Abstract –

Social media is a collection of internet-based platforms that allow users to create and interact with content in real-time. It enables individuals, groups, and organizations to connect, communicate, and share information. However, the relationship between social media and free speech is complex and can lead to the spread of hate speech and harmful content.

As the internet and social media continue to evolve, the problem of hate speech and offensive language on these platforms also evolves. Existing research highlights that hate is prevalent across various platforms, but there is a lack of models for detecting online hate using data from multiple platforms.

To gain a deeper understanding, we have compared different social media platforms by analyzing their datasets. Additionally, we conducted a rudimentary analysis to better comprehend these platforms' dynamics and the prevalence of hate speech. This research aims to contribute to a comprehensive understanding of the challenges and implications of hate speech in the ever-changing landscape of social media.

Introduction -

According to the United Nations (2019), hate speech is defined as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor." Despite this, multiple definitions of hate speech can exist. The term "Offensive language" is also used concurrently with hate speech. An observation that can be made from the way classifiers operate is that oftentimes hate speech is misclassified as offensive language and vice versa , from ( [Davidson et al., 2017](https://doi.org/10.48550/arXiv.1703.04009); [Mozafari et al., 2019](https://doi.org/10.48550/arXiv.1910.12574) ).

Hate Speech is defined as language used to express hatred towards a targeted group/individual based on protected attributes such as race or religion, offensive language contains offensive terms but is not targeting any group in particular, from ( [Yuan et al., WOAH 2022](http://dx.doi.org/10.18653/v1/2022.woah-1.1) ).

For the purpose of this research, we have used binary classification (hate, offensive / neither). This implies that offensive comments will be labelled as hate as well.

Lexicon –

* The researchers gathered five Spanish words associated with hateful speech towards women from the website hatebase.org. To expand their dataset, they used a technique called word embedding to find similar words in a large collection of Spanish texts. They excluded some words from the initial list as they were mainly related to animals. Afterwards, they used an online application to find synonymous words that were not already included. This helped them create three lexicons: one for misogyny, another for insults, and a third one for xenophobia. The lexicons were later translated into English. ( <https://dl.acm.org/doi/abs/10.1145/3369869> )

( <https://aclanthology.org/2022.woah-1.10/> )

* HURTLEX is a lexicon that includes offensive, aggressive, and hateful words from 53 languages. It contains approximately 1000 words manually chosen to cover 17 specific categories. To expand the lexicon, additional words were added using MultiWordNet synsets and Babelnet in a semi-automatic process. The lexicon was then refined and updated to create the modern version of HURTLEX.
* All the lexicons mentioned earlier were combined into a single, comprehensive lexicon. To ensure its accuracy and avoid repetition, the lexicon was filtered, and duplicate words were removed. This process resulted in a larger and more refined lexicon.

Datasets –

Twitter - Researchers used Hatebase.org's hate speech lexicon, identified tweets containing its terms via the Twitter API, and collected a sample of 85.4 million tweets from 33,458 users. They manually coded 25,000 sample tweets into hate speech, offensive (not hate speech), or neither. The majority decision from CrowdFlower workers (92% agreement) was used to label each tweet, resulting in 24,802 labeled tweets. Only 5% of tweets were coded as hate speech. (<https://ojs.aaai.org/index.php/ICWSM/article/view/14955> )