Hate speech detection as a chrome extension
Solution Outline

Dhar Padma Patanjali (2018201011)
K A Meghashree (2018201055)

Our System:

Our proposed system can detect hate speech and toxic content from social media data.

Implementation Details:

- The idea is to start by building a basic chrome extension outline using HTML and Javascript.
- Dataset-
    - We shall scrape data from various platforms like reddit and facebook amongst other social media.
    - For labelling the data that we collect, It would be using a crowdsourcing technique to decide if the speech is hate speech or not. This can be done by showing the sentences to students in our class and asking them for labelling it as hate speech or not.
- Then, we shall start including different hate speech detection codes into this extension and test its working. The goal is to make this chrome extension as versatile as possible.
- The final step would be to implement the following two instances:
    - One instance shall be when the user is typing any text in a text box to post it on any social media like facebook. When the person is going to post to the platform, if he/she has written a hate speech, the extension shall give a pop up. Please refer to the mockup for the example.
    - Second instance shall be when we open a web page to read. If the page contains contents of hate speech, there shall be a pop up to warn us to proceed with caution and that hate speech has been detected.

Progress so far:

We have built a basic chrome extension outline. We are currently exploring different hate speech algorithms that can be included in our project. We are also in the process of collecting data from social media sites like facebook and reddit.

Evaluation metrics:

This shall include the accuracy of whether the speech is hate speech or not, the precision of detecting hate speech and the recall of the same. That way we would know if our model is detecting hate speech as a percentage of the entire dataset, if the incorrect labels are less as a percentage of the hate speech instances or more and if we are incorrectly labelling non hate speech as hate speech and vice versa.

Mockup of the system is as given: