Toxic Comment Classification

Soham Kulkarni (17211395)

Abhishek Pandey (17210432)

Adesh Shendge (17211687)

Ashish Poigal (17211775)

Problem statement:

Social media platforms provide an environment where people can freely engage in discussions. Unfortunately, this also enables several problems, such as online harassment. Recently, Google and Jigsaw started a project called Perspective, which uses machine learning to automatically detect toxic language. A demonstration website has been also launched, which allows anyone to type a phrase in the interface and instantaneously see the toxicity score. Platforms struggle to effectively facilitate conversations, leading many communities to limit or completely shut down user comments. One area of focus is the study of negative online behaviors, like toxic comments (i.e. comments that are rude, disrespectful or otherwise likely to make someone leave a discussion). But the current models still make errors, and they don’t allow users to select which types of toxicity they’re interested in finding (e.g. some platforms may be fine with profanity, but not with other types of toxic content).

In this project, we will be building a multi-headed model that’s capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate better than Perspective’s current models. We will be using a dataset of comments from Wikipedia’s talk page edits. Improvements to the current model will hopefully help online discussion become more productive and respectful.

Dataset :

Dataset consists of large number of Wikipedia comments which have been labeled by human raters for toxicity. We will create a model which predicts a probability of each type of toxicity for each comment. The types of toxicity are:

* Toxic
* Severe toxic
* Obscene
* Threat
* Insult
* Identity hate

Existing dataset will be dived into 70 % and 30% for training and testing respectively.

Implementation :

* Dataset contains toxic comments in form of strings. In order to derive meaning out of it we need to process comments from training data with Natural language processing methods.
* Natural language processing will get rid of Stop words, punctuation marks, redundant symbols and whitespaces.
* Cleaned data will be fit to be transformed term frequency(TF) and Inverse document frequency(IDF). Alternatively, the comments will be converted to Word2vec using genism language modeling package/library.
* Elements from TF IDF matrix can be treated as features and further modeled against toxicity level. Machine learning algorithms like Naïve Bayes, SVM, ANN and CNN will be used and their accuracy will be compared.

Validation:

Accuracy will be measured by comparing predicted toxicity and available toxicity as testing dataset will have all the ratings.

Tools :

Jupyter notebook will be used for Natural language processing and training and testing the data.

Existing Work :

Google and jigsaw has come up with Perspective API which does the toxic comment classification using machine learning methods.