Identifying Toxicity in User Comments on Wikepedia Dataset

Himanshu Punetha, Suman Kumar and Vaibhav Garg

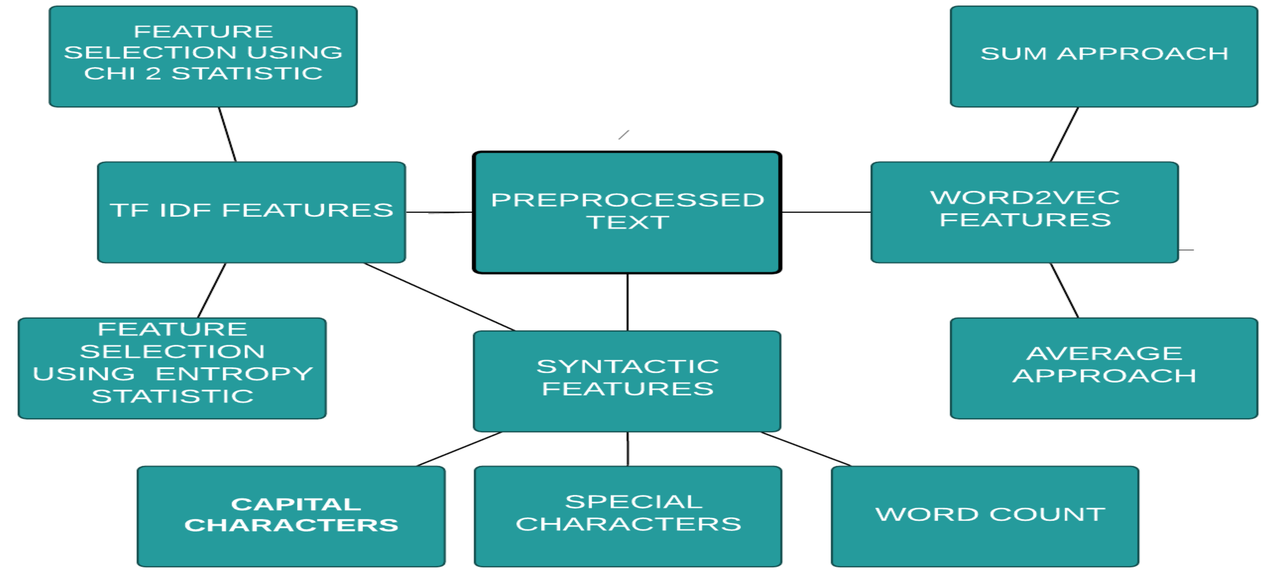
Aim

The aim of this project is to classify a given comment into different classes of toxicity (Toxic, Severe Toxic, Obscene, Threat , Insult and Identity Hate). This is a multilabel classification problem in which one comment can be classified into multiple classes or may belong to none of the classes.

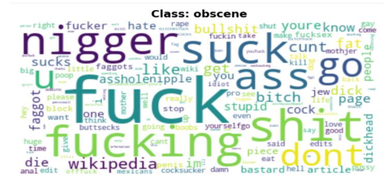
Motivation

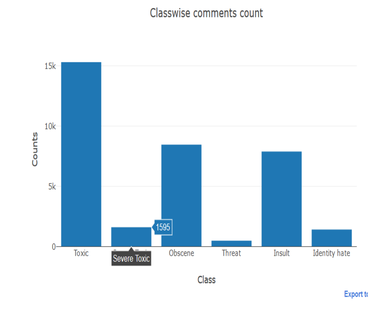
While interacting on online platforms, people tend to use toxic words in order to show their aggression. These toxic words can be categorized into obscene, threatening insulting words etc. Nowadays, online platforms have abundance of such toxic comments which can lead to cyberbullying or online harassment of users.. Thus there is a need of an hour to identify such toxicity in comments and prevent them from being posted on online platform.

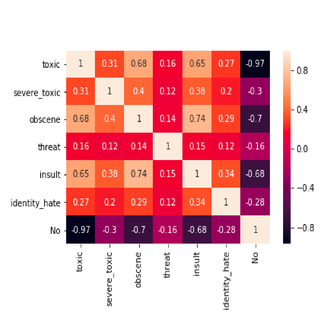
Methodology

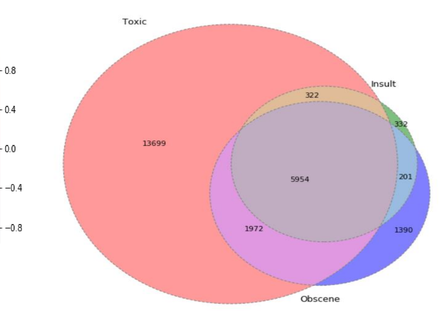


Data Visualization

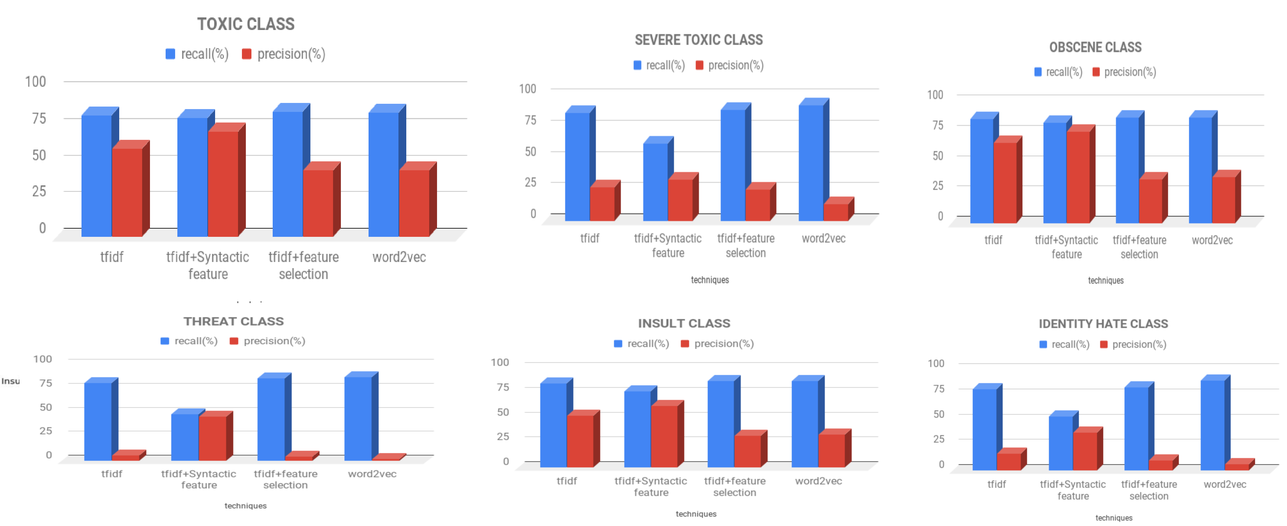








Results



References

Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak,Bernard J. Jansen.Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media.Proceedings of the Twelfth International AAAI Conference on Web and Social Media.