A comparative study of toxic language detection using NLP algorithm

Guide Name	Panel Head
Dr. G Ramya	Dr. T. Senthil Kumar

Faculty Advisor	Project Domain
Dr. G Usha	Natural Language Processing

Student(s) Details: Name

1. Shivam Pandey
2. Mayank Sinha

Passport size photo(s)

Registration Number(s)

1.RA1911003010383

2.RA1911003010386

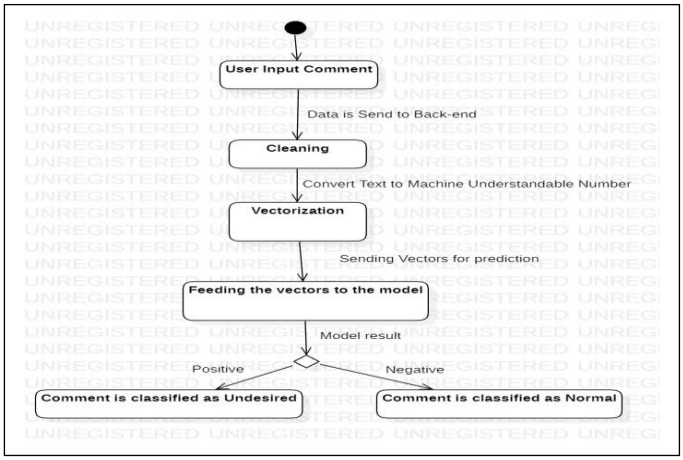
Email ID(s)&Mobile Number(s)

1: ms7651@srmist.edu.in	2: sp7287@srmist.edu.in
9934996311	9026524877

Abstract

As there are numerous algorithms utilised for toxic language detection in the field of natural language processing now, but they in this algorithm not developed considering production environment. The speed of categorization has a significant impact on how quickly an application is used in a production setting where the algorithm is running in real-time. The technique is only useful for theoretical applications if it performs classification tasks slowly. In the actual world, in this algorithm need algorithms that operates quickly and with excellent accuracy. Through this research, in this algorithm hope to identify an algorithm that works best in the actual world, meaning that there is no compromise in this algorithm speed and accuracy. For the classification job, in this algorithm will make use of the toxic language dataset. After classifiers have been trained, they will be put to the test using a variety of criteria, including accuracy, recall, precision, F1 score, and prediction time. In this algorithm shall categorise the flaws in different NLP algorithms based on the parameters

Architecture Diagram



Significance of the Project

Through this project we aim to devise an algorithm which is viable in real-world i.e., the algorithm is fast and accurate at the same time and there is no tradeoff between two. We will be conducting a study between various NLP algorithms, and we will aim to identify the shortcomings in the current algorithms. We will be using the toxic language dataset for the classification task. We will be training the classifiers and then we will be testing the trained classifiers based on variety of parameters like accuracy, recall, precision, F1 score, time taken for prediction. Based on the parameters we will define what are shortcomings in various NLP algorithms.

Conclusion

Through this project comparison was drawn between various NLP algorithms in terms of their training time,accuracy,precision,recall and prediction time in the runtime enviroment. In this project SVM was recognised as the best suited algorithm for the supervised problems. It had high accuracy and low prediction time in the runtime environement. In this project it was found that Logisitic Regression and K-nearest neighbors was found to be fast in execution but had low accuracy on the test set. On the other hand Random Forest and Decision tree took a lot of time in HyperParameter Tuning and the test results were also very low.

Conference/Journal Publication Details (If Any)

Submitted the paper in 3rd International conference on Computational Intelligence-ICCI 2022 (Dec 29-30,2022) at IIIT, Pune. Accepted paper will be published in the Springer book series “Algorithms for Intelligent Systems” (Scopus/WOS)