Toxic Comment Classifier

Aditya Sridhar, Chandan Shankarappa
College of Computer and Information Science, Northeastern University
sridhar.ad@husky.neu.edu, shankarappa.c@husky.neu.edu

Abstract

The objective of the project is to identify if a given comment is toxic in nature or not. This is a topic that has been dominating the news of late, given the extent to which social media and message boards are getting polarized. It is an inherently subjective task and it is hard to decipher without examining the context of the comments. As part of this project, we plan to evaluate commonly used sentiment analysis tools and some of the more recent deep learning techniques, to model this problem accurately.

Introduction

With the increasing dependency on and use of social media comes an increase in being exposed to rude and mean comments. We can all agree that not all use the discussion boards or comment section to provide quality conversations or discussions. Some men just want to watch the world burn.

These comments not only bring down the discussion at hand but might also lead to harrassing a person in severe cases. According to Pew 2017 Harassment Report every one in five users are subjected to being called offensive names online. After witnessing the harassment of others 27% refrained from posting online and 13% stopped using an online service. Some comment sections turn so toxic, filled with boorish behavior, that many turn off the comments section.

The project introduces the idea of using NLP and Machine Learning techniques to identify such toxic comments. This can help the moderator to delete such bad comments or have a bot autodelete comments that are tagged as toxic. A naively trained model will have some unintended biases in classifying. For example, 'The Gay and Lesbian Film Festival starts today.'. This sentence would get a high toxicity score because in general, 'gay' and 'lesbian' are used frequently in toxic comments on online boards. This project aims to mitigate these over generalizations.

Related Work

There are tons of papers and work going on in general that looks at texts for sentiment analysis but abuse classification is relatively new in the NLP domain. Abuse classification research with ML began by using Support Vector Machines as the classifier. But more recent studies have shown that deep learning performs much better instead when it comes to sentiment analysis. Recurrent Neural Networks worked well for poetry generalization as RNN uses a sequencing model. Other researchers have used convolutional neural networks for sentiment analysis. Even this approach does significantly better. Looking at the performamces of these two approaches it is clear that RNNs and CNNs are state of the art architectures for sentiment analysis. Given the similarities between toxic comment classification and sentiment analysis, we hope to use this research to fine tune our approach and methodology.

Currently a competition is being hosted by ConversationAI team with a prize pool of $35000 to build a multi-headed model thats capable of detecting different types of of toxicity like threats, obscenity, insults, and identity-based hate that performs better than their current models.

Methodology

As the first step in the project, we plan to perform exploratory data analysis and pre-processing on the corpora that we have collected so far. The dataset can be split into training, cross-validation, and test sets. Following this, we plan to process the data using NLTK - tokenization, indexing etc. Once we have sanitized and normalized the data, we can apply the commonly used machine learning techniques like Naive-Bayes, Support Vector Machines in order to define a suitable benchmark that we can improve on. Given the success that deep learning models like LSTM have had with most other natural language processing problems, we plan to evaluate the technique in this scenario.

Datasets and Evaluation

The main challenges associated with applying machine learning techniques to determine toxicity are the dearth of publicly available human-labeled data and the inherent bias in the labeling. Considering that the problem of identifying malicious comments is extremely common these days, given the proliferation of social media and message boards, there is a significant amount of public data. These include Wikipedia Human Annotations of Personal Attacks on Talk Pages, Wikipedia Human Annotations that have been made available via the ConversationAI project. These corpora include over 100k labeled discussion comments that are tagged manually using a crowd-sourcing platform. Some of these datasets also include demographic data about the workers involved in the labeling process.

The performance of the models can be tested on the validation set. We can create a baseline using a Naive Bayes model and determine the performance of other models in relation to that. Apart from the accuracy of the model, we can also examine the AUC plots for the models. Lastly, we can also test the best model against the test set published by the ConversationAI team for the online competition.

References

1. Ex Machina: Personal Attacks Seen at Scale - Ellery Wulczyn, Nithum Thain, Lucas Dixon - link
2. Toxic Comment Classification Challenge - Kaggle - link
3. Wikipedia Talk Labels: Personal Attacks - link
4. Wikipedia Talk Corpus - link
5. Wikipedia Talk Labels: Toxicity - link
6. Unintended Bias Analysis - Perspective - link