MALIGNANT COMMENTS CLASSIFIER PROJECT

A project report submitted in

partial fulfilment of the requirement for the award of internship provided by

FlipRobo Technologies, Bangalore

Submitted by:

ELTON GRIVITH D SOUZA

engardaelton.work@gmail.com

DECEMBER 2021

ACKNOWLEDGMENT

I would like to offer my sincere thanks to **FlipRobo Technologies** for providing the opportunity to intern under their company.

I would like to thank **Ms Khusboo Garg** for coordinating and providing valuable suggestions during the process of this project and internship.

Finally, I would like to express my sincere gratitude to my family and friends who have been the embodiment of love and support which helped us carry out this project in a smooth and successful way.

INTRODUCTION

Problem Statement

The proliferation of social media enables people to express their opinions widely online. However, at the same time, this has resulted in the emergence of conflict and hate, making online environments uninviting for users. Although researchers have found that hate is a problem across multiple platforms, there is a lack of models for online hate detection. Online hate, described as abusive language, aggression, cyberbullying, hatefulness and many others has been identified as a major threat on online social media platforms. Social media platforms are the most prominent grounds for such toxic behaviour.

There has been a remarkable increase in the cases of cyberbullying and trolls on various social media platforms. Many celebrities and influences are facing backlashes from people and have to come across hateful and offensive comments. This can take a toll on anyone and affect them mentally leading to depression, mental illness, self-hatred and suicidal thoughts. Internet comments are bastions of hatred and vitriol. While online anonymity has provided a new outlet for aggression and hate speech, machine learning can be used to fight it. The problem we sought to solve was the tagging of internet comments that are aggressive towards other users. This means that insults to third parties such as celebrities will be tagged as inoffensive, but “u are an idiot” is clearly offensive.

Our goal is to build a prototype of online hate and abuse comment classifier which can used to classify hate and offensive comments so that it can be controlled and restricted from spreading hatred and cyberbullying.

Data Sources and their formats

The data was provided by the company which contained a training dataset with approximately 159,000 samples and the test dataset with 153,000 samples. The data samples contain 8 features that identify the post, comments and the labels categorised into its level of offence .

Data Pre-processing

All features were first correctly formatted to provide proper input to the model.

The comments were converted to lower case to reduce intra-diversity.

WordNetLemmatiser was used to group different inflected forms of a word.

Wordcloud was used to sense which words were offensive.

VIF was used to check for multi-collinearity. The threshold was taken to be 5. The results acquired, showed that the data showed no signs of the said issue.

Outliers were removed using Z-Score Analysis. It reduced out data by a bit more than 3124 tuples. So data loss did not become an issue in this step.

Data Relationships

The relationships analysed from the data was done through exploring the correlation, distribution of feature values, distribution of feature sets, distribution of feature values in regard to the target feature etc.

The observations are documented extensively in the analysis notebook provided in the attachments.

Hardware and Software Requirements and Tools Used

The development of this project required the following tools:

- Python 3.8
- CUDA 10
- Scikit-learn 1.0

Model Development and Evaluation

Model Development

The model was trained on a 30% split of the training data which was fed to a pipeline of different classification algorithms. These models are listed below:

- Logistic Regression
- MultinomialNB
- Decision Tree Classifier
- K Neighbours Classifier
- XGB Classifier
- ADABoost Classifier
- Gradient Boosting Regressor

The model was evaluated based on the accuracy, precision, recall and CV using AUC/ROC plots.

Hyper Parameter Tuning *(HPT)*

Parameter tuning was done using GridSearchCV on the Logistic Regression. This model was chosen based on its CV scores. The grid parameters that were used are given below:

- ‘C’ : [0.001, 0.01, 0.1, 1, 10, 100]
- penalty : [‘l1’,‘l2’]
- max_iter : list(range(1,3002,500))
- solver : [‘newton-cg’,’lbfgs’,‘liblinear’]

The best parameters obtained from this process was then used to provide the submission model.

CONCLUSION

The submitted model provides consistent results that can be validated under other circumstances. The result obtained from the provided data can be assumed to be competitive and can provide better insights to the client. The provided solution satisfies all the requirements needed by the client.

Future Work could include cleaner, balanced and localised data with more dependent and independent feature sets to provide better estimations and analytical relationships.