Team: Overfitters Submission (Phase-1)

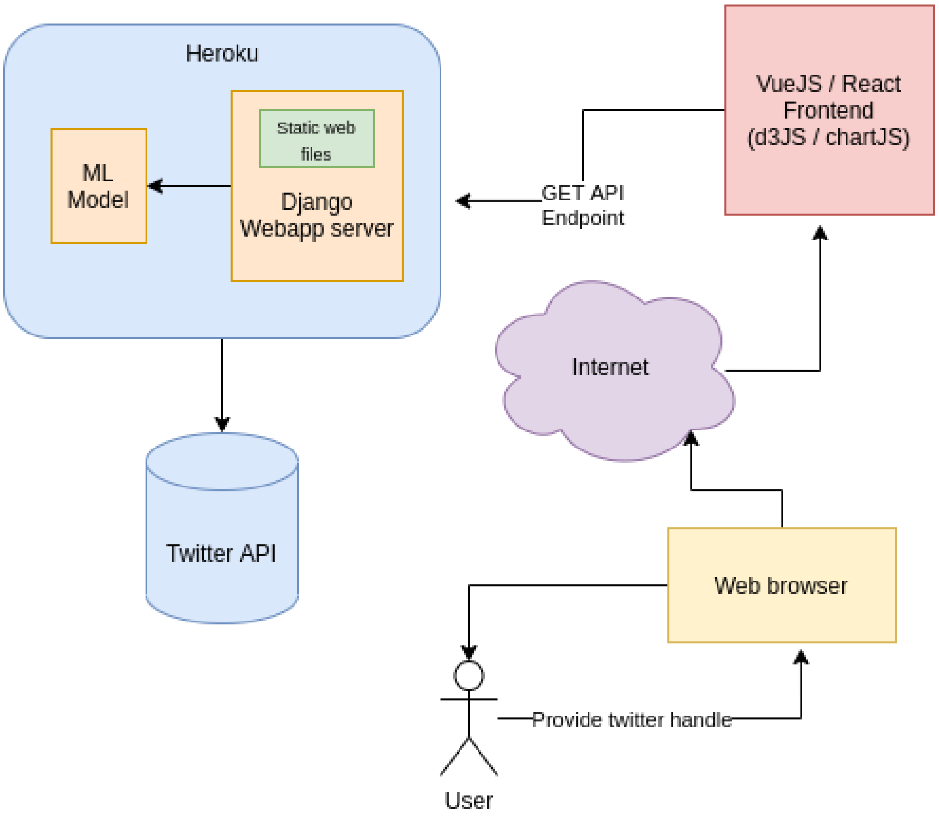
Developers:

- Mohit Balwani
- Mrulay Mistry
- Nauman Mansuri
- Prayush Dawda
- Safwan Mansuri

Problem statement

Nowadays, all the things are going online. Major concern of everything going online is safety, especially women safety. There are many issues people face on online platforms like trolling, teasing, etc. Can you come up with anything which can assure safety on such platforms. It need not necessarily be mobile or web interface but anything implemented by yourself.

Ideation



System Architecture

1. Users provide their twitter handle via the frontend in the web browser.
2. The frontend will make a GET API request to the backend along with the twitter handle.
3. The backend will make an API request to twitter's API endpoints in order to fetch all tweets/comments that mention the given twitter handle.
4. This data will be preprocessed and provided to the Machine Learning model which will then return sentiments of all the tweets in JSON format.
5. This data will then be visualized in the frontend using d3JS / chartJS. The frontend will display the following
    a. Stacked bar Graph of comments vs mean comments per day
    b. Most glorified haters list
    c. Most common words (Word Cloud)
    d. Most common region (If available using the twitter API)
    e. Kind of hate speech (offensive/hate speech/etc.)
    f. Common Hashtags
    g. Number of mentions in abusive posts

Sentiment Analysis

The way we have decided to detect trolls and teasers on social media is by using Natural Language Processing (more specifically, sentiment analysis). For this we needed an already annotated dataset. We chose to work on 2 datasets for this purpose:

1. https://data.world/thomasrdavidson/hate-speech-and-offensive-language: A data set of tweets in English Language that can be useful in detecting hate speech and offensive language.
2. https://docs.google.com/forms/d/1Y-JEdtEc6syMuxVB4oXNqmjUHOaymgA0GUFoqoyMalc/viewform?edit_requested=true: A data-set with Hindi-English mixed dialect to detect if any harassment is done by trolls in Hinglish Language.

Preprocessing: In order to pre-process these tweets, we need to perform a number of steps including, removing HTML tags, tweet mention tags, removing stop-words etc.

Modelling: For detecting hate speech in English and Hinglish, we plan to use two different models. To detect which language the tweet is in, we will use TextBlob and then vectorize them to perform sentiment analysis using machine learning or deep learning in order to detect if there is any trolling or teasing happening. We plan to use models such as Naive Bayes, BERT, RNNs, etc.

APIs

In order to extract tweets from the user given input (i.e. their social media handles), we will use social media APIs such as Twitter Developer API, Facebook API for Facebook and Instagram etc.