Toxic Comment Classification

Naikawadi Shivaprasad, Neerukonda Venkata Srinivas, Sabavath Purya
B20AI057, B20CS037, B20EE054
Github Repo: *PRML: Course Project*

Index Terms

Natural Language Processing, Text classification,Logistic Regression, Support Vector Machine, Extreme Gradient Boosting(XGBoost), Pipeline

I. Introduction

THE penetration of the internet in all domains of life has led to an increase in people's participation actively and give remarks as an issue of communicating their concerns/feedback/opinions in various online forums. Although most of the time these comments are helpful, sometimes these may be abusive and create hatred-feeling among the people. as these are openly available to the public and are being viewed from various sections of the society, people in different age groups, different communities, and different socio-economic backgrounds, it becomes our prime responsibility to filter out these comments in order to stop the spread of negativity or hatred within people. detecting Toxic comments has been a great challenge as the type of diverse data collected, Pre-processing Methods, models we use to predict Toxicity, and accuracy scores are all important issues in this study. The deployment of the model has been explained, and several feature vector combinations and machine learning models have been compared.

II. Data Description and Preprocessing

The dataset contains text comments collected from social media posts and the target toxicity and additional toxicity subtype attributes which are rated as continuous values between 0 and 1.

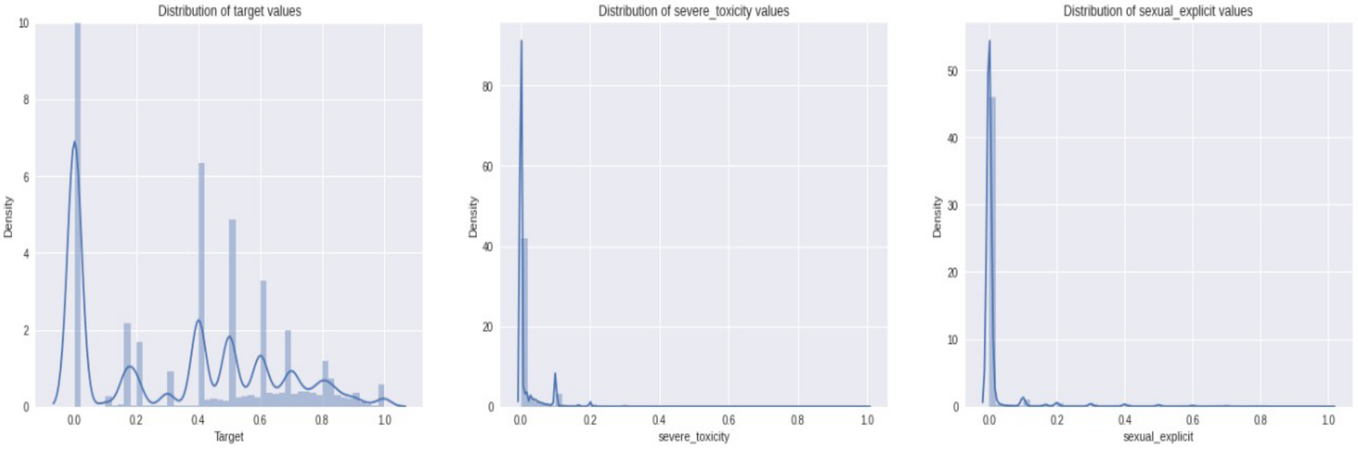

Fig. 1. a. Distribution of Target Toxcity, b. Distribution of severe toxicity,
c. Distribution of Sexual Explicit

Data Preprocessing

A. Performing NLP

- Converted the whole comments into lower case.
- Replaced punctuations and other unnecessary elements with spaces.
- Dropped links (like https, html, etc) and emails.
- stop words were removed using the NLTK library and lemmatized the text using the same NLTK library.
- Finally, cleaned posts were stored in a separate column to compare the original and pre-processed texts.

B. Labels transformation

The given values of the labels were continuous as the ”target toxicity” and other toxicity subtype attributes were rated in range of 0 and 1 as continuous values. So, the labels are converted into binary classes by considering the threshold value of 0.5, as most of the toxic comments have target toxicity value greater than 0.5. So, the labels with continuous value greater than equal to 0.5 are replaced with 1 and less that 0.5 are replaced with 0.

