Comparison of Classification Algorithms for Hate Comment Detection

PIYUSH RAJPUT

2019IMT-074

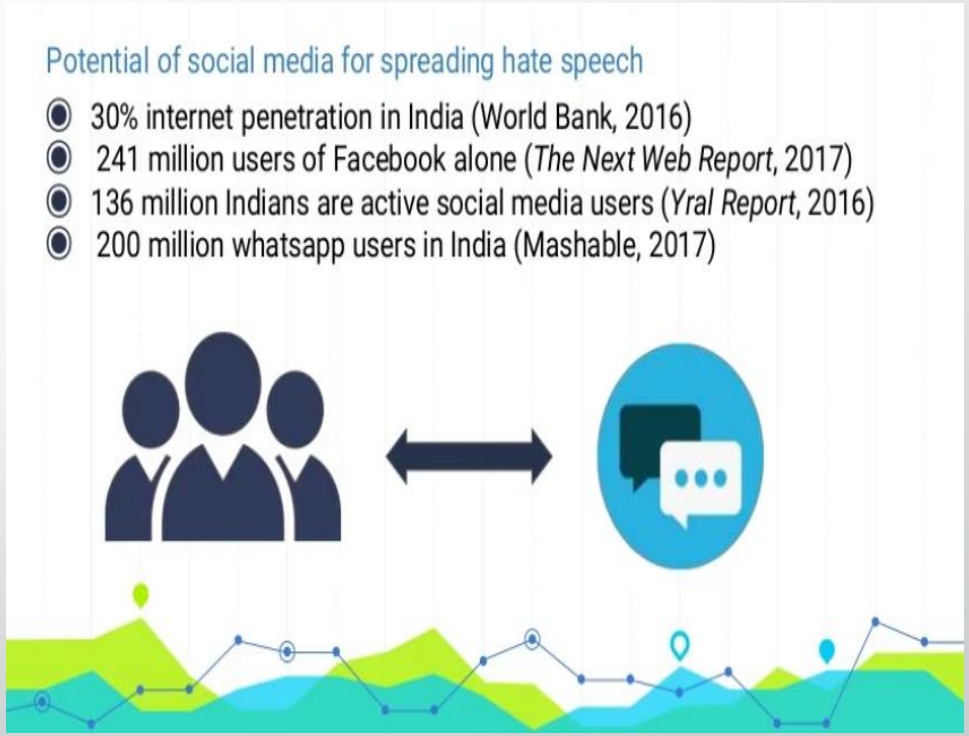
MENTOR: Dr. Anuraj Singh

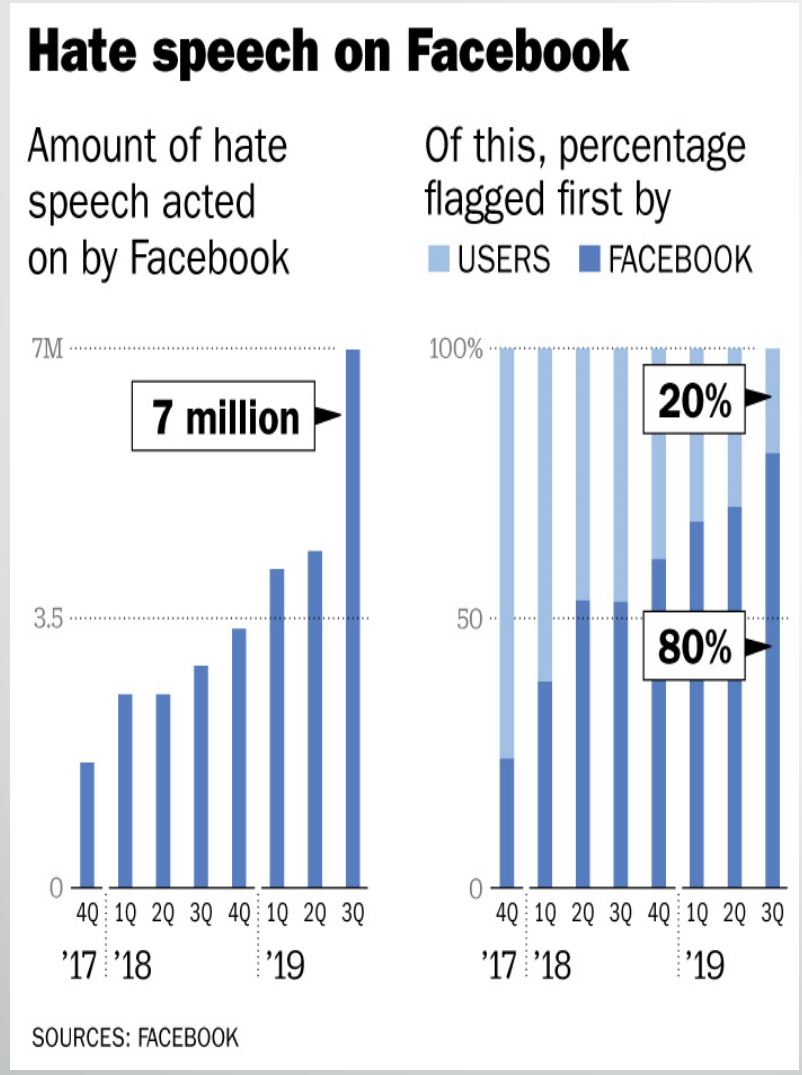
INTRODUCTION

- Hate speech is speech that attacks a person or group on the basis of attributes such as race, religion, ethnic origin, national origin, gender, disability, sexual orientation.
- The development in the field of technologies and the boom of the Internet has driven the popularity of social networking sites to the limit that the social media giant Face-book alone has an active monthly user of 2.9 billion people.
- Although expressing personal views have became easy, hateful and offensive texts or comments have increased significantly on social media.
- Natural language processing techniques can assist with monitoring online hate speech

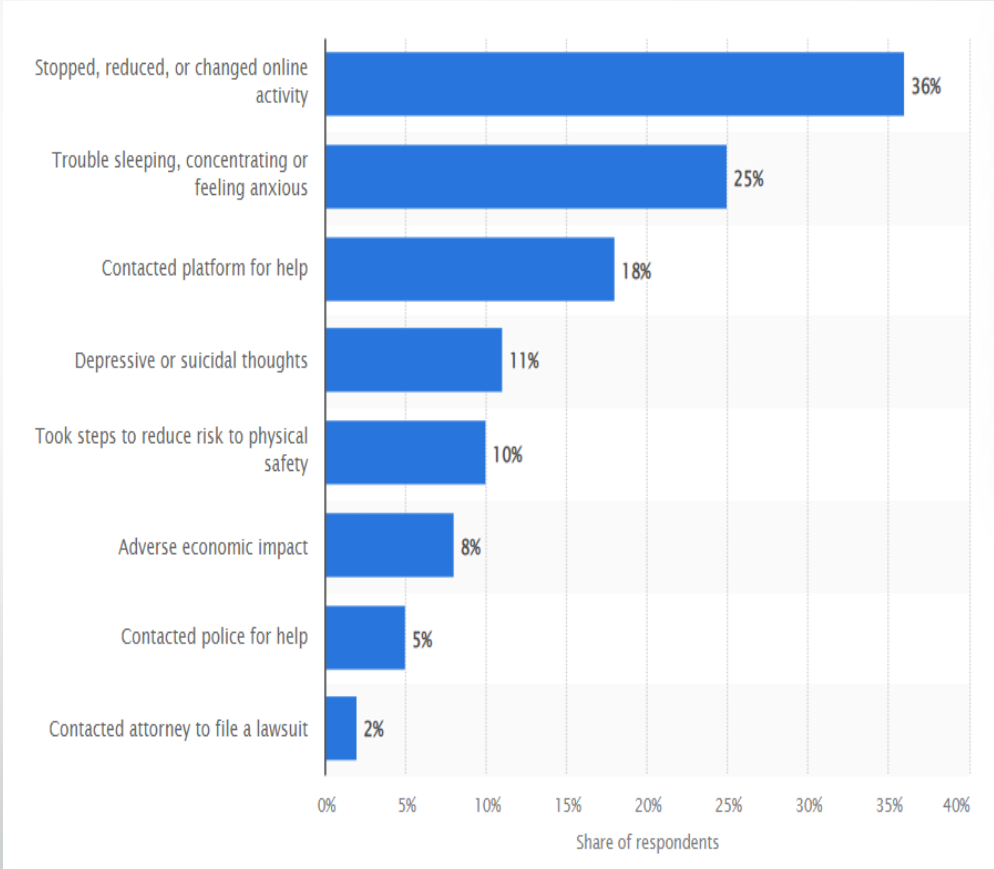
MOTIVATION

- Twitter “actioned” 1,126,990 different accounts between July and December 2020 for infringing its hateful conduct policy, a 77% increase over the prior six-month period.
- Abusive online content can cause users of social media platforms emotional and psychological trauma, which has caused some of them to delete their accounts and, in the worst circumstances, commit suicide.





