1. Bio

Group Name: VioletGo

Member: Xinyu Wang

Email: xw3080@nyu.edu

Country: United States

Institution: New York University

Specialization: NLP

2. Problem Description

Our aim is to categorize comments and phrases from Twitter based on their potential as hate speech. This project involves training a Transformer model to discern if a specific Twitter comment should be flagged as hate speech, i.e., an offensive attack from an individual or group.

3. Business Understanding

Developing a model that can accurately identify hate speech on social media platforms can aid platform administrators in punishing misbehaving users. This not only safeguards the rights of users, fostering a more friendly and inclusive environment, but it also assists platforms in adhering to governmental regulations. Consequently, this can potentially attract more users, increase user retention rates, enhance user engagement, and ultimately, drive profits.

4. Project Lifecycle & Deadlines

The project is slated to run for one month:

- Proposal Submission: 20th June
- Data Intake Report: 27th June
- Data Preprocessing & EDA: 3rd July
- Initial Model Version: 8th July
- Final Model Version: 15th July
- Submission of Final Assignments: 20th July

5. Data Intake Report

Dataset: Twitter Hate Speech

Overview:

The dataset, purposed for sentiment analysis, comprises approximately 32k records. Each record signifies a unique text string tagged with a sentiment value.

Data Source: Twitter Hate Speech Dataset on Kaggle (https://www.kaggle.com/datasets/vkrahul/twitter-hate-speech?select=train_E6oV3lV.csv)

Data Collection:

The exact methodology behind the data collection is unspecified. More information would be essential to understand potential biases in the dataset.

Data Description:

The dataset is made up of three fields: 'id', 'label', and 'tweet'. The 'id' field acts as a unique identifier, the 'label' field contains the sentiment value, and the 'tweet' field houses the text data for sentiment analysis.

Data Quality:

The dataset appears well-curated with no apparent missing or mismatched values. However, some duplicate values in the 'tweet' column require further investigation.

6.Github Repository Link:

https://github.com/VioletGo319/sentimental_analysis