Detecting Toxicity in Social Media

Sindhuja Kodaparthi (skodapar@gmu.edu)

Venkata Nikhil Chakravarthy Korrapati (vkorrap@gmu.edu)

Saranya Chakravarthy Korrapati (skorrap@gmu.edu)

Motivation

Problem to Solve:

- Cyber-bullying has become a pervasive issue across social media, contributing to mental health problems and fostering toxic online environments.

Goal:

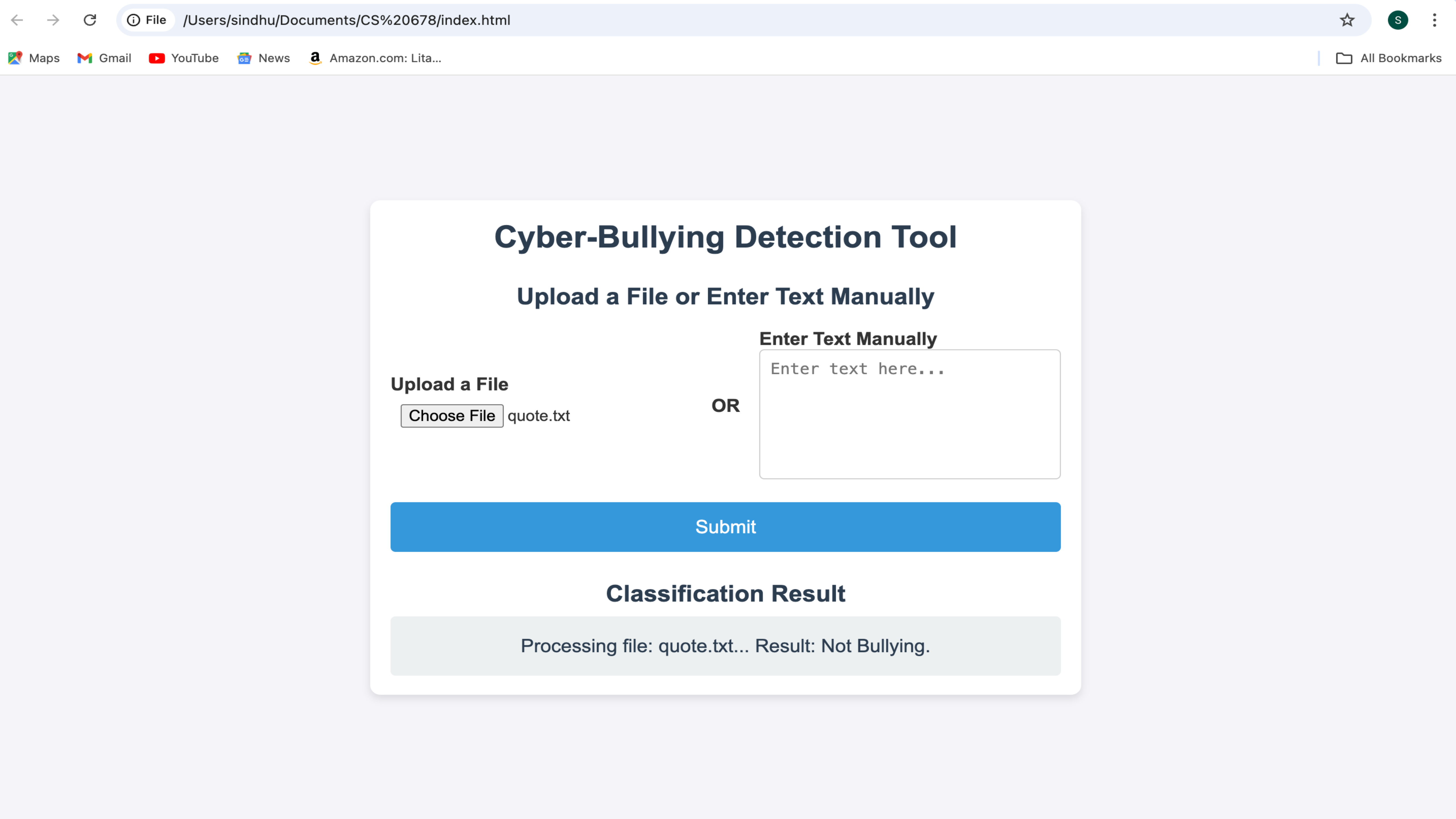
- The objective is to develop a model that can reliably detect cyber-bullying in various forms, including hate speech, aggression, insults, and toxicity, within text data.
- The goal is to ensure the model achieves high accuracy in correctly classifying whether a given piece of text contains cyber-bullying or not.

Target Audience

- Social Media Platforms
  - Automate the detection and removal of harmful content, reducing cyber-bullying, leading to healthier online communities.
- Law Enforcement and Legal Authorities
  - This tool can help track and report cyber-bullying incidents. It can provide evidence for legal investigations and support the enforcement of laws against online harassment.
- Educational Institutions
  - Monitor online interactions among students, helping to identify and address cyber-bullying incidents, thus ensuring a safer digital environment for students.

Due Diligence

- The data for this project will be sourced from publicly available datasets, with platforms like Kaggle offering datasets, while the remaining sources, such as Twitter, Wikipedia Talk pages, and YouTube, are the media from which the data is collected.
- Compute requirements will depend on the scale of the model and the dataset size.
- Will be using pre-trained transformer model like BERT for text classification task. If the pre-trained models do not yield satisfactory performance, will be fine-tuned further to better detect cyber-bullying in specific contexts.
- The development of this machine learning project is expected to take around 2 months, including tasks such as data collection, feature engineering, model training, evaluation, and deployment. Each phase may vary in duration based on the complexity and specific requirements of the project.



Cyber-Bullying Detection Tool

Upload a File or Enter Text Manually

Upload a File

OR

Enter Text Manually

Classification Result

Processing file: quote.txt... Result: Not Bullying.

Evaluation

Metrics

- Accuracy
- Precision
- Recall

Unit Tests

- Upload functionality
- Text Preprocessing
- Model prediction