CS 410 - Progress Report

Team Information

Team Name: Waiting for the Next MP
Captain: Charlene Zhang (yuqianz6)
Team Members: Gregory Znoyko(gznoyko2), Joel Feddes (jfeddes2), Itay Gozalzani(itayg2)
Free Topics: Hate Toxic Comment Detector

Completed Tasks

Our goal over the past 3 weeks has been to finish the first three milestones we had set in our proposal. The first milestone involved researching and finding an appropriate dataset to use for our topic. We choose the Toxic Comment Classification Challenge dataset, as well as a dataset for stop words. We chose this dataset as it was a labeled dataset that contained both non-toxic and toxic behavior to help develop our models. We got a dataset with stop words to help filter them out/account for them for our TF-IDF weighting. After getting these datasets, we proceeded to do some data pre-processing and explore the data before cleaning and tokenizing the data. We noticed that there were many instances of different words being used as the same word with an example being YOU, You, and you. We addressed this issue by performing data transformation before tokenizing the data. After this, we went ahead and implemented our TF-IDF code and got word embedding in as well.

Future Tasks

For our future work, we plan to finish the last two milestones in our proposal. This means we need to begin working on the model training and tuning so we can most accurately determine when toxic behavior is demonstrated or not. We plan to train and evaluate three kinds of models: traditional ML models (Naive Bayes or SVMs), deep learning models, transformer + adversarial training, then observe their performance. The last milestone involves testing the model as well as evaluating the accuracy, recall, precision, F1 scores, etc. We still aim to have our model reach an accuracy rate of 80-90%. This should all take approximately 30 hours to accomplish.

Challenges and Solutions

We faced several challenges during the first three milestones of our project. One of the key challenges was setting up the Colab Python notebook so that we could access the dataset and work on the code simultaneously. We overcame this hurdle by storing the Kaggle API in a shared folder and loading the dataset in Colab. During the pre-processing phase, one of these was the previously mentioned problem involving instances of different words being used as the same word with one example being YOU, You, and you. Additionally, we identified some characters such as "/n" and " " that were not relevant and needed to be removed after tokenization. We tackled this by normalizing and removing all of these noises.

Team Contributions

Yuqianz6: Completed environment setup, data loading, data preprocessing; helped revise the report

Jfeddes2: Implemented TF-IDF weighting for tokenized strings, word embedding Word2Vec

Gznoyko2: Helped review the code and write up the report