Toxic Comment Classification

Problem

The 21st century is the age of social media. On one hand, many small businesses are benefiting and on the other, there is also a dark side to it. Thanks to social media, people are aware of ideas they are not used to. While few people take it positively and make an effort to get used to it, many people start going in the wrong direction and start spouting malicious words. So many social media apps take the necessary steps to remove these comments in order to predict their users, and they do so using NLP techniques.

Introduction

This effort aims to make the internet a safer place. Toxic comments on social media can be quickly reported and deleted by looking for them. In the long term, this would improve interpersonal connections in this increasingly digital environment. Because it is difficult for hard-coded algorithms to interpret and recognize human speech and expensive for businesses to engage humans to identify these comments, a neural network is the most effective way to solve this problem.

Toxic-Comments Dataset

Data is used from the “Toxic Comment Classification Challenge” on Kaggle, This specific dataset consists of a large number of Wikipedia comments which have been labeled by human raters about each comment's toxicity. Train-set and test-set include 159571 and 153164 comments respectively. The purpose is to build a model that classifies the comments of the test set into six different categories regarding the types of toxicity. More precisely, the basic goal of this assignment is to train a Natural Language Processing model which predicts the probabilities of each type of toxicity for each comment. The following table provides information about the label frequencies in the train set. It can be easily observed that the most multitudinous class is "Toxic" which is present in 15294 comments. This is reasonable because toxic is a very general label compared to the others.

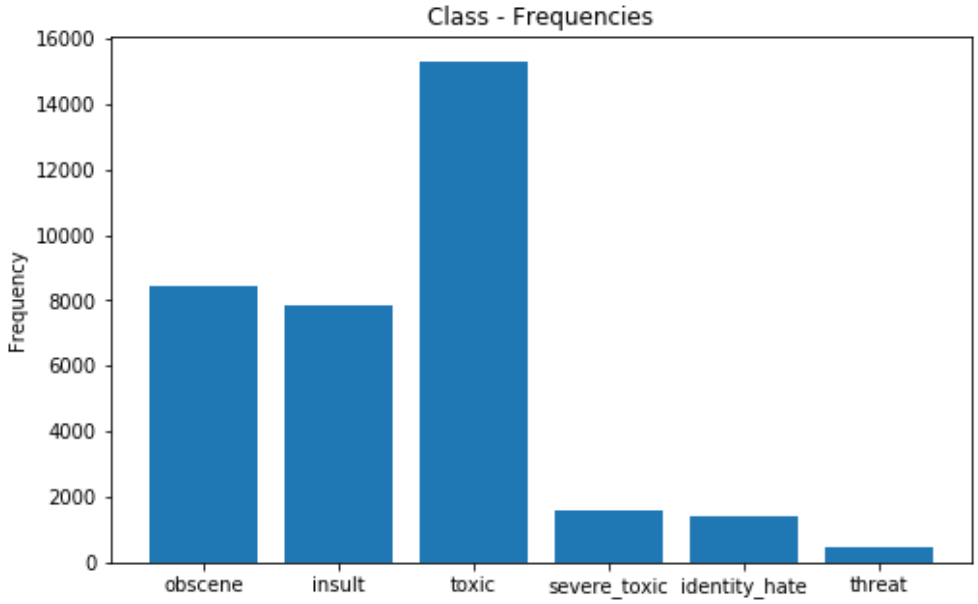
Table: Train set Class-Frequencies

Obscene	Insult	Toxic	Severe toxic	Identity hate	Threat
8449	7877	15294	1595	1405	478

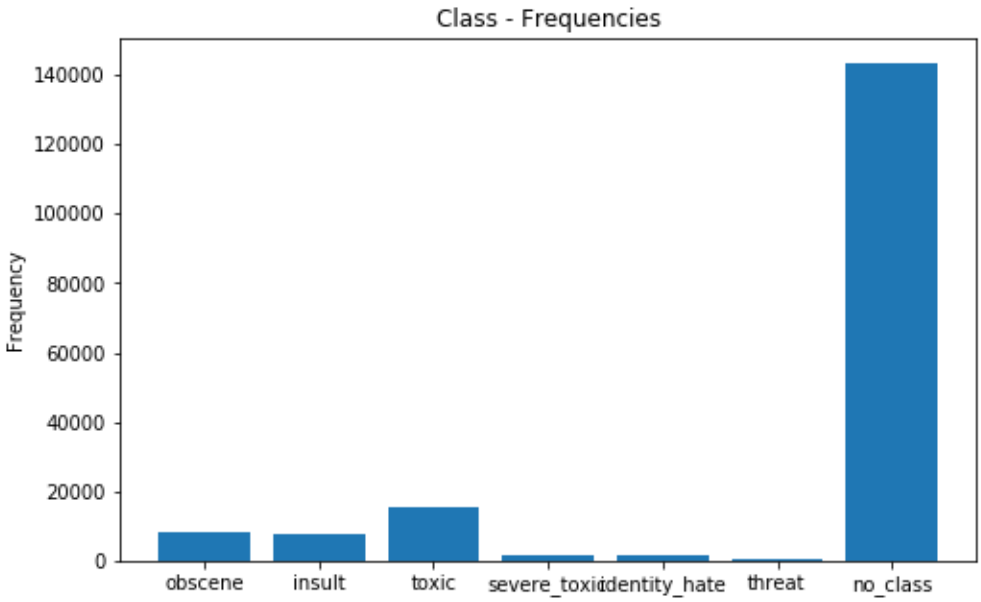
Imbalanced Data and Bias

The characteristic of the data set is a large number of unmarked comments. About %90 of the entire dataset consists of comments that do not belong to any class. This fact can have a strong impact on classification models. There are many solutions to the problem of imbalanced data, such as undersampling/oversampling and general methods of artificially balancing data. Unfortunately in this case this is very difficult due to the linguistic nature of the data. Also, comments cannot be discarded without captions because they would lose important information.

The chart below visualizes the frequency of each class in the training set.



The chart below visualizes the frequencies of imbalanced data



Method

This dataset can be used to classify the comments as toxic or non-toxic. For this project, textual data preprocessing techniques need to be used first. After that, basic NLP methods like TF-IDF for converting textual data into numbers could be performed, and then machine learning algorithms can be used to label the comments.