm Recognition:** In sarcastic text/comments, people express their negative sentiments using positive words. Thus sarcasm makes it difficult to identify the polarity of the text. It easily deceives sentiment analysis models. Sarcasm is difficult to understand for machines and thus it becomes impossible for the system to classify the offensive sentiment.

III. METHODOLOGY

3.1 Data Cleaning

The preprocessing of the text data is an important step because it makes the raw text ready for mining, i.e., it becomes easier to extract information from the text and apply machine learning algorithms to it. If we skip this step then there's a better chance that you simply are working with noisy and inconsistent data. The objective of this step is to wash noise those are less relevant to seek out the sentiment of tweets like punctuation, special characters, numbers, and terms which don't carry much weight age in context to the text. In one among the later stages, we'll be extracting numeric features from our Twitter text data. This feature space is made using all the unique words present within the entire data. So, if we preprocess our data well, then we might be ready to get a far better quality feature space.

Removing Twitter Handles (@user), the tweets contain lots of twitter handles (@user) that is how a Twitter user acknowledged on Twitter. We will remove of these twitter handles from the info as they don't convey much information. Removing Punctuation's, Numbers, and Special Characters As discussed, punctuation's, numbers and special characters do not help much. It is better to get rid of them from the text even as we removed the twitter handles. Here we'll replace everything except characters and hash tags with spaces.

Removing Short Words we've to be a touch careful here in selecting the length of the words which we would like to get rid of. So, I even have decided to get rid of all the words having length 3 or less. For example, terms like "hmm", "oh" are of little or no use. It is better to get rid of them.

- **Tokenization:** Now we'll tokenize all the cleaned tweets in our dataset. Tokens are individual terms or words, and tokenization is that the process of splitting a string of text into tokens.
- **Stemming:** In stemming the suffixes are stripped or trimmed, based upon rules,("ing", "ly", "es", "s" etc) from a word. For example – "play", "player", "played", "plays" and "playing" are the various words which may have a root word - "play".

3.2 Story Generation and Visualization from Tweets

Exploring and visualizing data, no matter whether its text or the opposite data, is a crucial step in gaining insights. Before we start exploration, we must think and ask questions related to the data in hand a couple of probable questions. Understanding the common words utilized in the tweets: Word Cloud, now I would like to ascertain how well the given sentiments are distributed across the train dataset. A method to accomplish this task is by understanding the common words

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

by plotting word clouds. A word cloud could also be a visualization wherein the foremost frequent words appear in large size and thus the less frequent words appear in smaller sizes.

3.3 Feature Extraction

Extracting Features from Cleaned Tweets to research a preprocessed data, it must be converted into features. Depending upon the usage, text features are often constructed using assorted techniques – Bag-of-Words, TF-IDF, and Word Embedding. Bag-of-Words Features: The Bag of Words (BoW) model is that the simplest sort of text representation in numbers, just like the term itself, we'll represent a sentence as a bag of words vector (a string of numbers).

- **Step 1:** Determine the Vocabulary we first define our vocabulary, which is that the set of all words found in our document set.
- **Step 2:** Count to vectorize our documents, all we've to try to is count what percentage times each word appears. TF-IDF Features this is often another method which is based on the frequency method but it's different to the bag-of-words approach within the sense that it takes into account, not just the occurrence of a word during one document (or tweet) but within the whole corpus. TF-IDF works by penalizing the common words by assigning them lower weights while giving importance to words which are rare within the whole corpus but appear in good numbers in few documents.

IV. Literature Review

Warner and Hirschberg detected hate speech on the basis of 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 news groups and AJC gave URL 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 stereotype and thus decided to make language model for stereotypes to mark hate speech. They made an antisemitic speech classifier first. They identified 9000 paragraphs matching to their regular expression and then removed those paragraphs that were not offensive. Then further seven categories were chosen to annotate the data. After this annotation for their gold corpus, they used two fold cross validation classifier to find a refined data set.

Motivated by work done in Kwok and Wang, (2013) proposed a method for detecting hatred 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 prepossessed to correct spelling variation, remove stop words and eliminate URL etc. In order to classify tweets, NB classifier highlighted racist and non-racist tweets and prominent feature were identified from those tweets. The classifier showed an accuracy of 86%.

