**Hate Speech Detection**

**Abstract and Scope**

Hate speech, defined as an "abusive speech targeting specific group characteristics, regarding ethnicity, religion, gender or social issue", is a critical problem plaguing websites that makes social media a thriving platform for negativity and misinformation. Consequently, the community is affected as a whole, deteriorating the society on a behavioral and economical front.

In order to address this issue, in this project, we aim to develop a model that will be able to classify a given text as hate speech or not using state-of-the-art transformer models through deep learning techniques.

**Feasibility Study**:

In this day and age cyberbullying is a serious issue among youth of the nation. Around 85% of kids on social media are subjected to some form of cyberbullying and online abuse. Hate speech refers to the use of aggressive, violent or offensive language targeting different groups of people. By detecting hate speech we can avoid offensive behavior.

**Design Approach/ Methodology/ Planning of work**

**Data Collection:** Tweets is initially taken as input to the model.

**Data Pre-processing:** The initially collected data is cleaned to remove NULL, redundant and inconsistent data. Further punctuation cleaning is performed using regular expression. Stop words are removed in order to get a concise representation of data appropriate for further analysis and modelling.

**Word Embeddings generation:** Word embeddings play a crucial role in correctly classifying hate speech tweets as they convey contextualised meaning of the words to the model. We have used the following embeddings :

* FastText
* GloVe
* RoBerta

These word embeddings have been passed to a neural networks with three Bi-directional LSTM layers or to BERT architecture.

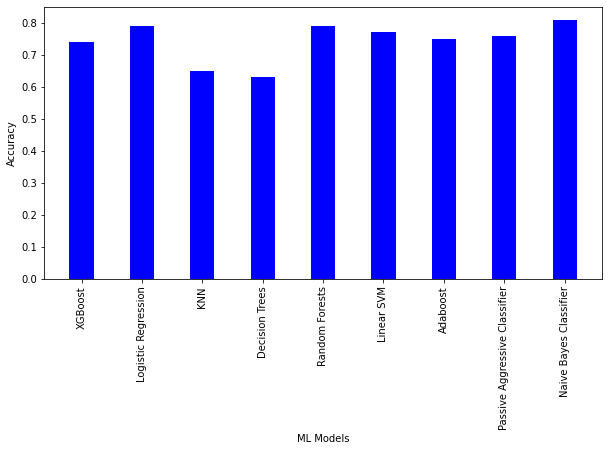
**Modelling:** Training the model on the train dataset

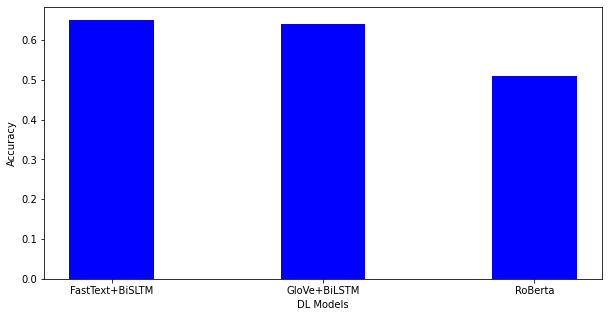
**Hyperparameter Fine Tuning:** Validating and improving accuracy of the model.

**Models used**

* Naive Bayes Classifier
* Logistic Regression
* Random Forests
* Passive Aggressive Classifier
* Adaboost
* Linear SVM
* K- Nearest Neighbours
* XGBoost
* Decision trees

**Results**:





### References

* <https://arxiv.org/pdf/2010.12472v2.pdf>
* <https://openaccess.thecvf.com/content_WACV_2020/papers/Gomez_Exploring_Hate_Speech_Detection_in_Multimodal_Publications_WACV_2020_paper.pdf>
* <https://peerj.com/articles/cs-598/>
* <https://www.researchgate.net/profile/Fabio-Del-Vigna/publication/316971988_Hate_me_hate_me_not_Hate_speech_detection_on_Facebook/links/591af15d0f7e9beed7f5ff61/Hate-me-hate-me-not-Hate-speech-detection-on-Facebook.pdf>
* <https://ieeexplore.ieee.org/abstract/document/4722931>
* <https://arxiv.org/abs/1712.07525>
* <https://ieeexplore.ieee.org/abstract/document/9658978>