CS 410 - Project Proposal

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

Introduction

Social media is a double-edged sword. While it serves as a platform for individuals to freely express their opinions, it also paves the way for negative experiences. Some people abuse the liberty and anonymity nature of the internet to disseminate verbal harassment, hate speech, racist remarks, obscene content, and the like. In recent years, an increasing number of scientists have employed NLP to delve into the classification of toxic comments. However, such research has not yet been widely adopted in the market. Recognizing the need to create a safer and more respectful online environment, our team proposes the development of a real-time detector aimed at identifying and flagging toxic comments.

Objective

Our project seeks to harness the capabilities of Natural Language Processing (NLP) to categorize and highlight comments that may be deemed offensive or harmful. By integrating this system into social media platforms, such as YouTube, we aspire to prompt users to reconsider their choice of words before submitting potentially harmful content. The ultimate goal is to reduce the incidence of online harassment and create a more inclusive digital community.

Data and Methodology

We will utilize the dataset from the Toxic Comment Classification Challenge, which comprises comments annotated by human raters based on various categories of toxic behavior. Using this dataset and our program, we intend to find and flag any example comment or social media post that includes toxic content. We will use Python and below are the approach details:

1. Data Preprocessing:
    a. Cleaning Text: Remove irrelevant characters, numbers, or symbols, and convert all cleaned text to lowercase.
    b. Text Tokenization: Convert text into tokens.
2. Feature Extraction: TF-IDF, and other word embedding models such as Word2Vec.
3. Model selection and Training: Model TBD.
4. Result Evaluation and Model Turning

Challenges and Solutions

One of the primary challenges we anticipate is the subjective nature of "toxic content" and determining the appropriate weight for each word or phrase. To address these issues, we will leverage our labeled dataset, incorporate N-grams, and utilize word embeddings to enhance the contextual understanding of our model.

Expected Outcomes

We are targeting an accuracy rate of 80-90% in identifying toxic comments within our test sets. The reason why we chose this range was that we wanted to have higher recall over high precision. Since we are only flagging and warning the user of the existence of toxic words in their message, we are more ok with having false positives over having false negatives. Through this initiative, we hope to contribute to the ongoing efforts to mitigate online harassment and foster a healthier, more respectful virtual discourse.

Timeline and Workload

Our project is planned over an 80-hour period, distributed as follows:

- 20 hours of in depth research into the dataset and what resources we have to interpret and utilize it
- 40 hours of actual implementation by the entire team
- 20 hours of testing to finalize and generate a complete product