Hate speech detection as chrome extension
Project Scope

Dhar Padma Patanjali (2018201011)
KA Meghashree (2018201055)

Our System:
Our proposed system can detect hate speech and toxic content from social media data. We shall be making a chrome extension that is versatile, can detect hate speech on various platforms and is highly accurate with a high recall and a high precision. The interface shall be kept extremely simple and user friendly. At the backend, various models can be used eg. google perspective API.
A novel feature that we are thinking of implementing if possible and time permits is to recommend a less toxic or a completely non toxic version of what the user wants to convey.

Our model will be working on two stances:

1. 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.
2. 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.

Other implementations for detecting toxic speech:
Other implementations of the toxic speech detection included approaches from simple and easy detections like maintaining a dictionary of profane words and certain stereotypical words and the section of the society which is victimised via those stereotypical words. This however, cannot truly decipher if a speech has elements of toxicity and hate to it or not, as hate speech may not always require negative or stereotypical words. More complex and higher accuracy models include SVM, Naive Bayes with TFIDF features , Logistic regression among others. However, none of these models could give an alternate to the speech detected as hate speech, that is toning down the language to make it less hateful and less radicalised. Eg. Instead of generalising an entire section of the society, we could change it to some members of the particular section of the society may commit an act.

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.

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: