Hate Speech Detection As A Chrome Extension

Padma Dhar
Meghashree K A

Link: https://drive.google.com/file/d/1oNpQAy8nBz-O6kxle5Gp2jrFD12mCJFt/view?usp=sharing

Scope

We are creating a chrome extension that can detect hate speech in social media posts and warn the user about the same.

What is Hate Speech?
Any hateful or toxic words especially oriented towards a group a people based on a common factor like race, religion, etc.

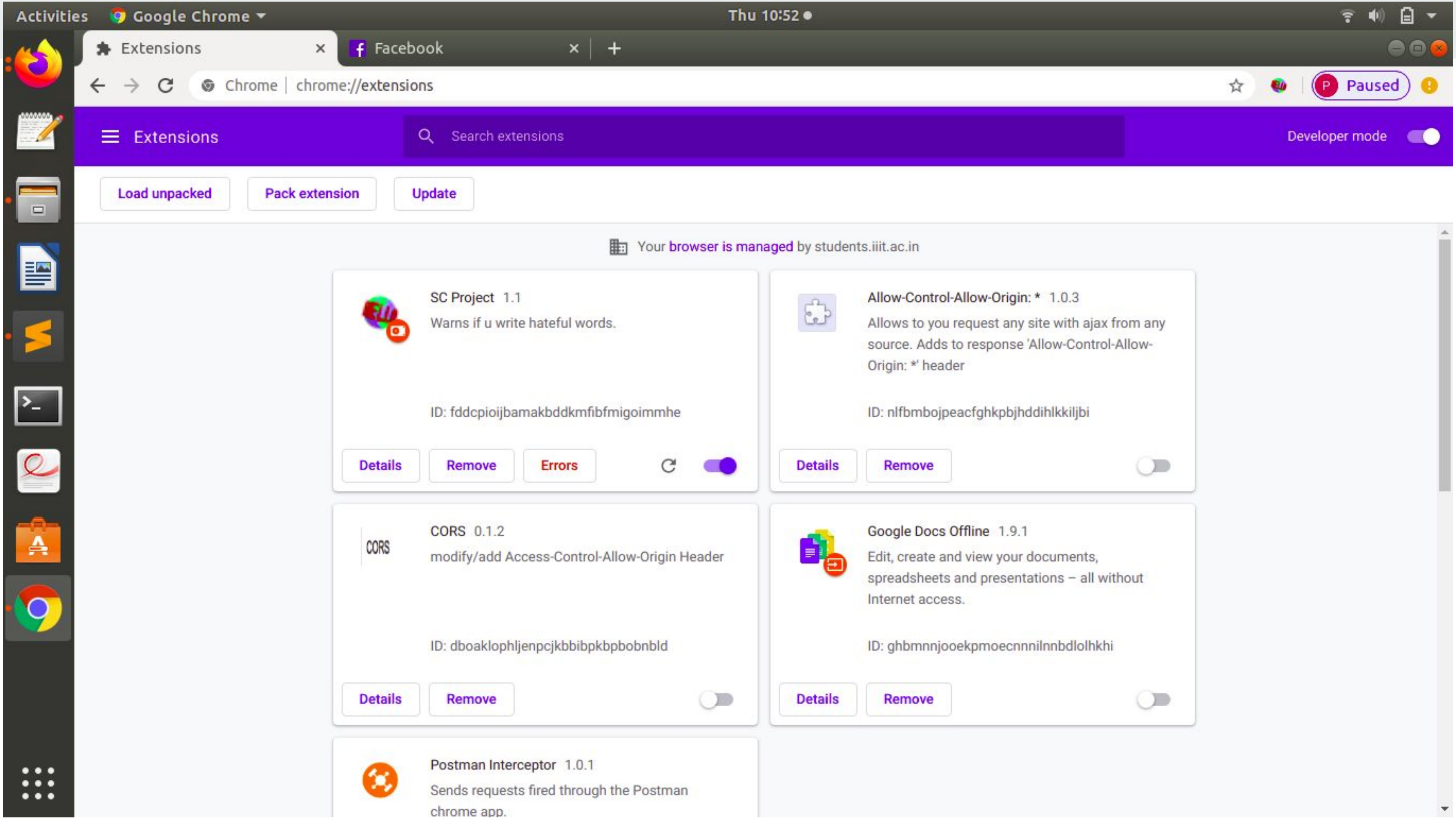
Why is it harmful?
It can cause riots and negative impression, negative feelings and animosity towards another group of people. It may incite racist attacks or others. It can cause losses for a company or a targeted by hate speech.

