HATE SPEECH DETECTION AS A CHROME EXTENSION

PADMA DHAR (2018201011)

MEGHASHREE K A (2018201055)

Abstract

Our proposed system can detect hate speech and toxic content from social media data. Our chrome extension is versatile, can detect hate speech on various platforms and is highly accurate with a high recall and a high precision. The interface is extremely simple and user friendly. When the user is typing any text in a text box to post it on any social media like facebook, if the person is going to post a hate speech to the platform, the extension shall give a pop up.

Literature Survey

Detecting hate speech is very important in various fields like controversial event finding, reporting and blocking users that spread hate and misinformation amongst the masses. The complexity of the task is very challenging as we need to take care of natural language constructs.There are many baseline methods for detecting hate speech. The most naive methods simply suggest classifying words as good or bad. So if anyone uses a word/words that are included in the bad word dictionary, they would be informed of it. There is presently a google chrome extension working mostly on this principle called Hate Free. The developers have maintained bad words and other words which if appear along with it may be hurtful to some section of the society.
We read some technical papers about various hate speech detection techniques. They are as follows:

1. HateMiners: Detecting Hate Speech Against Women

In this paper, the authors have classified posts as misogynistic in twitter data set by generating
sentence embeddings which generated a vector representation of the text. TFIDF vectors were generated on the preprocessed text followed by BoWv(Bag of words vector) for each tweet. These were passed through a logistic regression classifier to classify tweets as misogynistic.

2. Hate Speech Classification in Social Media Using Emotional Analysis

In this paper, the authors have presented a combination of lexicon-based and machine learning approaches to predict hate speech contained in a text, using an emotional approach through sentiment analysis.

Using the emotional information contained in text helps to increase the accuracy on hate speech detection. This analysis still has limitations that lead to exciting future research directions. It is reasonable to question the definition of hateful content, in the sense that it is not clear what is the threshold a published text shared in social media has to violate to be considered hateful due to the subjectivity of the definition of hate-speech.

3. Identification of Hate Speech in Social Media

The authors in this paper have spoken about various models that can be used to detect hate speech including various lexical analysis models in which machines are fed patterns of languages to detect hate words and negative polarity words. The second approach they looked into was various machine learning techniques. Logistic regression with L2 regularisation and 10 fold cross validation was used for this purpose. They also looked into hybrid methods which used the combination of 3 different classification features (Maximum Entropy, SVM and Random Forest). They however found that the best performance was obtained using the Naive Bayes classifier with TFIDF features. Thus, they used both supervised as well as unsupervised methods to get results for detection of hate speech.

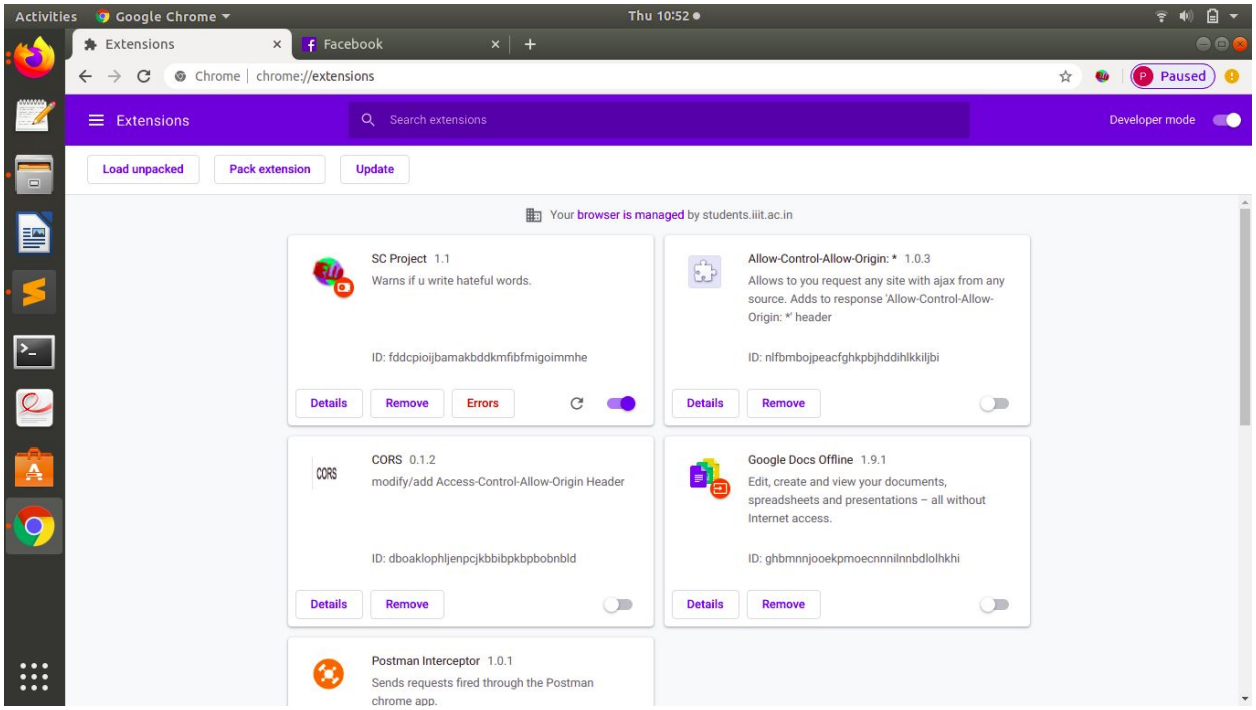
4. Hate speech detection: Challenges and solutions