Burnap et al. (2013) developed a rule-based approach to classifying antagonistic content on Twitter and they used associational terms as features. They also included accusation and attributional terms targeted at a person or persons following a socially disruptive event as features, in an effort to capture the context of the term use. Their results demonstrated an improvement on standard learning technique.

Ting et al. (2013) proposed architecture for discovering hate groups over Facebook with the help of social network and text mining analysis. They extracted features including keywords that are frequently used in groups.

Sureka et al. (2012) proposed an approach based upon the data mining and social network analysis for discovering hate promoting videos, users and their hidden communities on YouTube.

V. System Design And Flow

The project is efficiently divided into four modules such as Data Extraction, Data Preprocessing and Feature Extraction, Training data and Results.

- **Data Extraction:** Users can post anything on social media platforms (Twitter) in the form of text, images and videos. In this module the data is extracted from Twitter in the form of text, images and videos and are stored on a local disk. Required text is extracted from images and videos and that textual data is stored on a local disk.
- **Data Preprocessing and Feature Extraction:** Data is cleaned and transformed in this module. During the data cleaning process punctuation symbols, URLs, numbers, stop words are removed and words are converted to

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

lowercase and to their root meaning. Necessary features are extracted from the data after data preprocessing by transforming the data into a vector matrix on the basis of frequency count, represented by a column of matrix.

- **Training Data:** After the necessary features are extracted from the data, the data is divided into training and testing parts. The model is trained by Naive Bayes algorithm and Random forest algorithms. After training, testing can be done on the data to achieve the accuracy.
- **Results:** The results are available as offensive sentiment polarities such as Positive, Negative or Neutral.

