Chat Application with Hateful Text Classification

Anvita Karne
CS Department
Virginia Tech
Falls Church, VA, USA
anvitakarne@vt.edu

Atharva Chouthai
CS Department
Virginia Tech
Falls Church, VA, USA
atharvachouthai@vt.edu

Kshitij Narvekar
CS Department
Virginia Tech
Falls Church, VA, USA
kshn@vt.edu

Rajat Belgundi
CS Department
Virginia Tech
Falls Church, VA, USA
rajatrb@vt.edu

Yash Kulkarni
CS Department
Virginia Tech
Falls Church, VA, USA
yashpandharish@vt.edu

ABSTRACT

The term "hate speech" refers to any kind of communication, whether spoken, written, or symbolic, that denigrates, threatens, or incites violence or hostility towards one or more individuals or groups on the basis of that person's race, ethnicity, religion, sexual orientation, gender identity, or other characteristics. This project aims to solve the problem of mitigating hate/toxic comments on a messaging application by flagging a toxic message. For this project we intend to build a web application analogous to a basic messaging app that will categorize hateful comments into “toxic”, “severe toxic”, “obscene”, “threat”, “insult”, and “identity hate”. Bidirectional Encoder Representation of Transformers (BERT) will be used for classifying hateful comments in this project which has proven to be an effective tool for various NLP applications.

KEYWORDS

BERT, Toxic comments

1 Introduction

Cyberbullying is the practice of harassing, intimidating, or harming others via digital communication technologies like social networking platforms, instant messaging, or forums on the internet. It involves the deliberate and repetitive use of technology to target someone, causing them emotional distress,embarrassment, or social isolation. Cyberbullying can take various forms, including sending threatening or derogatory messages.The motivation behind this project is to try and mitigate the online abuse which is caused by messages on social media messaging apps. None of the existing harmful content and enabling users to have more courteous and productive interactions, you may improve the general user experience on online platforms. The issue of toxic remarks and hateful comments has also arisen as a result of the increased internet contact. Online groups and people can be severely harmed by toxic remarks, which can contain a variety of disrespectful and unpleasant words such as extreme toxicity, profanity, threats, insults, and identity hatred.

The goal of this project is to create a reliable and accurate toxic comment classifier that can recognise and identify toxic remarks according to certain criteria, such as identity hatred, obscene, threat, toxic, and severe toxic. By tackling this issue, we want to:

1. Improve Online Safety: By automatically recognising and flagging toxic remarks, you can make the internet a safer place for everyone while also promoting healthy online discourse.
2. Enhance User Experience: By removing harmful content and enabling users to have more courteous and productive interactions, you may improve the general user experience on online platforms.
3.Moderation Efficiency: By automating the first identification and classification of poisonous remarks, platform administrators and moderators may more effectively moderate user-generated material.

For this project, we aim to use a multi-labeled BERT model which is capable of detecting different types of toxicities like obscenity, insults and identity-based hate and fine-tune the

model according to our need. We would be training the model on a combination of datasets covering a variety of platforms along with an adequate number of comments / messages. In our chat application, we are developing a dual functionality:

1. Toxicity amount Flagging: We will show a sliding bar reflecting the amount of toxicity of the content being written as the user inputs words in the chat window. The user will receive immediate feedback about the possibly hazardous stuff they are accessing.
2. Text Categorization: When a user submits a message, we will classify it into one of six toxicity categories that have been previously established. Users and moderators will be able to swiftly recognize and address various forms of offensive or dangerous material with the use of this labeling system.

The model latency as of now is under 0.1 seconds to give out scores for string length of about 18 characters excluding whitespaces. So the trained model would be available on the server and we would run a python script to get a score of the available input string after word completions.

In conclusion, our chat application will classify user messages into one of the six toxicity categories and provide real-time feedback on the comment's toxicity level.

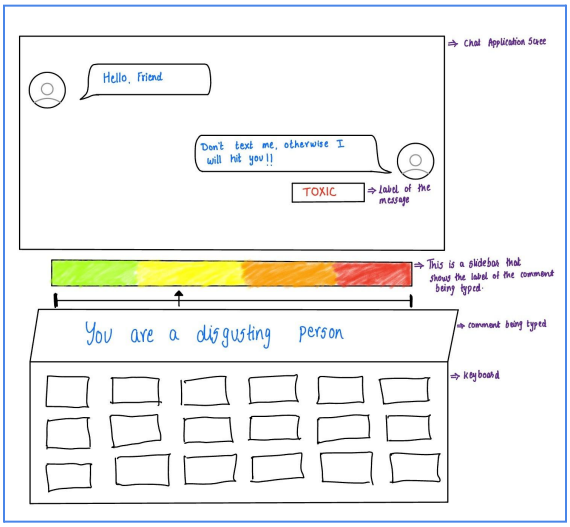

Figure 1

After the literature survey, we found a need for a medium to classify hateful comments posted on social media applications to mitigate the spread of online hate. Hence, decided to proceed with fine-tuned Distil BERT for classification of comments and develop a web-based chat application for flagging the comments.

2 Related Work

Hate speech, which insults a person or people based on their race, religion, gender, or handicap, is a common issue. In the paper [1], they have performed a sentiment analysis on two data sets; one data set is collected by tweets made by people from all over the world, and the other data set contains the tweets made by people of India. In the paper [2], the performance of BERT-based classifiers is then compared with that of bag-of-words approaches to determine the effectiveness of BERT-based classifiers. The evaluations performed using Yelp shopping reviews show that fine-tuned BERT-based classifiers outperform bag-of-words approaches in classifying helpful. The authors in [3] propose to use the BERT model in combination with the Bayesian network to build a binary classifier. They have used a Bayesian network to first classify the comment in two categories to get its approximate category and then deployed a BERT model to predict the exact category of the text. In Paper [4], they have built a hate-o-meter based on Convolutional Neural Network, in combination with NLP techniques. Through the model, they have displayed the percentage of hate displays the bias of the statements. Their model classifies text with 80.35% accuracy. The research work in paper [5] proposes a tool to raise awareness on the persistent hate speech in blogs, online-forums, and newspapers. The primary aim of this research work is to highlight the content that promotes violence or hatred against individuals or groups based on religion, gender, ethnicity or disability.

REFERENCES

[1] Singh, Mrityunjay, et al. "Sentiment Analysis on the Impact of Coronavirus in Social Life Using the BERT Model." Social Network Analysis and Mining, vol. 11, no. 1, Springer Science+Business Media, Mar. 2021, https://doi.org/10.1007/s13278-021-00737-z.

[2] Bilal, Muhammad, and Abdulwahab Ali Almazroi. "Effectiveness of Fine-tuned BERT Model in Classification of Helpful and Unhelpful Online Customer Reviews." Electronic Commerce Research, Springer Science+Business Media, Apr. 2022, https://doi.org/10.1007/s10660-022-09560-w.

[3] Liu, Songsong, et al. "Text Classification Research Based on Bert Model and Bayesian Network." IEEE Xplore, Nov. 2019, https://doi.org/10.1109/cac48633.2019.8996183.

[4] A. Chaudhari, A. Parseja and A. Patyal, "CNN based Hate-o- Meter: A Hate Speech Detecting Tool," 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 940-944, doi: 10.1109/ICSSIT48917.2020.9214247.

[5] F. M. Plaza-Del-Arco, M. D. Molina-González, L. A. Ureña- López and M. T. Martín-Valdivia, "A Multi-Task Learning Approach to Hate Speech Detection Leveraging Sentiment Analysis," in IEEE Access, vol. 9, pp. 112478-112489, 2021, doi: 10.1109/ACCESS.2021.3103697