C. Data splitting

Splitted the cleaned data into train and validation data in the ratio 75:25 so that the model can be trained on train data and validated using validation data.

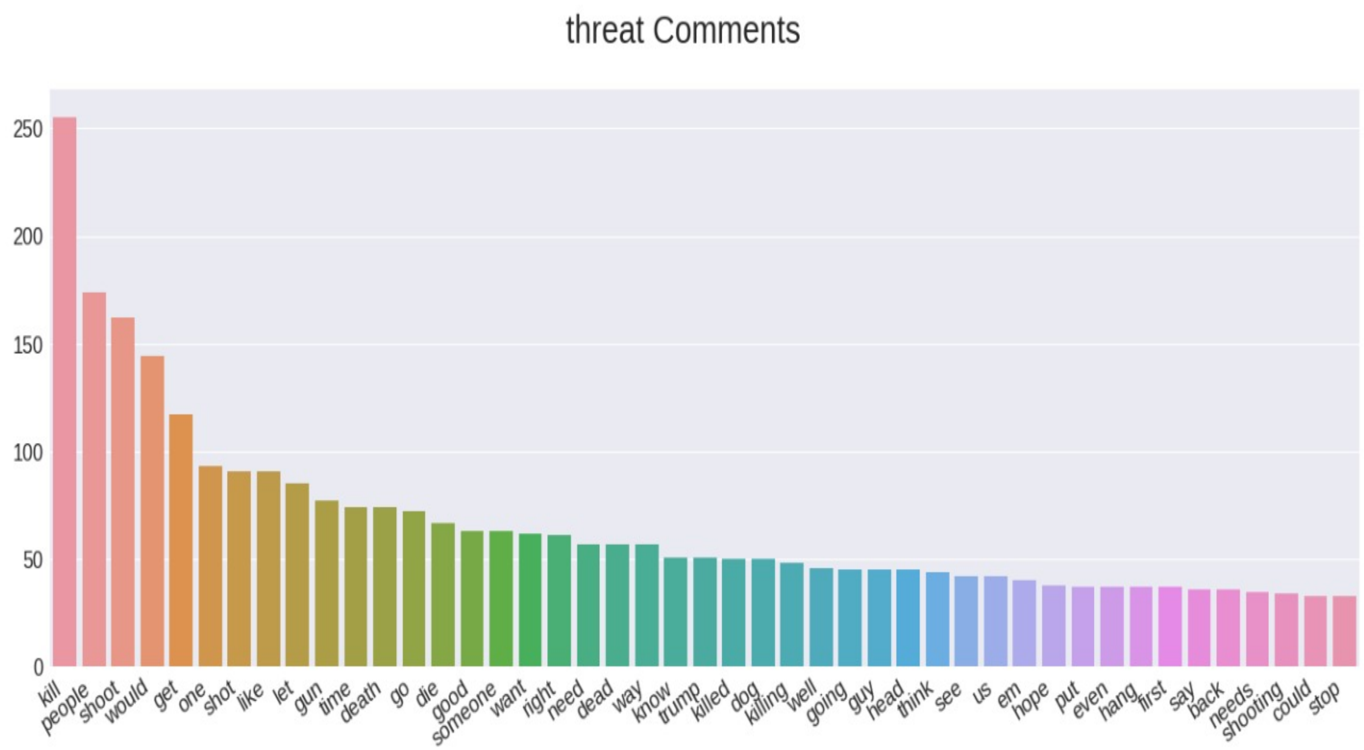

Fig. 2. Visualization of data

III. Machine Learning Models

A. Multinomial Naive Bayes

In Natural Language Processing(NLP), the Multinomial Naive Bayes method is a common Bayesian learning approach. Using the Bayes theorem, the programme estimates the tag of a text, such as an email or a newspaper piece. It assesses the likelihood of each tag for a given sample and returns the tag with the highest probability. The Naive Bayes classifier is made up of several algorithms, all of which have one thing in common, each feature being classified is unrelated to any other feature. The presence or absence of one trait has no bearing on the inclusion or exclusion of another.

B. Linear SVC

The Linear Support Vector Classifier (SVC) approach uses a linear kernel function to classify data and works well with huge datasets. When compared to the SVC model, the Linear SVC adds additional parameters such as penalty 1(i.e, C=1) and normalization (L1 or L2), and loss function.It is designed to fit to the data you provide and provide a ”best fit” hyperplane that divides or categorises your data. Following that, you may input some features to your classifier to check what the ”predicted” class is after you've obtained the hyperplane. we got an accuracy of 94.82 by using this model

C. XGBoost Classifier

It is an ensemble learning technique which builds a strong classifier by combining different weak classifiers. At first a model is built using training data and the second model is built in a way that it tries to correct the misclassification done in first and this keeps on repeating till the training data is correctly classified or till maximum number of models is reached.XGBoost classifier was used to train the model and the model was evaluated. we got an accuracy of 93.72 using this model

D. Logistic Regression

Logistic Regression is a classification system based on Machine Learning. It's a method for predicting a categorical dependent variable from a set of independent variables. Solver used was “sag” as it performs well when there is a large dataset and the cost function used was “sigmoid function” which converts the predicted values into the probabilities between the range 0 and 1 , we got accuracy of 94.41 using this model

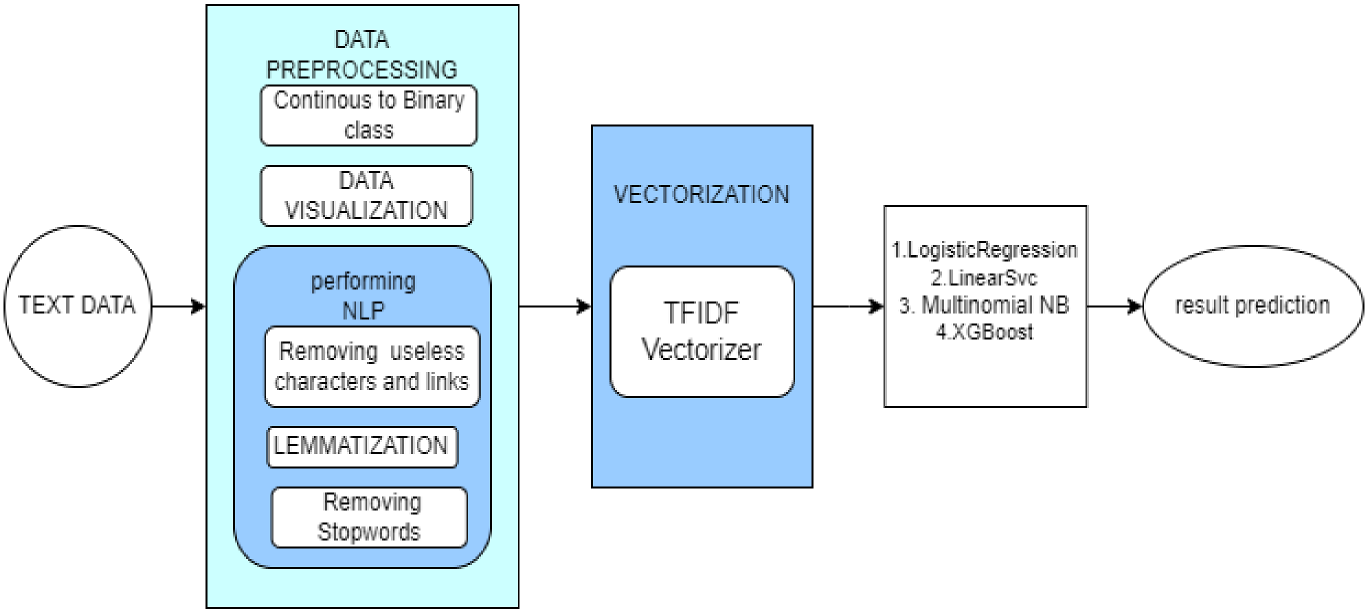

Fig. 3. whole architecture of the project

IV. Model Building

A. Pipeline

A pipeline was created which takes the comment texts and performs pre-processing using a class which uses sklearn base library and performs NLP on the text comments and sends the output of the pre-processed text to the vectorizer which was created using TFIDF vectorizer which vectorizes the comments text and sends the vectorized text comments to the Machine learning model.

