**HATE SPEECH**

**DETECTION**

**Machine Intelligence**

**Abstract**

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.

**Literature Survey**

The main aim of the paper, Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection, is to classify the tweets in three categories :

a) Clean: Class of speech which is neutral and has no hate b) Offensive: Class of speech which is offensive but not necessarily have hate or any racial words

c) Hateful: this class includes tweets which are offensive,

and present hate, racist and segregative words and expressions.

Three datasets were collected and combined : Two datasets from crowdflower, and one from github. All three of them are manually annotated. The training set had 21000 tweets, while the testing and validation set had 2,010 tweets each. Sentiment polarity is one of the main features used to indicate hate in this paper. The classification is performed using the machine learning algorithm ‘‘J48graft”. This classifier presents better performances than other classifiers (even powerful ones such as Support Vector Machine (SVM) and Random Forest, etc.). Binary classification (hateful and offensive) and ternary classification(clean, offensive and hurtful) are done and results are obtained.TP Rate, FP rate, Precision, Recall and F1 score is calculated for all the features along with the confusion matrix for both the classifications.

Main aim of the paper, A Lexicon-based Approach for Hate Speech Detection, is to generate a lexicon of sentiment expressions using semantic and subjectivity

features with an orientation to hate speech and then use these features to create a classifier for hate

speech detection. The first stage involves subjectivity detection, and is intended at isolating sentences that have subjective expressions from those that generally express objective sentiments. In the second stage a

lexicon of hate related words is built using a rule-based method using subjective features identified from the sentences and semantic features learned directly from the corpus. Then in the final stage a classifier that

utilizes features created from the lexicon and is used to test for hate speech in a document.

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

**Data Collection:** Tweets are 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 modeling.

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

● FastText

● GloVe

● RoBerta

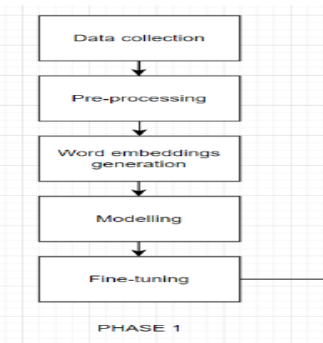
These word embeddings have been passed to a neural network with three Bi-directional LSTM layers or to BERT architecture.

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

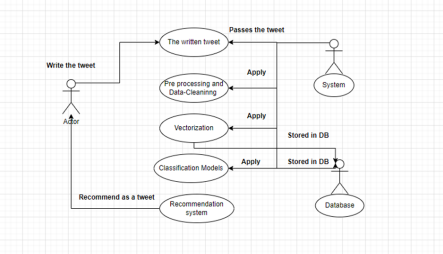
**Hyperparameter Fine Tuning:** Validating and improving accuracy of the model.

**Methods**

Training data has been extracted from Huggingface which is a labeled dataset indicating whether a given tweet represents hate speech or not. The dataset has been cleaned and pre-processed to remove redundant and inconsistent data. It has been split into training and testing sets. The data is mapped to word embeddings using linguistic models. Deep learning models have been trained on this data and the results are validated. Testing set is used to obtain test results through classification metrics and confusion matrix.



***Figure:*** *Control flow diagram*

*****Figure:*** *Use case diagram*

**Machine Learning Models**

● Naive Bayes Classifier

● Logistic Regression

● Random Forests

● Passive Aggressive Classifier

● Adaboost

● Linear SVM

● K- Nearest Neighbors

● XGBoost

● Decision trees

**Deep Learning Models**

● Bi-directional LSTM

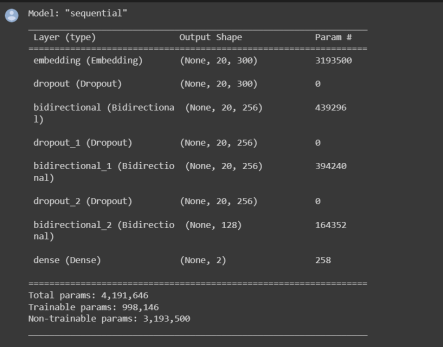
● BERT

**Word embeddings for deep learning**

● FastText

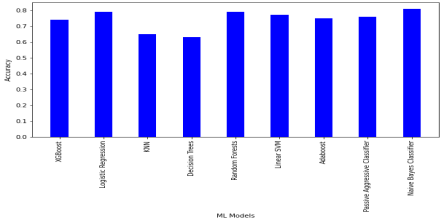
● GloVe

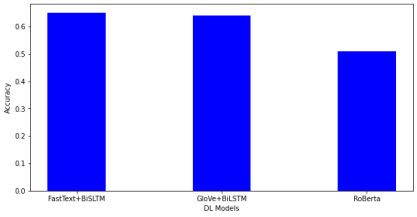
● RoBerta

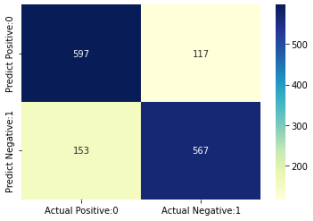
***Figure:*** *Architecture of the network for Bi-LSTM models*

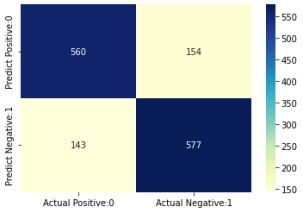
**RESULT**

The metrics used for evaluating the models included confusion matrix and classification report. Naive Bayes Classifier outperformed other models with an accuracy of 80% follows by Logistic Regression, Random Forests and Naive Bayes Classifier.

***Figure:*** *Accuracy metrics of Machine Learning models*

*****Figure:*** *Accuracy metrics of Deep Learning models*

*****Figure:*** *Confusion matrix for Naive Bayes Classifier*

*****Figure:*** *Confusion matrix for Logistic Regression*

**References**

● https://arxiv.org/pdf/2010.12472v2.pdf

● https://openaccess.thecvf.com/content\_WACV\_2020/pap ers/Gomez\_Exploring\_Hate\_Speech\_Detection\_in\_Multi modal\_Publications\_WACV\_2020\_paper.pdf

● https://peerj.com/articles/cs-598/

● https://www.researchgate.net/profile/Fabio-Del-Vigna/pub lication/316971988\_Hate\_me\_hate\_me\_not\_Hate\_speec h\_detection\_on\_Facebook/links/591af15d0f7e9beed7f5ff 61/Hate-me-hate-me-not-Hate-speech-detection-on-Face book.pdf

● https://ieeexplore.ieee.org/abstract/document/4722931 ● https://arxiv.org/abs/1712.07525

● https://ieeexplore.ieee.org/abstract/document/9658978