Work Done

- Created a basic working chrome extension for hate speech detection
- 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 have currently 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 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 relation between words in sentence. 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



Your browser is managed by students.iiit.ac.in

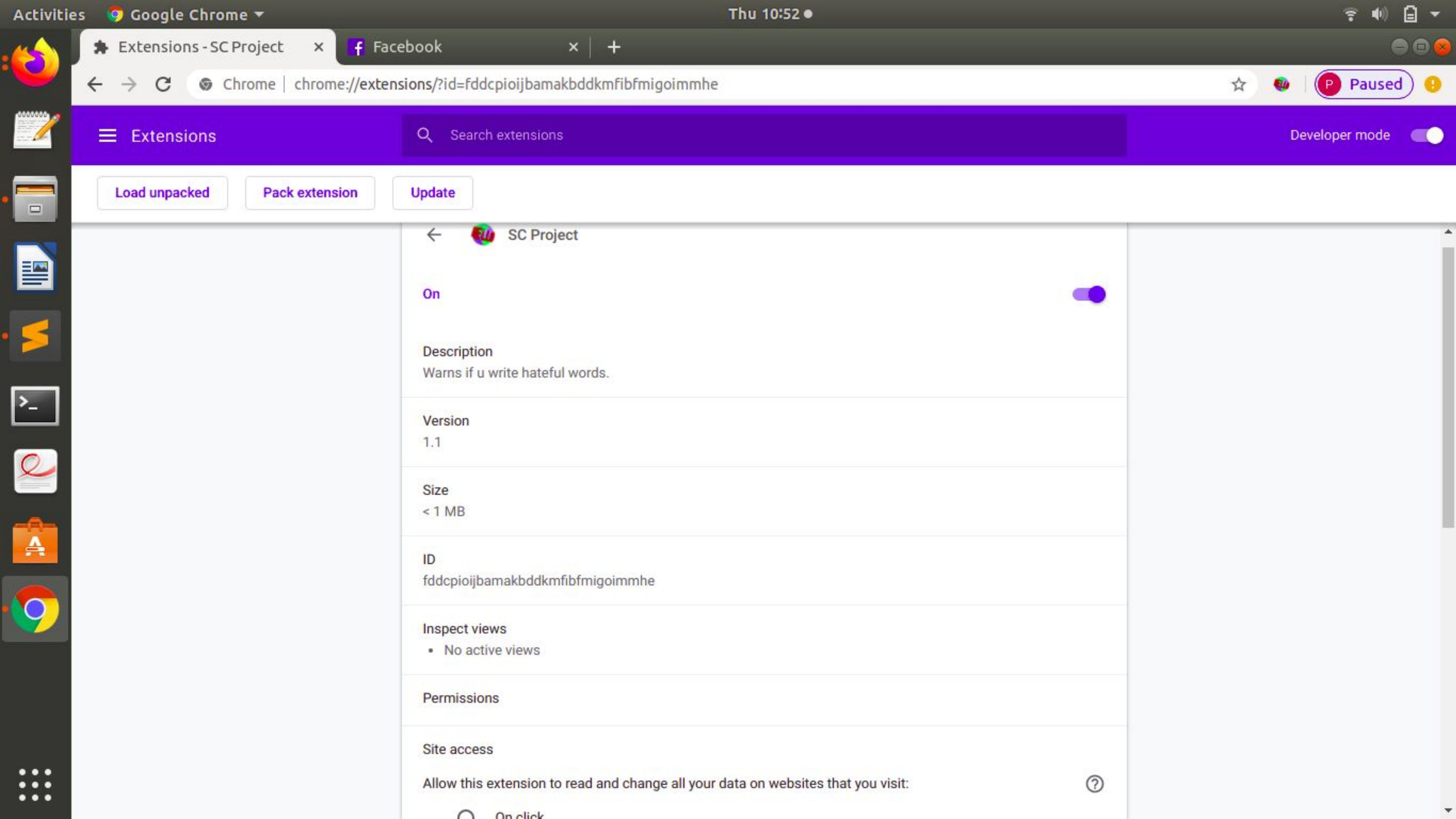
SC Project 1.1
Warns if u write hateful words.
ID: fddcpioijbamakbddkmfibfmigoimmhe

Allow-Control-Allow-Origin: * 1.0.3
Allows to you request any site with ajax from any source. Adds to response 'Allow-Control-Allow-Origin: *' header
ID: nlfbmbojpeacfghkpbjhddihlkkiljbi

CORS 0.1.2
modify/add Access-Control-Allow-Origin Header
ID: dboaklophljenpcjkbbibpkbpbobnbld

Google Docs Offline 1.9.1
Edit, create and view your documents, spreadsheets and presentations – all without Internet access.
ID: ghbmnnjooekpmoecnnnilnnbdlolhkhi

Postman Interceptor 1.0.1
Sends requests fired through the Postman chrome app.