B. Model training

The selected machine learning models (i.e., Logistic Regression, Linear SVC, Multinomial Naive Bayes, XGBoost) were trained using training data by creating separate pipeline for each model and reported the performance for each model.
Evaluation metrics used: Accuracy score, F1 score, Precision, Recall

During the process of training the models along with the ”target toxicity” other additional toxicity subtype attributes like obscene, identity attack, insult, threat and others were also predicted to classify the toxic comment more precisely based on subtype attributes. The accuracies reported finally were the values of ”target toxicity” and different models were compared based on the same label.

V. Classifier Comparision

We have compared the four different Machine Learning models (Random forest, Linear SVC, MULTINOMIAL Naive Bayes, XG Boost). For these models, along with accuracy, measures like precision, recall, F1-score were compared. The comparision is shown below:

Classifier	Accuracy%	Precision	Recall	F1-Score
Logistic Regressor	94.41	0.9389	0.9441	0.9339
XG Boost	93.72	0.9337	0.9372	0.9203
Linear SVM	94.82	0.9428	0.9482	0.9428
Multinomial Naives Bayes	92.49	0.9233	0.9249	0.892

ROC CURVES and COMPARISION of ACCURACIES

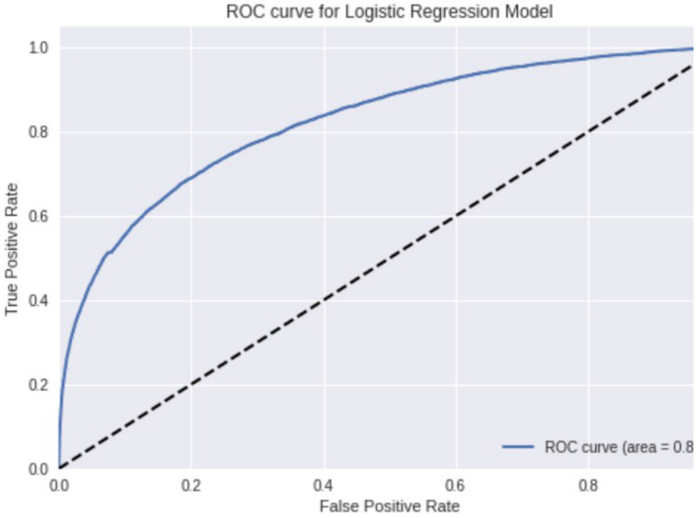

Fig. 4. ROC Curve for Linear Regression Model on cross-validation

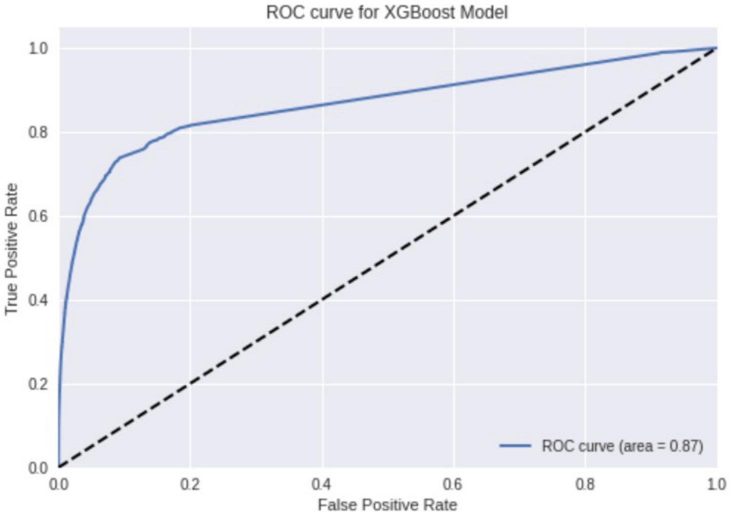

Fig. 5. ROC Curve for XG Boost model on cross-validation

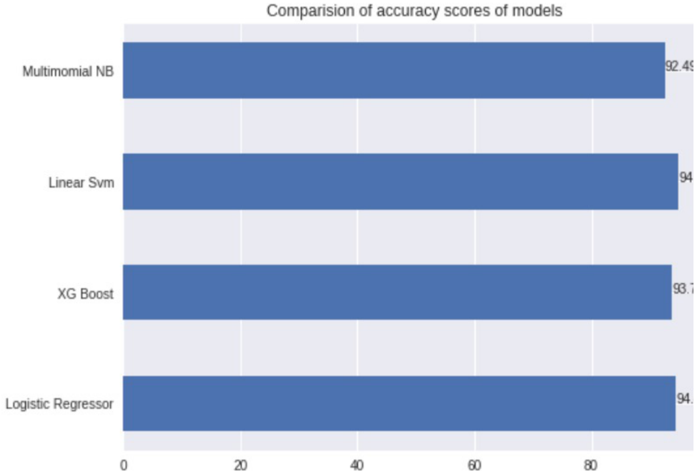

Fig. 6. The comparison of accuracies of various ML