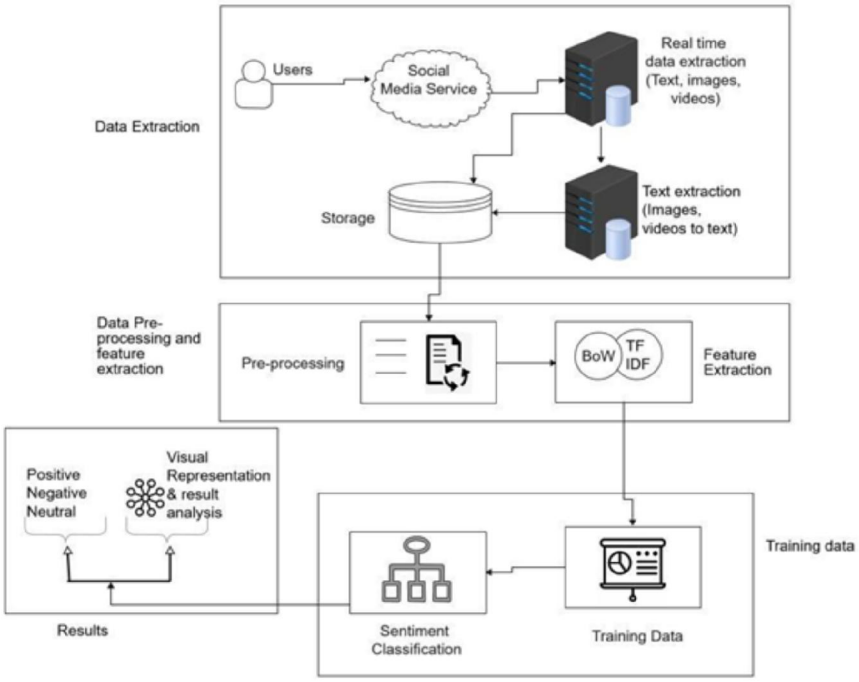

Figure 1: System Architecture

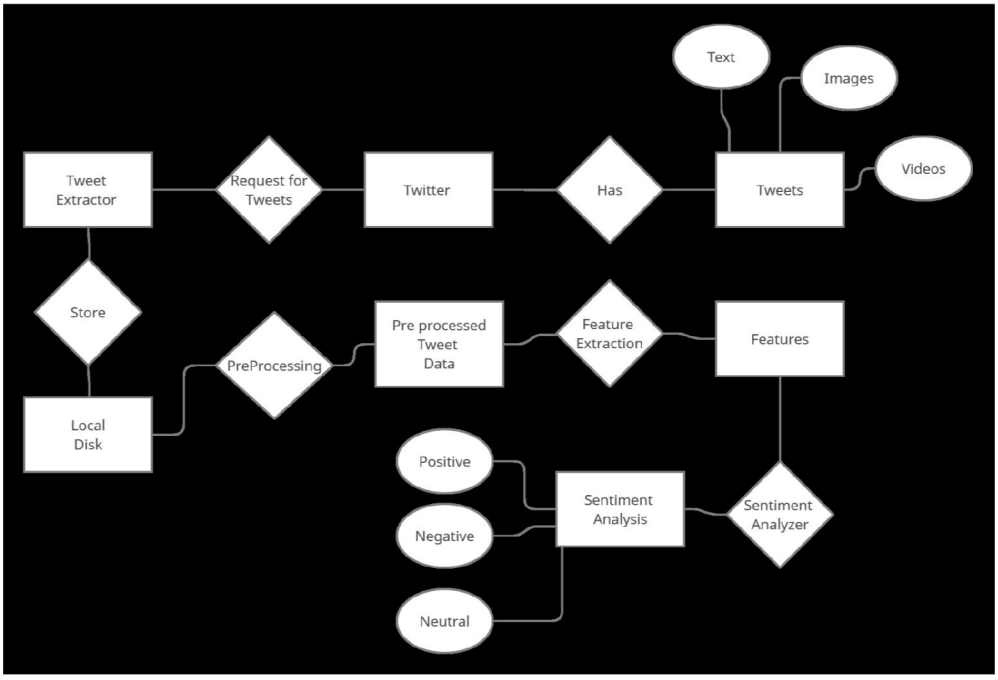

Figure 2: ERP Diagram

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

VI. Project Implementation

6.1 Algorithm Details

A. Naive Bayes

Naive Bayes classifiers are a gaggle of classification algorithms supported Bayes' Theorem. It is not one algorithm but a family of algorithms where all of them share a standard principle. Naive Bayes algorithms are mostly utilized in sentiment analysis, spam filtering, recommendation systems etc. They are fast and straightforward to implement but their biggest disadvantage is that the need of predictors to be independent. In most of the important life cases, the predictors are dependent, this hinders the performance of the classifier.

Multinomial Naive Bayes: Feature vectors represent the frequencies with which certain events are generated by a multinomial distribution.

B. Random Forest Classifier

Random forest, like its name implies, consists of an outsized number of individual decision trees that operate as an ensemble. Each individual tree within the random forest spits out a category prediction and thus the category with the foremost votes becomes our model's prediction. The low correlation between models is that key.

The rationale for this excellent effect is that the trees protect one another from their individual errors (as long as they don't constantly all err within the same direction). While some trees could even be wrong, many other trees are becoming to be right, so as a gaggle the trees are ready to move within the proper direction.

Therefore the prerequisites for random forest to perform well are: There must be some actual signal in our features so as that models built using those features do better than random guessing. The predictions (and therefore the errors) made by the individual trees got to have low correlations with one another.

Figure 3: Word Cloud - Positive Words

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

Figure 4:- Tree Map - Positive Words

Figure 5:- Word Cloud - Negative Words

Figure 6:- Tree Map - Negative Words

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

Figure 7:- Word Cloud - Neutral Words

Figure 8:- Tree Map - Neutral Words

```
[58]: #Test the pipeline with a sample tweet
      pipeline.predict(['killer'])

[58]: array([0])
```

