Detection of abusive comments

Chitransh Bose

IIITD

Delhi Okhla Industrial Estate Phase III, New Delhi, India, 110020

[chitransh22096@iiitd.ac.in](mailto:chitransh22096@iiitd.ac.in)

Varun Jhunjhunwala

IIITD

Delhi Okhla Industrial Estate Phase III, New Delhi, India, 110020

[varun22087@iiitd.ac.in](mailto:varun22087@iiitd.ac.in)

Ranit Pal

IIITD

Delhi Okhla Industrial Estate Phase III, New Delhi, India, 110020

[ranit22119@iiitd.ac.in](mailto:ranit22119@iiitd.ac.in)

Soumyajyoti DAS

IIITD

Delhi Okhla Industrial Estate Phase III, New Delhi, India, 110020

[soumyajyoti22075@iiitd.ac.in](mailto:soumyajyoti22075@iiitd.ac.in)

**Abstract**

In today’s world where social media is taking an edge over any other source of information and communication, the need for putting appropriate restrictions over it has also increased significantly. People of all ages whether children or elderly people all have the same access to social media sites and micro- blogging sites thus there is a necessity to filter out the comments appropriately. Thus, we are developing a machine learning based model for classifying that if comments done on Twitter are abusive or not. For this we have used various machine learning models and have procured an F1- Score of accuracy of 97.58% using Logistic Regression, 97.66% using SVM and 97.64% using Voting Classifier that is an amalgamation of Logistic regression and SVM~~.~~

**Introduction:**

With the increase in the use of social media sites like Facebook, Instagram and micro-blogging sites like Twitter people are now have more ways to express their views and ideas and have encouraged mass communication significantly. But not all users use the technology in an appropriate ways and use it to express their anger using abusive and derogatory comments. Thus, there should be a mechanism using which we can detect such comments and filter out the abusive comments. Thus, we have proposed a machine learning based model that detects if a comment done is abusive or not. In our work, we have used different machine learning models in order to obtain the model which can determine the comments as abusive and non-abusive efficiently. For this we have used “Hate speech and offensive language dataset” [ 1] consisting of 24,782 samples and 7 features out of which we have extracted features relevant to our work and had modified the labels as per the requirement. This dataset consists of comments obtained from Twitter data and is used to classify the data as hate-speech, offensive and neither. We have modified this dataset as the comments having offensive and non-offensive language and have removed the samples that were hate-speeches.

**Literature Survey:**

Determining if a comment on a post is abusive or not is very tedious and difficult tasks and many of the social media sites and platforms are still working to get more accurate and efficient ways to handle such comments. A lot of work is being conducted in this area. Mukul Anand[2] et.al in his paper has used deep learning based model and classified the comments in various categories namely, severe threat, toxic, obscene, insult and identity hate with the help of techniques like CNN, LSTM, etc. Ji Ho Park[3] et al. in their paper has implemented a two- step approach of classification of abusive comments and their classification and has compared it using the traditional single step approach using English Twitter corpus. The detection of comments is not confined to English only and has been done using other languages also like, Muhammad Pervez Akhter[4] et al. in his paper titled “Abusive language detection from social media comments using conventional machine learning and deep learning approaches” have used deep learning based models like CNN, LSTM, BLSTM and CLSTM for detecting abusive comments in Urdu and Roman Urdu languages. Similarly, Tanjim Taharat Aurpa[5] et al. in their paper has proposed an approach to classify the comments in Bengali as an abusive. M. Anand[6] et al. has used feature selection techniques and multilingual offensive language detection using deep learning techniques to classify the language as offensive or not after removing noise, stopwords, from them and performing filtering and segmentation. For feature selection they have Fuzzy based CNN. Monirul Islam Pavel in his work have implemented a CNN and NLP based ensemble method for that performs segmentation on non-toxic and toxic comments in the first phase and then classify them into 6 types, for the preprocessing part they have performed using NLP techniques like tokenization, vectorization, stemming, etc.

**Methodology:**

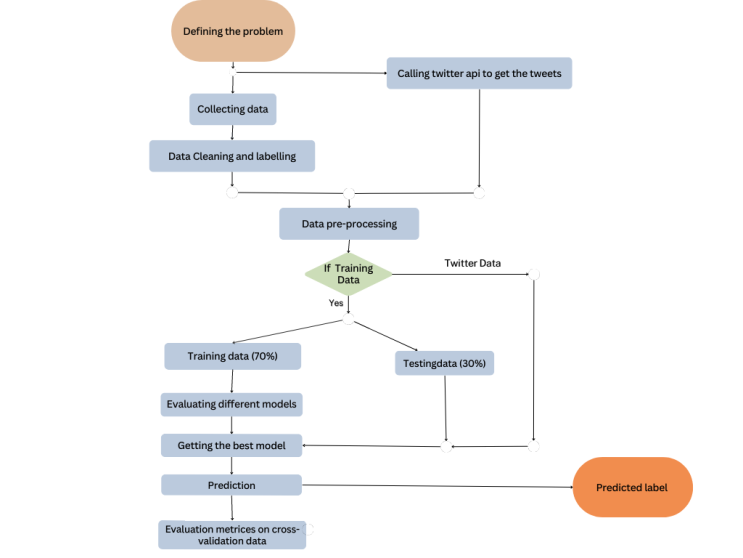


Fig. 1 Framework of the model.

In our work, we have various machine learning models for the analysis of the text data and classify them as abusive or non- abusive based on the features obtained after preprocessing the data. For training the data we have used models like Naïve Bayes, SVM, decision tree, logistic regression, random forest. Our framework takes a data as an input and performs the various preprocessing on the data. For the preprocessing part we have done the analysis of the data and removed the features which were not of use for our work. Then we have labelled our data based on the comment if it is abusive or not. Further we have applied NLP techniques for the preprocessing of the dataset. After that the data has been split into the training and testing dataset for cross validation and the training dataset have been trained on various models mentioned above. Then the testing dataset has been passed to the trained model and the model predicts the output for the testing data along with the evaluation metrics like accuracy, precision, recall and confusion matrix and classification report. After the cross-validation the overall model has been trained on voting classifier and the unknown testing data has been taken in real time from the Twitter using its API and after the preprocessing of the testing data the final predictions are made on the Twitter data. Our framework has been shown in Fig. 1

**Preprocessing:**

In the preprocessing part we have used NLP technique like lemmatization, lower casing of the sentences, tokenization, have removed unwanted space, removed tags, removed frequent words, removed non-english characters, removed https strings and removed stop words from the dataset.

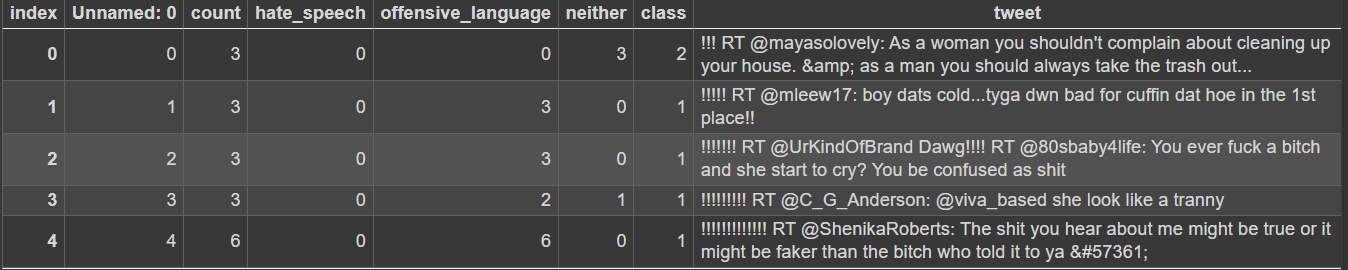


Fig. 2 Unprocessed initial data.