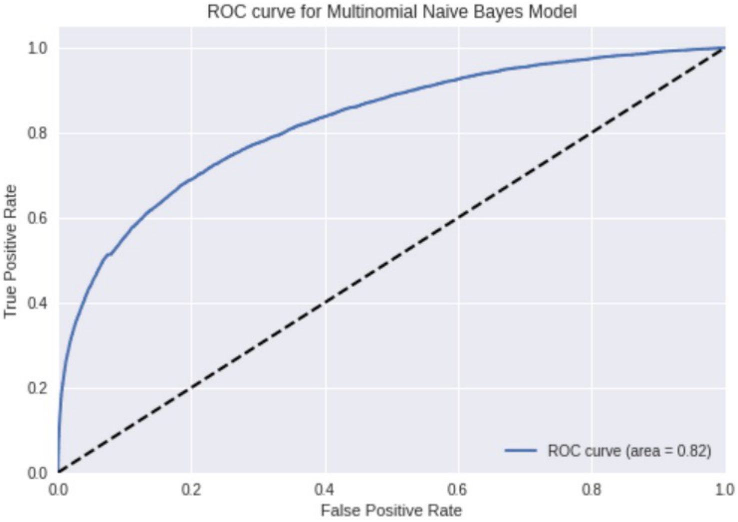

Fig. 7. ROC Curve for Multinomial Naive bayes Model on cross-validation

VI. Conclusion

The comparision table of the models shows that all the classifiers had nearly equally efficient performance and among all the classifiers used, Linear SVC model was giving result in less time and performance of it was also little better compared to other models. As the dataset contains more data samples with less toxicity compared to the data samples with high toxicity due to which measures like F1 score,precision and recall values were less accurate. We can conclude that Linear SVC model is preferred.

Contribution

The learning and planning was done as a team.The individual contributions are as given,

- Naikawadi Shiva Prasad (B20AI057) - Data pre-processing and exploratory data analysis, performing NLP on text data, Pipeline implementation, Report.
- Neerukonda Venkata Srinivas (B20CS037) - Logistic Regression, XG Boost models implementation, Data Visualization based on NLP, models comparision and analysis, Report.
- Sabavath Purya (B20EE054) - Linear SVC, Multinomial NB models implementation, data pre-processing,Pipeline, report.

References

https://github.com/jayspeidell/ToxicCommentClassification-/blob/master/ToxicComments$_{E}DA.ipynb$
https://www.analyticsvidhya.com/blog/2021/12/different-methods-for-calculating-sentiment-score-of-text/
http://scikit-learn.org/stable/modules/generated/sklearn.naive
https://saejournal.com/wp-content/uploads/2021/07/Personality-Prediction-Using-Machine-Learning.pdf