SC Project

On

Description

Warns if u write hateful words.

Version

1.1

Size

< 1 MB

ID

fddcpioijbamakbddkmfibfmigoimmhe

Inspect views

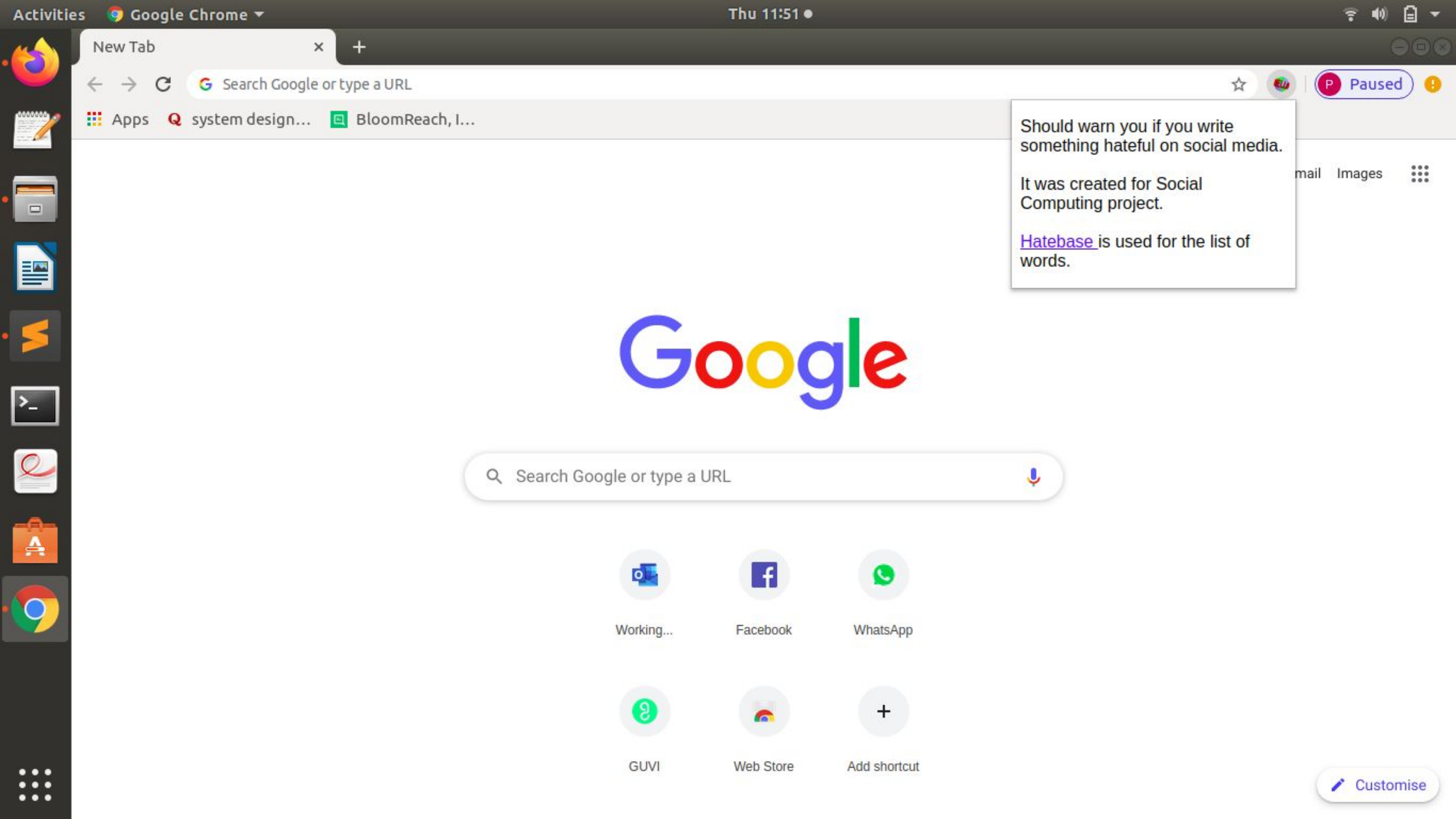
- No active views

Permissions

Site access

Allow this extension to read and change all your data on websites that you visit:

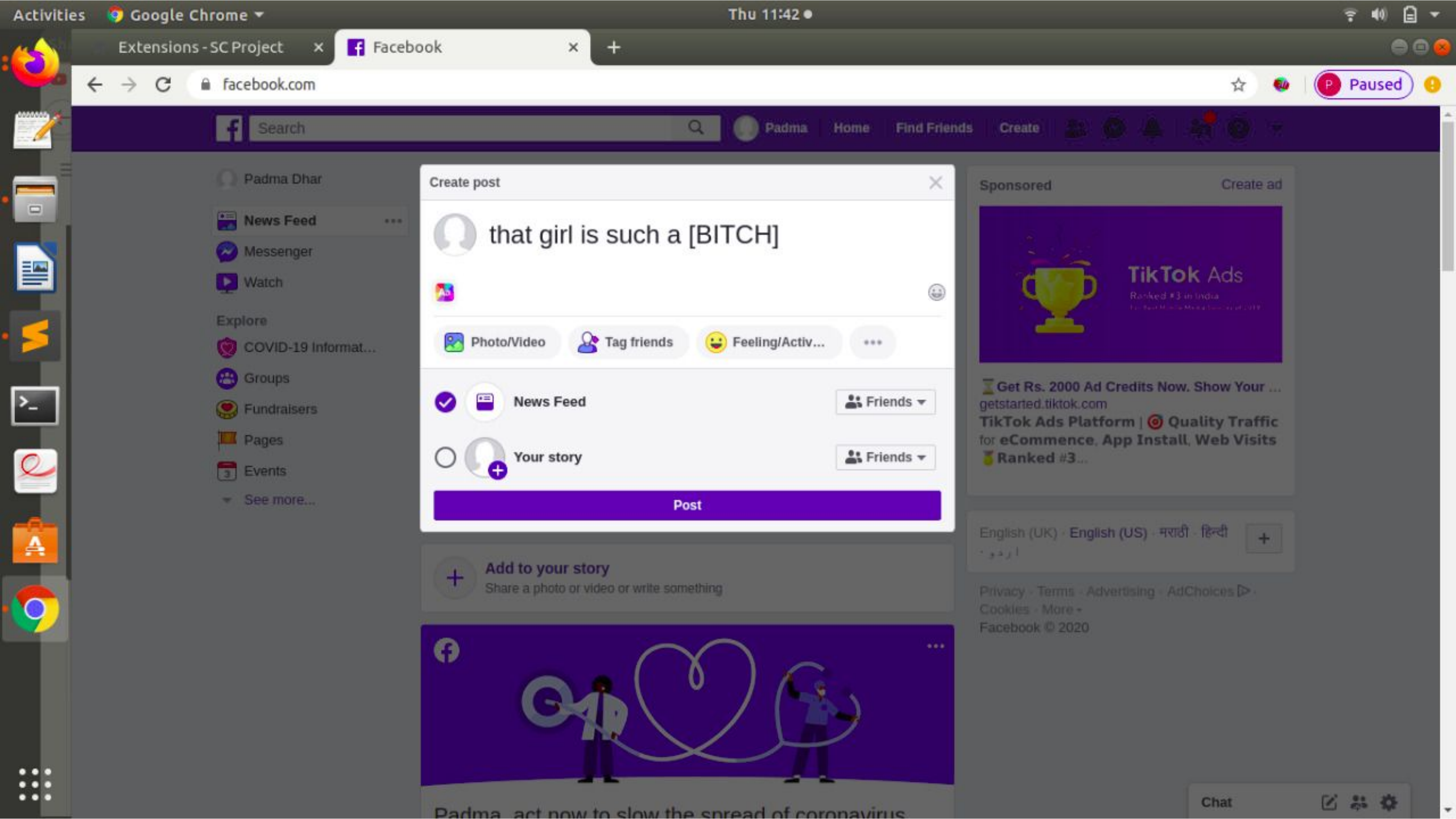
○ On click



Should warn you if you write something hateful on social media.

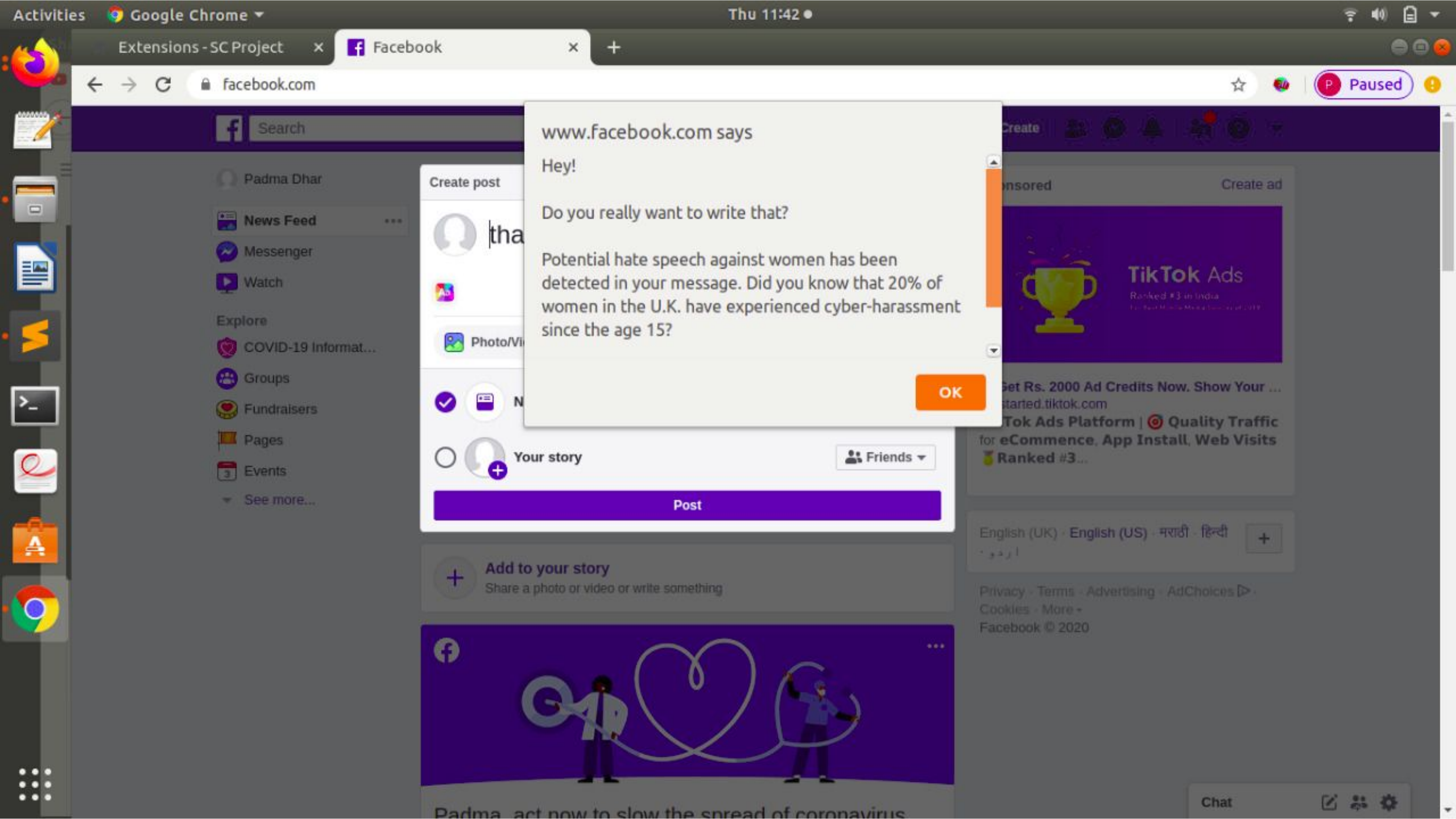
It was created for Social Computing project.

Hatebase is used for the list of words.



Create post

that girl is such a [BITCH]



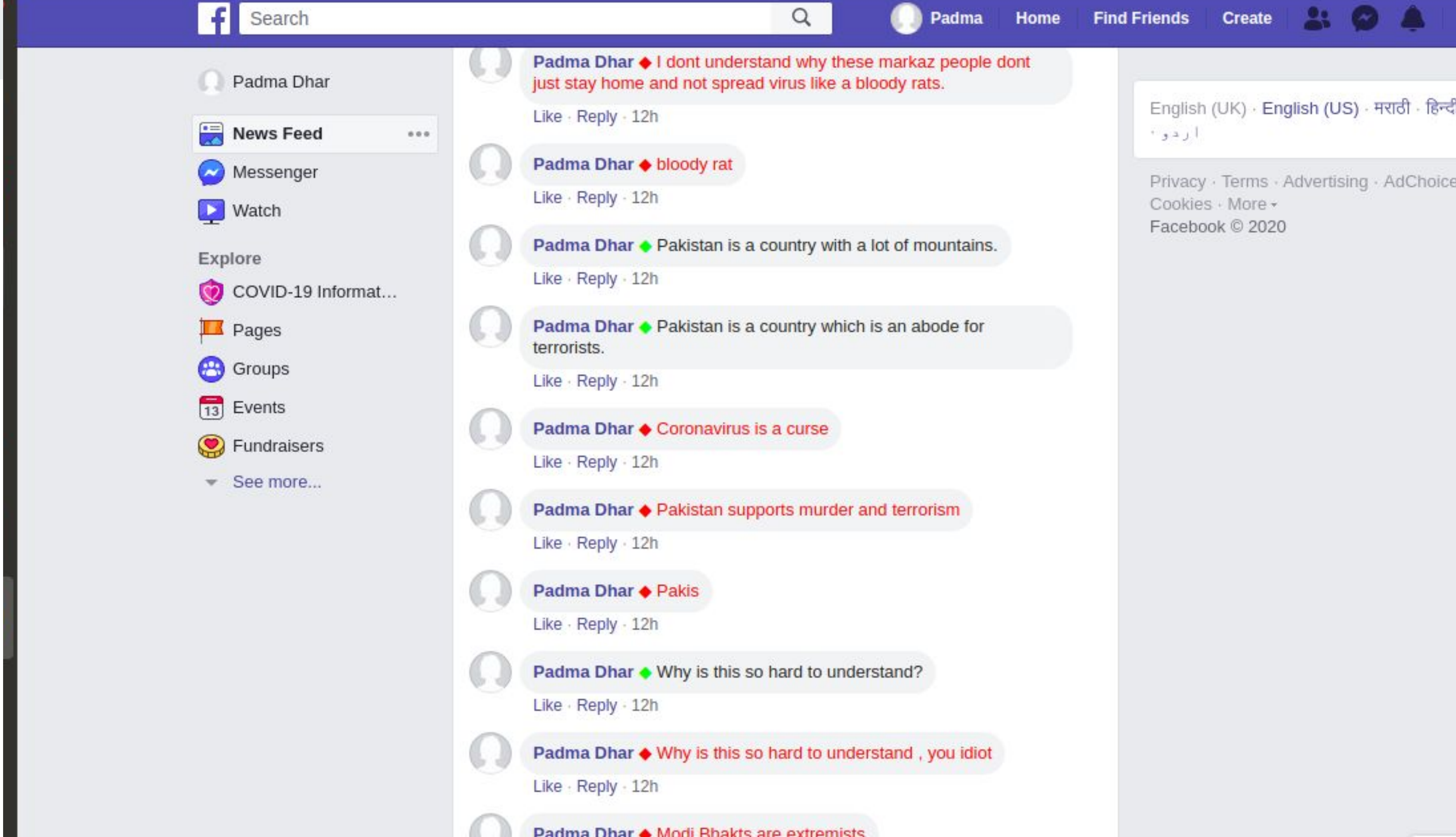
www.facebook.com says

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



Padma Dhar 🟥 I dont understand why these markaz people dont just stay home and not spread virus like a bloody rats.
Like · Reply · 12h

Padma Dhar 🟥 bloody rat
Like · Reply · 12h

Padma Dhar 🟩 Pakistan is a country with a lot of mountains.
Like · Reply · 12h

Padma Dhar 🟩 Pakistan is a country which is an abode for terrorists.
Like · Reply · 12h

Padma Dhar 🟥 Coronavirus is a curse
Like · Reply · 12h

Padma Dhar 🟥 Pakistan supports murder and terrorism
Like · Reply · 12h

Padma Dhar 🟥 Pakis
Like · Reply · 12h

Padma Dhar 🟩 Why is this so hard to understand?
Like · Reply · 12h

Padma Dhar 🟥 Why is this so hard to understand , you idiot
Like · Reply · 12h

Padma Dhar 🟥 Modi Bhakts are extremists

English (UK) · English (US) · मराठी · हिन्दी
اردو ·
Privacy · Terms · Advertising · AdChoice
Cookies · More ▾
Facebook © 2020

Major 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 is shown as a 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.)

Thank You