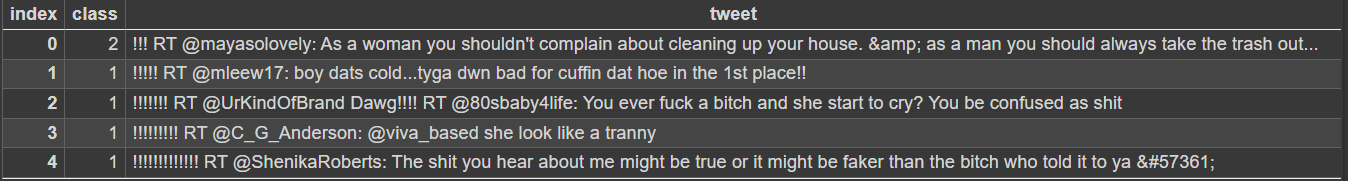


Fig. 3 Data with required features.

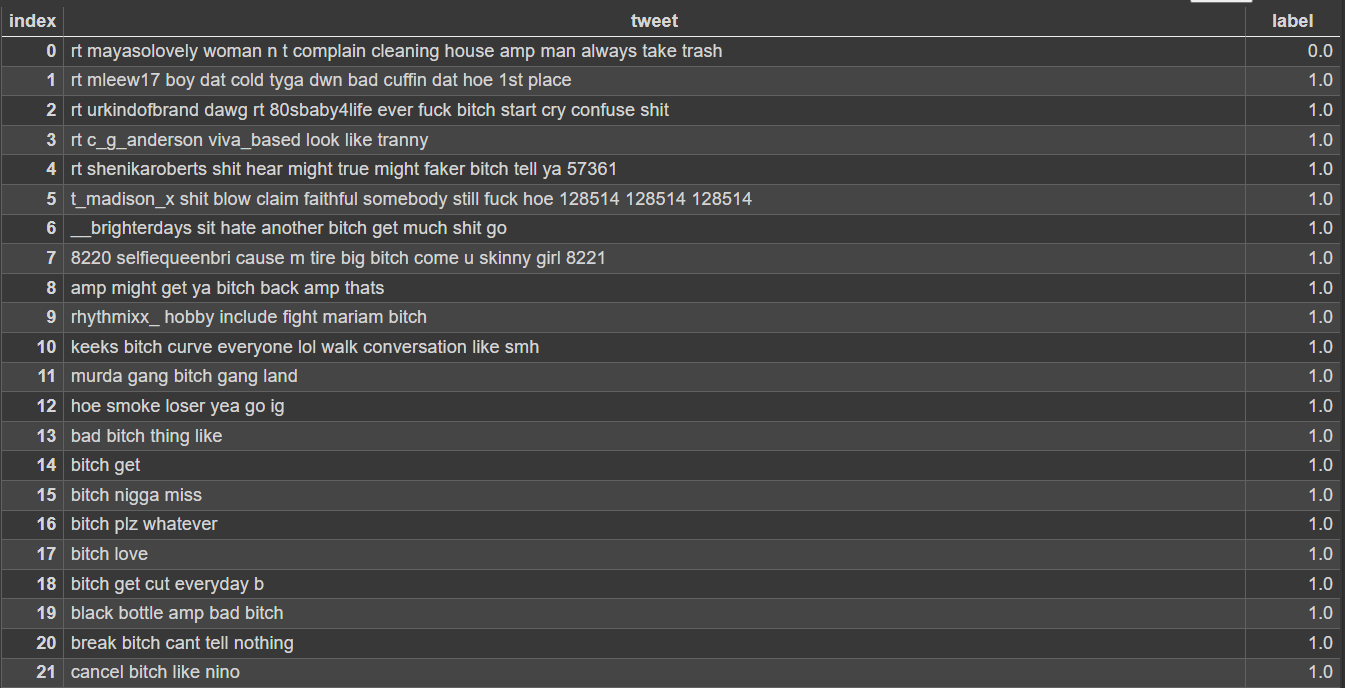


Fig. 4 Processed data.

