CLASSIFYING TOXIC USER COMMENTS USING BOOSTING ALGORITHMS

TASNEEM SULEMAN, SUPERVISOR - DR. LEI SHI
INSTITUTE OF TECHNOLOGY, CARLOW
MSc. In Data Science, Department of Computing and Networking

INTRODUCTION

In the recent times, one of the most important tasks for social platforms has been to moderate the user comments to ensure healthy and constructive discussions and maintain a balance between freedom of speech and the quality of comments[1]. This task of sentiment classification is one of the major applications in the Natural Language Processing domain.

Fig. 1 - Example of toxic and non-toxic comment[1]

A large proportion of the content posted online contains several typographical errors. This may be a deliberate attempt to come up with abusive comments that seem creative[2].

ABOUT THE DATASET

The dataset consists of 1,59,570 comments from Wikipedia Talk page published by Google Jigsaw in December 2017, that have been labelled as toxic by human moderators. These comments are to be classified into one or more levels namely, toxic, severely toxic, obscene, hate, insult and threat.

RESEARCH OBJECTIVE

The study would be an attempt to compare the performance of Boosting techniques on the comments which contains words that are out of vocabulary, multi-word phrases and long-range persistence. The objective of this research would be to determine if removal of such discrepancies i.e. pre-processing of data helps in producing better results.

LITERATURE REVIEW

Traditional Machine learning techniques have been applied widely to classify sentiments but boosting algorithms – XGBoost and AdaBoost have not been explored much in the classification sphere.

A recent work at Intel was done using various classification algorithms like Logistic Regression, Naïve Bayes with SVM, XGBoost and FastText algorithm with Bidirectional LSTM. The study also focusses on studying the usefulness of 35 transformations applied on the data[2].

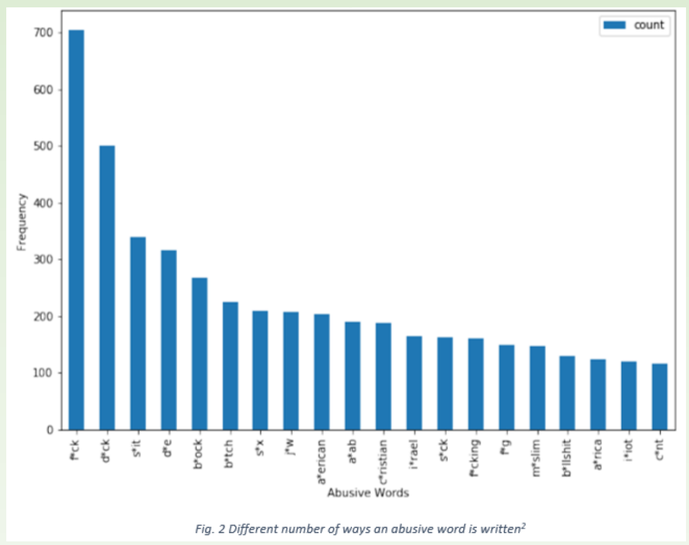
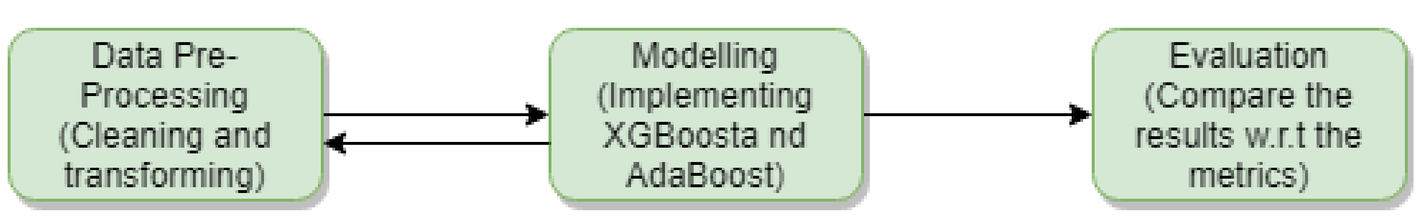


Fig. 2 Different number of ways an abusive word is written[2]

Another work at Penn State presents an AdaBoost model which was used to identify features useful in predicting sentiments of cancer survivors[3].

METHODOLOGY



NEXT STEPS

Moving forward, the immediate step would be to prepare the data for modelling which includes removal of redundant characters, eliminating stopwords, stemming, lemmatization and finalizing the evaluation metrics. Later, the chosen models would be implemented and evaluated against the decided metrics.

CONTACT:
Tasneem Suleman
Phone: +353-899881019
Email: tasneemnew95@gmail.com
[https://www.linkedin.com/in/tasneem-suleman/]

REFERENCES

1.Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G. & Plagianakos, V. P. Convolutional Neural Networks for Toxic Comment Classification. in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence - SETN '18* 1–6 (ACM Press, 2018). doi:10.1145/3200947.3208069

2.Mohammad, F. Is preprocessing of text really worth your time for online comment classification? *arXiv:1806.02908 [cs]* (2018).

3.Qiu, B. *et al.* Get Online Support, Feel Better – Sentiment Analysis and Dynamics in an Online Cancer Survivor Community. in *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing* 274–281 (2011). doi:10.1109/PASSAT/SocialCom.2011.127