Consequences of online hate and harassment according to internet users in the United States as of January 2020



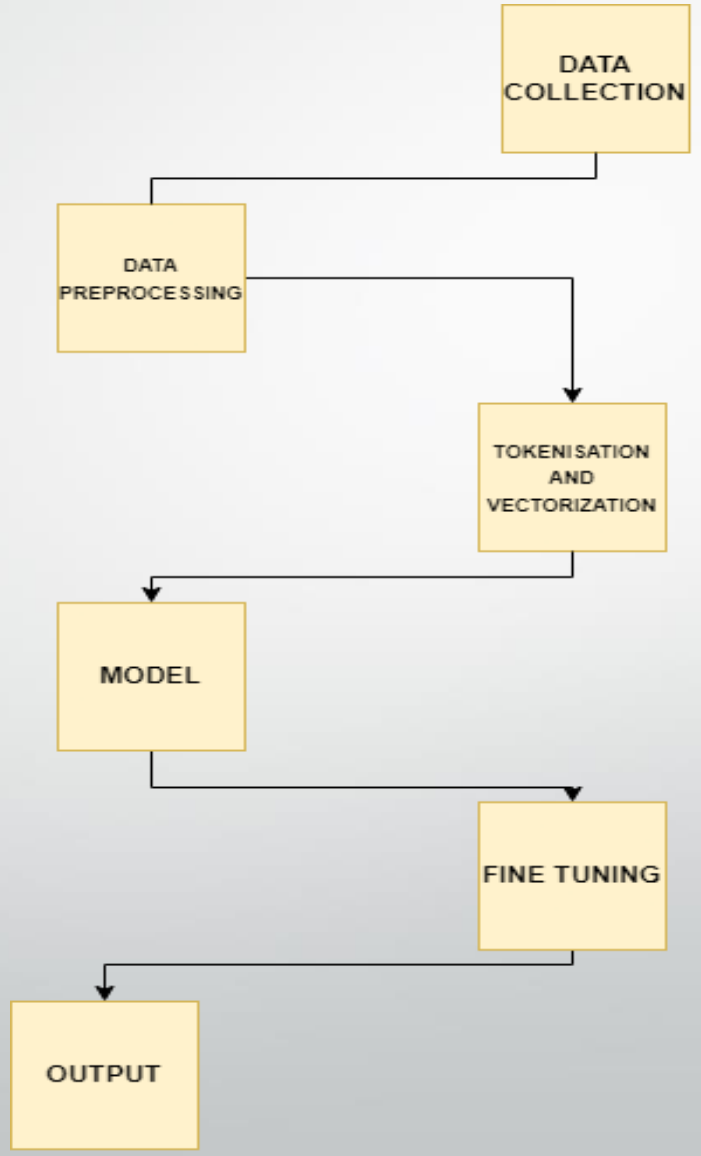
LITERATURE REVIEW

Banik and Rahman	2019	Detection of toxic text in Facebook bangla comments using SVM
Hussain et al.	2018	It presented a root level algorithm to identify offensive comments. The research is limited by the absence of comparisons with conventional classifiers and the small dataset.
Eshan and Hasan	2017	SVM with linear kernel is utilized along with TF-IDF vectorizer for detection of abusive text.
Rahul et al.	2020	Logistic regression is utilized for the detection of offensive text in a sentence.
Anand and Eswari	2019	CNN glove and LSTM is used to perform the task and the result was good.

OBJECTIVES

- To train the machine learning models in order to predict whether a comment or speech contains any sort of hate or offensive language.
- To evaluate the performance of different models like Logistic Regression, Naive Bayes, Support Vector Machines, Random Forest Classifier etc.
- To compare the results obtained by the models.