Figure 9:- Model Predictions

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

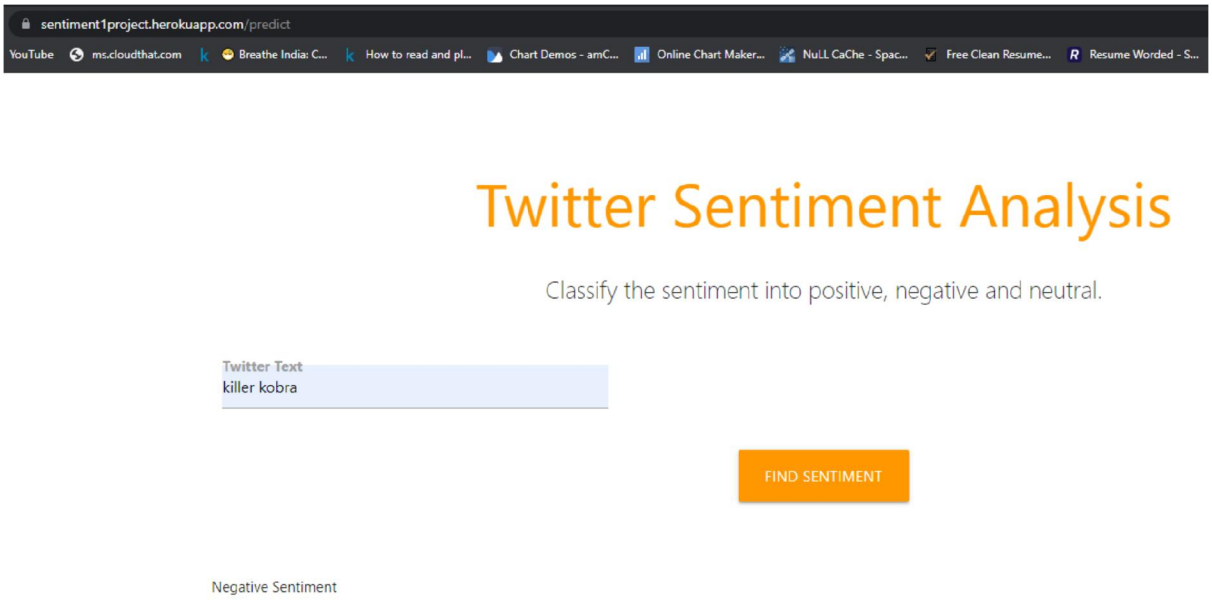

Figure 10: Sentiment Analysis - Model Deployment Heroku

VII. Advantages & Disadvantages

7.1 Advantages

- Uses freely available data.
- Saves man power on data analysis.
- Extended with other platform.

7.2 Disadvantages

- Heavily depend on social media.
- Multiple words can have same meaning.

VIII. Conclusion & Future Work

The aim of the study was to evaluate the performance for sentiment classification in terms of accuracy, precision and recall. We compared various supervised machine learning algorithms of Naive Bayes' for sentiment analysis and detection of the hate tweets in twitter. This provides us with an interesting insight into the usage pattern of hate-mongers in terms of how they express bigotry, racism and propaganda. The proposed project takes freely available twitter data and finds the sentiment and depression levels from the data. Model is deployed on the web so other users can test their data for sentiment analysis.

Attempt to increase the accuracy of the model. We have only focused on the sub-task A that is the detection of offensive language for this research project. But on the same data set there are two more sub-tasks A and B which focused on determining the type and the target of the offensive post. This can even lead us to further insights and analysis of offensive posts detected on the social media to take the relevant actions. In future, we would like to solve these both sub-tasks using variety of approaches to contribute more in the domain.

Human language is very much diverse and certain posts may not look offensive from the surface, but actually they are when analysed by the human annotator. In future we will like to conduct an analysis to capture the syntactic and semantic features along with their combination and other pre-trained features.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 3, May 2022
Impact Factor: 6.252

ACKNOWLEDGEMENTS

The completion of our project brings with it a sense of satisfaction, but it is never complete without those people who made it possible and whose constant support has crowned our efforts with success. One cannot even imagine our completion of the project without guidance and neither can we succeed without acknowledging it. It is a great pleasure that we acknowledge the enormous assistance and excellent co-operation to us by the respected personalities.

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

[3]. 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.

[4]. Freund, Y; Schapire, R.E.(1999). Large margin classification using the perceptron algorithm. Machine Learning, 37(3):277–296.

[5]. Kim, Y.H. et al. (2000). Text filtering by boosting naive Bayes classifiers. ACM SIGIR Conference:p168-175.

[6]. Parikh R, Movassate M. (2009). Sentiment analysis of user-generated Twitter updates using various classification techniques. CS224N Final Report;pages. 1–18.

[7]. Pak A, Paroubek P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In: LREC. vol. 10; pages. 1320–1326.

[8]. Gaudette L, Japkowicz N. (2009). Evaluation methods for ordinal classification.In Advances in Artificial Intelligence. Springer; p. 207–210.

[9]. Go A, Bhayani R, Huang L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford;p. 1–12