**Naïve Bayes**

We have used Naïve Bayes in our model because it is one of the best known algorithms for text classification and it also doesn’t require much of the dataset to be trained and can be used very well for making real time predictions as it is fast since it is a probability based algorithm based on Bayes theorem. The equation of Bayes Theorem is given in equation 1:

(1)

**SVM**

The reason behind using the SVM is that it is very robust and they work very well with higher dimension space, there is also no need of selecting the features in the case of SVM thus, making the classification of text based data easier. For higher dimension space it uses kernel trick and separates the non-linearly separable data in the higher dimension.

**Decision Tree**

We have used decision tree because, it requires minimal need of data preprocessing and can be modeled in very less time and is very good for the classification tasks. The text data can have noise present in it.

**Random Forest**

To overcome this problem, we have used random forest as it is one of the most suitable algorithms to handle such kind of data as it has a capability to deal with noisy data in the higher dimension space. It takes output of multiple decision trees and based on the voting done on the output of each tree it predicts the output of the data, thus the chance of right predictions are high.

**Logistic Regression**

Since we have only two classes that are needed to be predicted and the data we have is linearly separable. It is a very effective and efficient algorithm when we need to perform binary classification and can give a very high accuracy score. It uses sigmoid function to bind the data in the range of 0 to 1.

The equation of the sigmoid function is given in equation 2, 3:

(2)

Where, z = (3)

**Result and discussion:**

The models have been trained on the training dataset of 70% of the total dataset and have been cross validated on the 30% of the remaining dataset and has been carried on the system having configuration as Intel® Core™ i5-5200U CPU @ 2.20GHz processor with 8GB RAM and has been trained on “Hate speech and offensive language dataset” of Kaggle[1] that consists of 24,782 samples and 7 features initially and has been filtered as per the requirement. The models have been trained on SVM, Decision Tree, Random Forest, Logistic Regression, Naïve Bayes and has been combined using Voting Classifier with the objects of SVM and Logistic Regression. The datasets are to be classified into two classes abusive and non-abusive. We have chosen Twitter data for the prediction because its API’s are easy to be available and are one of the foremost dataset used in this field.

We have compared our models based on the various evaluation metrics and the formulae are shown in equations 4, 5, 6, 7.

(4)

(5)

(6)

(7)

Table 1 shows the comparative analysis of the different classifiers used based on the accuracy, precision, recall, F1-score obtained on various models.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classifier** | **Accuracy (%)** | **Precision (%)** | **Recall (%)** | **F1-Score (%)** |
| **Logistic Regression** | 96.04 | 98.11 | 97.03 | 97.58 |
| **SVM** | 96.21 | 98.52 | 96.82 | 97.66 |
| **Decision Tree** | 82.14 | 82.18 | 99.79 | 90.13S |
| **Random Forest** | 81.74 | 81.74 | 1.00 | 89.95 |
| **Naïve Bayes** | 90.75 | 90.80 | 98.67 | 94.57 |
| **Voting Classifier** | 96.16 | 97.82 | 97.46 | 97.64 |

**Table 1. Comparative analysis of classifiers based on accuracy, precision, recall, F1-score.**

Table 2 shows the comparative analysis of the different classifiers used based on the TP, TN, FP, FN and accuracy obtained on various models and Fig 5 shows the sample comments obtained from the twitter before preprocessing it.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Classifier** | **TP** | **TN** | **FP** | **FN** | **Accuracy (%)** |
| **Logistic Regression** | 1172 | 5557 | 107 | 170 | 96.04 |
| **SVM** | 1196 | 5545 | 83 | 182 | 96.21 |
| **Decision Tree** | 40 | 5715 | 1239 | 12 | 82.14 |
| **Random Forest** | 0 | 5727 | 1279 | 0 | 81.74 |
| **Naïve Bayes** | 707 | 5651 | 572 | 76 | 90.75 |
| **Voting Classifier** | 1155 | 5582 | 124 | 145 | 96.16 |