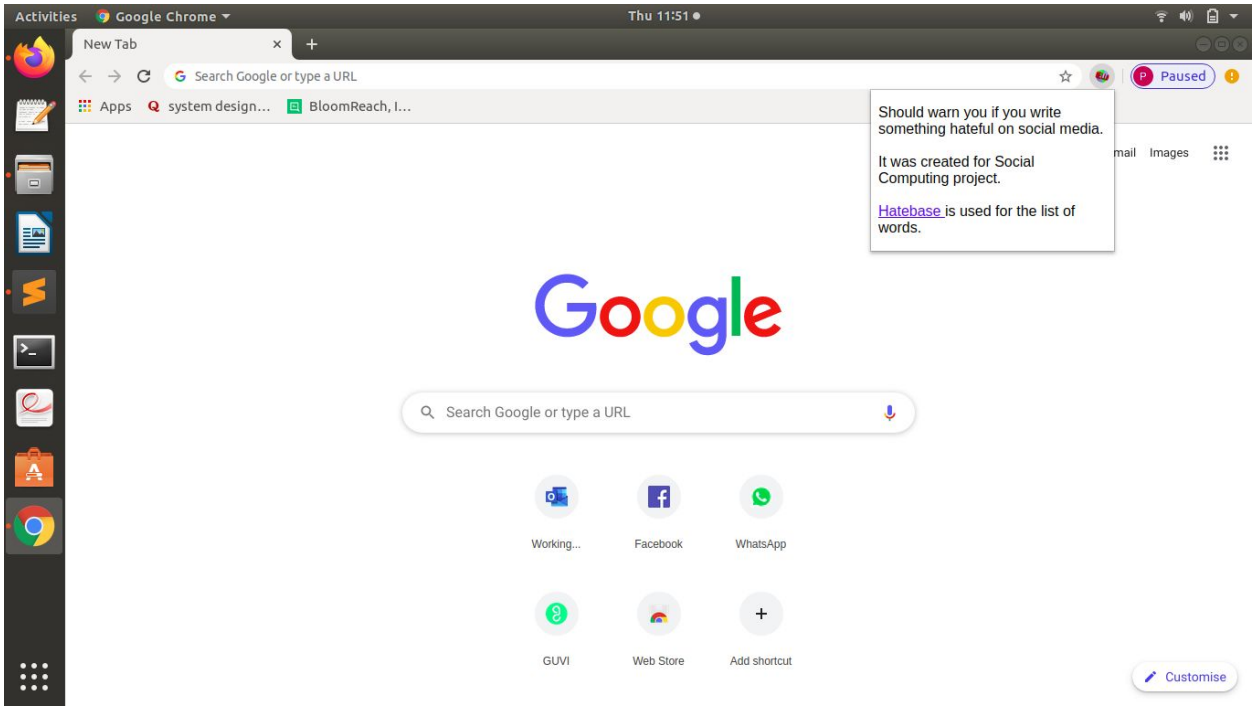
As online content continues to grow, so does the spread of hate speech. This paper identifies and examines the challenges faced by online automatic approaches for hate speech detection in text. Among these difficulties are subtleties in language, differing definitions on what constitutes hate speech, and limitations of data availability for training and testing of these systems. Furthermore, many recent approaches suffer from an interpretability problem—that is, it can be difficult to understand why the systems make the decisions that they do. The paper proposes a multi-view SVM approach that achieves near state-of-the-art performance, while being simpler and producing more easily interpretable decisions than neural methods. Both technical and practical challenges for this task are discussed as well.

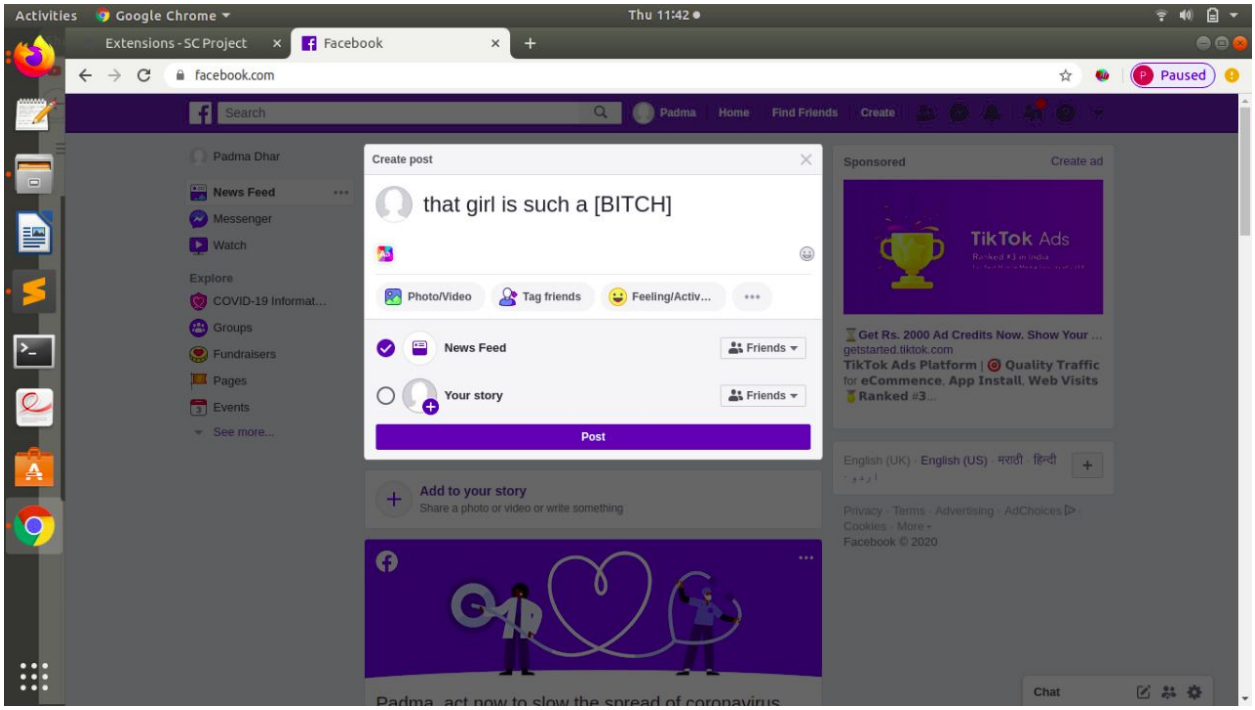
Methodology

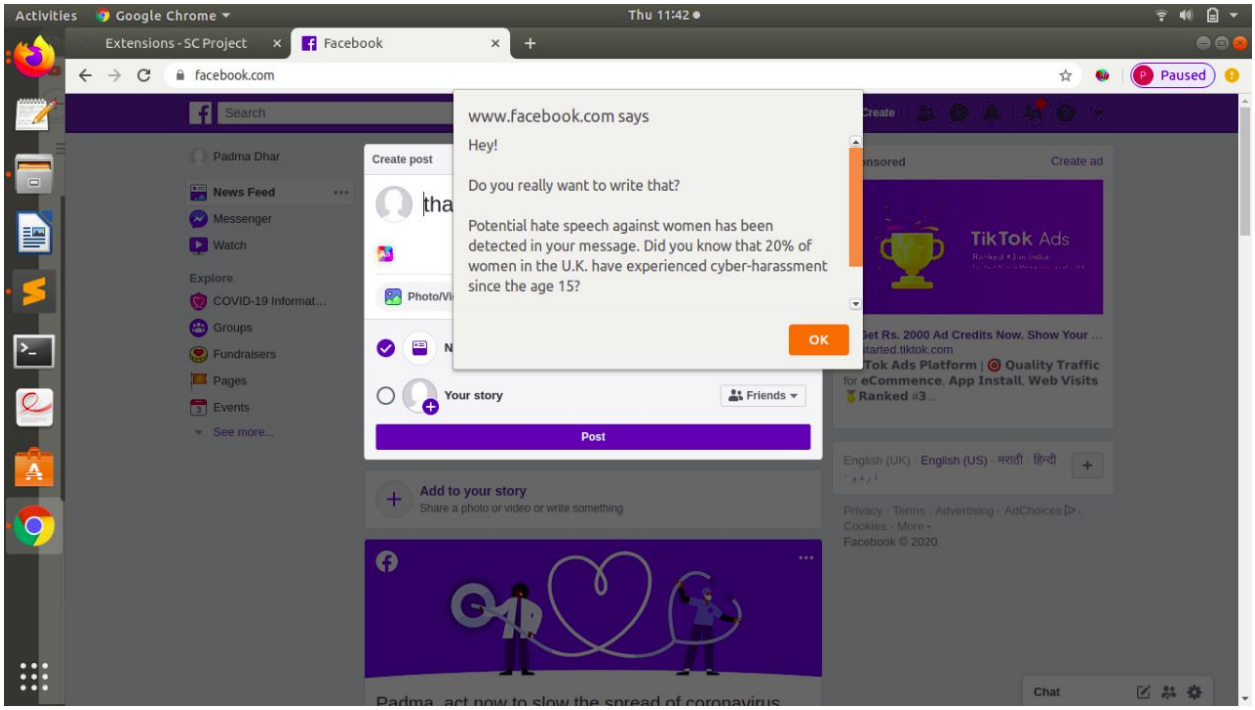
- The database used has been obtained from HateBase, an online repository.
- On a social media site like Facebook or Twitter, when the user types any inappropriate hate based content and clicks on post, our extension provides a pop up detecting hate speech
- We extracted a list of words that classify as “hate words” and we use it to match with the content typed by the user. However, this is to be considered only as a first level filter.
- Second filter we have included is the usage of a NLP Model based on the dataset of twitter.
- Some types of feature spaces were engineered including TFIDF, typed dependency and sentiment score.
- Preprocessing to remove punctuation and character conversion to lower case.
- Stanford Parser used to identify syntactic relations between words in sentences. It returns typed dependency for each tweet as a dictionary.
- The relationship within each tweet identified by the Stanford Parser resulted in features, with each storing the counts of each different typed dependency
- We also allocated sentiment scores as negative or positive.
- Normalization of positive and negative term frequency, TF-IDF scores, used as the only feature set to be included in the baseline model, are created based on the term frequency and inverse document frequency.
- For each tweet, TF-IDF values are calculated for each word and assigned a weight of 1 if the word appears in the hate base dictionary, 0 otherwise.
- The weightings are simply set at 1 and 0 since we're only concerned about words appearing in the dictionary. All TF-IDFs are then added together after multiplying their corresponding weight as TF-IDF score for each tweet.
- Our system takes any text written by the user on certain social media like facebook and using NLP techniques of preprocessing it gets the words individually which are then passed through the filters.
- The first filter checks on the content the user is writing.It checks for profane words from a dictionary of profane words provided by HateBase. On detection of an objectionable word, it will warn the user that they must not write such.
- The second filter shall read comments posted on facebook and shall warn which comments are safe, which are hateful and which are mildly offending. This is done via the algorithm explained above

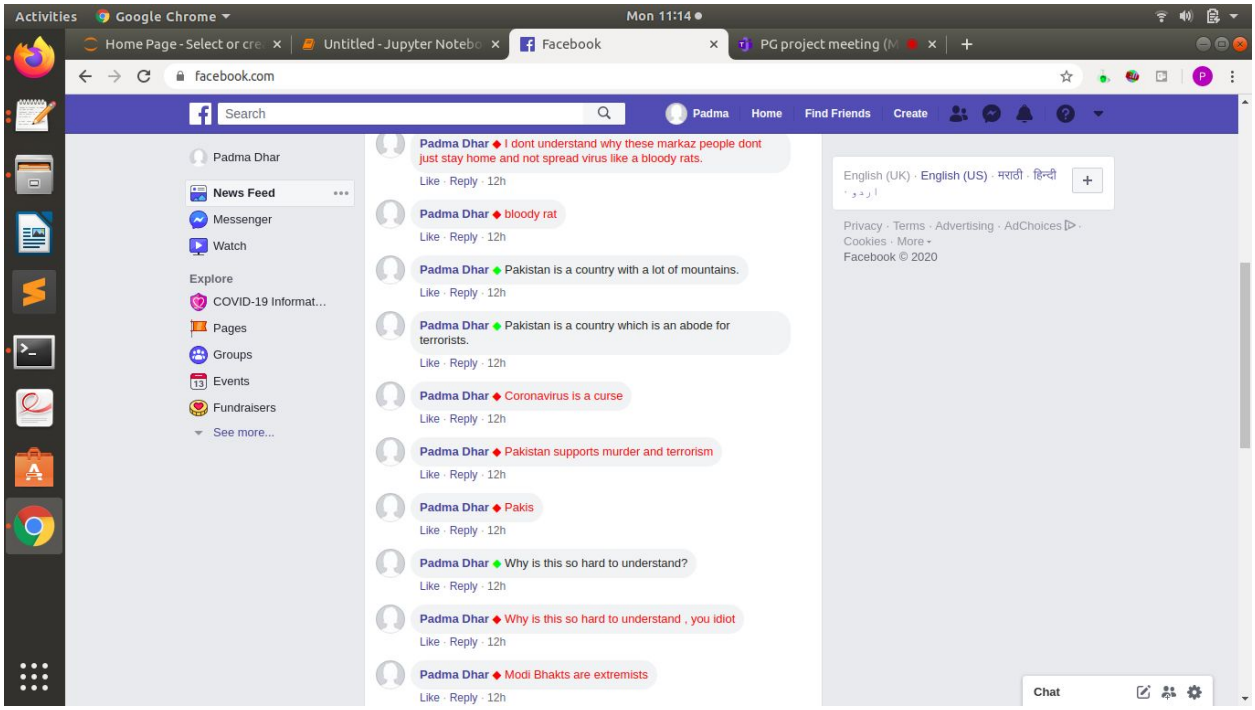
Screenshots






Activities
Google Chrome
Thu 11:42
Extensions - SC Project
Facebook
facebook.com
Paused
Search
Padma
Home
Find Friends
Create
Padma Dhar
News Feed
Messenger
Watch
Explore
COVID-19 Informat...
Groups
Fundraisers
Pages
Events
See more...
Create post
that girl is such a [BITCH]
Photo/Video
Tag friends
Feeling/Activ...
News Feed
Friends
Your story
Friends
Post
Add to your story
Share a photo or video or write something
Padma, act now to slow the spread of coronavirus
Sponsored
Create ad
TikTok Ads
Get Rs. 2000 Ad Credits Now. Show Your ...
getstarted.tiktok.com
TikTok Ads Platform | Quality Traffic for eCommerce, App Install, Web Visits
Ranked #3...
English (UK) · English (US) · मराठी · हिन्दी
Privacy · Terms · Advertising · AdChoices
Cookies · More
Facebook © 2020
Chat


Activities
Google Chrome
Thu 11:42
Extensions - SC Project
Facebook
facebook.com
Paused
Search
Padma Dhar
News Feed
Messenger
Watch
Explore
COVID-19 Informat...
Groups
Fundraisers
Pages
Events
See more...
Create post
Your story
Friends
Post
Add to your story
Share a photo or video or write something
Padma, act now to slow the spread of coronavirus
www.facebook.com says
Hey!
Do you really want to write that?
Potential hate speech against women has been detected in your message. Did you know that 20% of women in the U.K. have experienced cyber-harassment since the age 15?
OK
Sponsored
Create ad
TikTok Ads
Ranked #3...
English (UK) · English (US) · मराठी · हिन्दी
Privacy · Terms · Advertising · AdChoices
Cookies · More
Facebook © 2020
Chat



Challenges

It is not possible for the system to realise if the subject being talked about is a person, community or a negative thing that should be abolished.
Eg. Coronavirus is a curse that is shown as hate speech. If instead it was a person or a community, then it would be a form of hate speech, however, coronavirus is a negative thing that should be abolished.
Context may not always be implicit. We can never be certain how a person may react to a sentence. (Joke/insult/threat.)

Future Scope

Few improvements that can be made to this project are:

- 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.
- A novel feature would be to recommend a less toxic or a completely non toxic version of what the user wants to convey.

References

1. HateMiners: Detecting Hate Speech Against Women
https://arxiv.org/abs/1812.06700

2. Hate Speech Classification in Social Media Using Emotional Analysis
https://ieeexplore.ieee.org/document/8575590

3. Identification of Hate Speech in Social Media
https://ieeexplore.ieee.org/document/8615517

4. Hate speech detection: Challenges and solutions
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6701757/