**Identifying hate speech in social media platforms**

**Data**: Public dataset

Datasets(GV):

<https://huggingface.co/datasets/hate_speech18>

<https://drive.google.com/drive/folders/1hKLOtpVmF45IoBmJPwojgq6XraLtHmV6>

<https://github.com/dadangewp/SWAD-Repository>

<https://github.com/ipavlopoulos/toxic_spans>

<https://github.com/Vicomtech/hate-speech-dataset>

**Objective :**

1. Identify Hate speech from social media (Binary Classification, Clustering)
2. Sentiment analysis on Hate speech
3. Analysis of Hate speech across demographics(gender, race, religion)(provided we have it in the public dataset)

**Next steps:**

Look for dataset (GV,JK)

Methodology discussion (by Feb 10th) (algo, who should do what)

**Introduction**

Hate speech is defined any form of communication that’s aims at diminishing an individual or a group of people based on their physical characters. Now days, hate speech is one of the common form of cyberbullying being experienced by several social media users. Hate speech has the ability to cause severe negative psychological effect on the target individuals. Hate speech can make one feel unwanted or less important in life. Most victims of hate speech end up being depressed and losses their self-confidence.

Several studies aiming at exploring ways on identifying and combating hating speech have been done. Most of these studies usually work by comparing the obtained words to existing words that are considered as hate speech. The problem with these techniques is that new, informal and non – English hate speech words might be left out.

**What we need to do?**

In this project we are aiming at improving identification of hate speech from social media by analyzing both the words and the reaction types given to the words. Firstly, using document similarity techniques, will compare the collected words with existing dataset of hate speech words. If a word is found we will then classify it as a hate speech. The newly introduced technique in this project is the analyzing of the reactions given to the words. We will analyze the comments associated with the given words and compare the comment with words of complain or sorry feeling. We will determine a percentage at which a word will be consider hated based on comparison of comments to words of complain.

We will then compare the results obtained for each method. The last part will involving applying both techniques to collected words and compare the results to the one obtained in individual application of the techniques. In this project, we will focus on hate speech on gender/religion/racial/

**Literature**

1. [**https://ieeexplore.ieee.org/abstract/document/8355039**](https://ieeexplore.ieee.org/abstract/document/8355039)
2. [**https://ieeexplore.ieee.org/document/8920923**](https://ieeexplore.ieee.org/document/8920923)
3. [**https://ieeexplore.ieee.org/document/9544554**](https://ieeexplore.ieee.org/document/9544554)