Toxicity

Luke Newman

Project Design

Being able to maintain civility in online environments is a growing problem in our digital age. I believe everyone has the right to engage in online conversations without the fear of being harassed or the target of unwarranted abuse. In this project I made a model that can detect toxic comments in hopes that it can be used to improve civility online.

The process into obtaining my final model consisted of several steps including careful inspection of each model to understand it as best I could. The first step was preprocessing the text data. This included removing unwanted characters, tokenizing the text, removing stop-words and contractions, part of speech tagging, and lemmatization. I made a new column in my DataFrame at each step to make it easy to go back to a certain step if further preprocessing was required. The next step was to make my Minimal Viable Product. I used simple embeddings here and tried not to do anything fancy. I used CountVectorizer for embeddings, LSA to reduce dimensions so I can visualize my data and see if there is any separation. Then used Logistic Regression because of its short fitting time and relatively easy explain-ability. Then inspected the model by looking at the confusion matrix and the most important words my model was using to classify toxic comments. I used this same process for two more models except I changed the embeddings using Tf-IDF and Word2Vec. Tf-Idf embeddings with logistic regression boosted the performance of my model. Word2Vec did significantly worse than both TF-IDF and CountVecotirzer. I tried using random oversampling to handle the class imbalance as well. My final models were LSTM neural networks. One with transfer leaning and one without. The Neural Network with transfer learning improved my model significantly and was my final model.

Now back to how my model can be used to improve civility online. For online gaming, you can detect toxic users and then the gaming company can decide what sanctions should be placed on these users. However, for social media platforms you might not want to limit people’s freedom of speech so I proposed an idea of a toxicity tracker which will tell you how toxic your comment is in real time as you type it. With this people can be more self-conscious of what they’re posting and hopefully will second guess the toxic comment they were about to post.

Tools

- Python
  - Pandas, sklearn, seaborn, imblearn, keras, nltk, matplotlib, numpy
- PowerPoint

Data

In this project we are downloading a dataset from kaggle. A Civil Comments platform which shutdown in 2017 releasing their ~2million comments to the general public which we are going to use to classify toxicity. Our data has many columns but the only ones we're going to look at is the comment text and the target. This data has been originally labeled for how toxic a comment is on a scale from 0-1. We make a binary class from the target column with <0.5 being labeled as 0 (Not Toxic) and >=0.5 being labeled as a 1 (toxic).

Further Work

For further work, I'd want to build an unbiased model. my model classified some comments as toxic Just because words identifying marginalized groups were mentioned. To do this we would need more data pertaining to non-toxic comments with specific identity words. Also, I would like to build the toxicity tracker I talked about in the last slide in hopes that we can make our online environments less toxic and more civil.