**Table 2. Comparative analysis of different classifiers based on TP, TN, FP, FN and accuracy.**

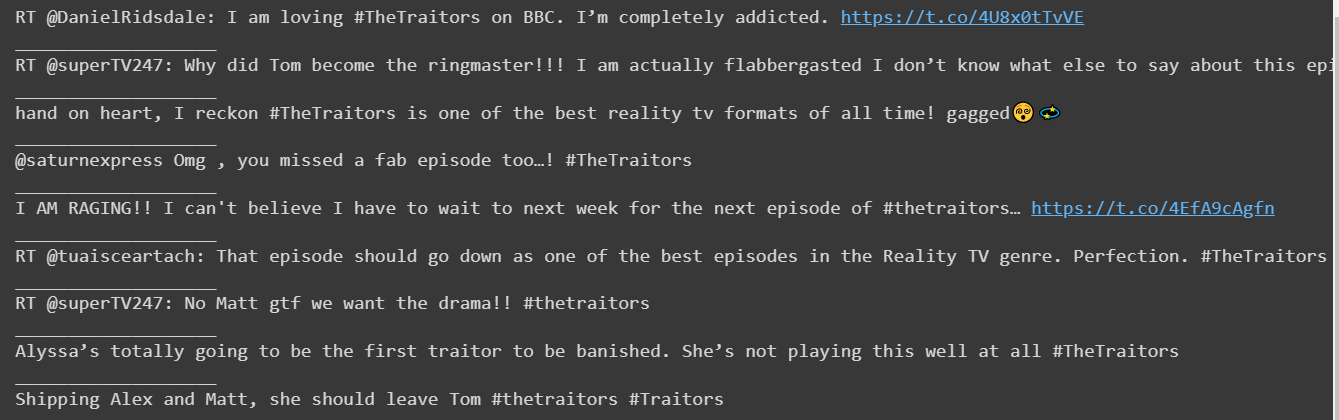


Fig. 5 Sample comments.

**Conclusion:**

In this project, we have proposed a machine learning based approach for detecting if the comment is of abusive nature or not. Our approach is capable enough of identifying the comments based on its nature into the suitable class with a very high F1-score. We have procured An F1-score of 97.58% using Logistic Regression, 97.66% using SVM, using Decision Tree, 89.95% using Random Forest, 94.57% using Naïve Bayes and 97.64% using Voting Classifier. We can see that the accuracies of Decision Tree and Random Forest are although high (>80%) but still both are not a good classifiers for this dataset as they both are having the problem of accuracy paradox.

**Individual Contribution:**

**1. Chitransh Bose:**

**2. Varun Jhunjhunwala:**

**3. Ranit Pal:**

**4. Soumyajyoti Das:**

**References:**

1. Dataset: https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive-language-dataset?resource=download

2. Mukul Anand; R. Eswari. Classification of Abusive Comments in Social Media using Deep Learning. Published in: 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 10.1109/ICCMC.2019.8819734

3. Ji Ho Park, Pascale Fung. One-step and Two-step Classification for Abusive Language Detection on Twitter. Published in: 1st Workshop on Abusive Languageof the Association of Computational Linguistics (ACL) 2017 (Vancouver, Canada), August 4th, 2017. https://doi.org/10.48550/arXiv.1706.01206

4. Akhter, M.P., Jiangbin, Z., Naqvi, I.R. et al. Abusive language detection from social media comments using conventional machine learning and deep learning approaches. Multimedia Systems 28, 1925–1940 (2022). https://doi.org/10.1007/s00530-021-00784-8

5. Aurpa, T.T., Sadik, R. & Ahmed, M.S. Abusive Bangla comments detection on Facebook using transformer-based deep learning models. Soc. Netw. Anal. Min. 12, 24 (2022). https://doi.org/10.1007/s13278-021-00852-x

6. M. Anand, Kishan Bhushan Sahay, Mohammed Altaf Ahmed, Daniyar Sultan, Radha Raman Chandan, Bharat Singh. Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. ISSN 0304-3975. https://doi.org/10.1016/j.tcs.2022.06.020.