METHODOLOGY



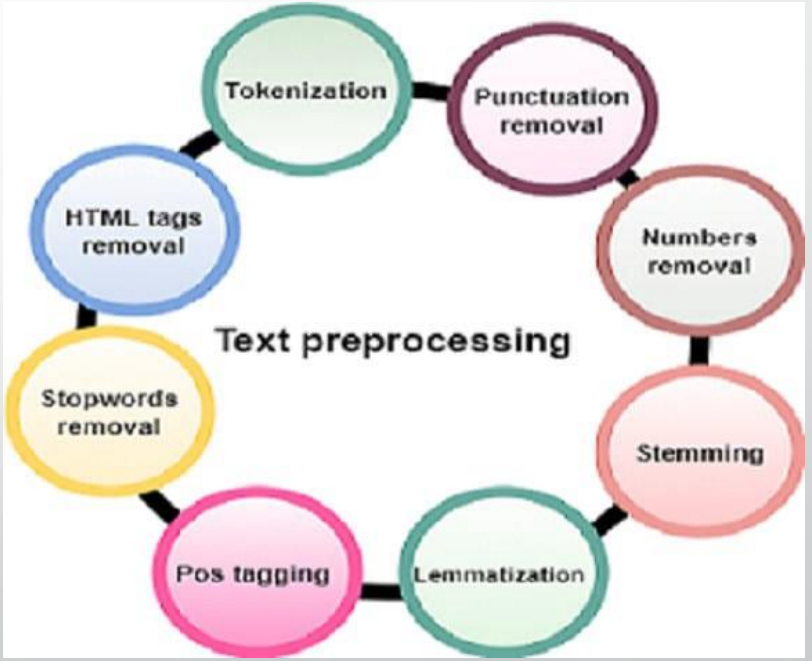
DATA COLLECTION

- The dataset we are using was created using Twitter data to research hate-speech detection.
- The text is classified as hate speech, offensive language, and neither.
- The dataset size is 24783 tweets.

#	#count	#hate_speech	#offensive_language	#neither	#class	tweet
Index	Number of CrowdFlower users who coded each tweet (min is 3, sometimes more users coded a tweet)	Number of CF users who judged the tweet to be hate speech	Number of CF users who judged the tweet to be offensive	Number of CF users who judged the tweet to be neither offensive nor non-offensive	Class label for majority of CF users. 0 - hate speech 1 - offensive language 2 - neither	Text tweet

DATA PREPROCESSING

- Removing punctuations like . , ! $( ) * % @
- Removing URLs
- Removing Stop words
- Lower casing
- Stemming
- Lemmatization



VECTORIZATION

- **BOW**: A bag-of-words model, or BoW for short, is a way of extracting features from text for use in modeling. It involves two things, a vocabulary of known words and a measure of the presence of known words.
- **TF-IDF**: TF-IDF stands for Term Frequency - Inverse Document Frequency. Frequency of a particular term relative to that document is known as term frequency. IDF looks at how common (or uncommon) a word is against the corpus.

MODELS

- Naïve Bayes
- Logistic Regression
- Random Forest Classifier
- Linear SVM
- BERT

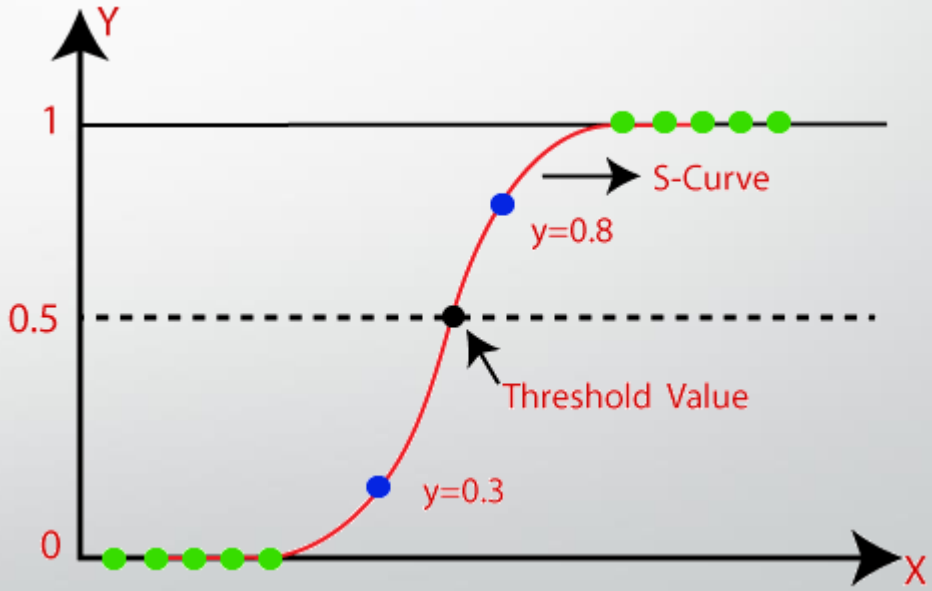
NAÏVE BAYES

- It is a probabilistic learning model and is widely used in NLP problems.
- It is based on Bayes Theorem.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

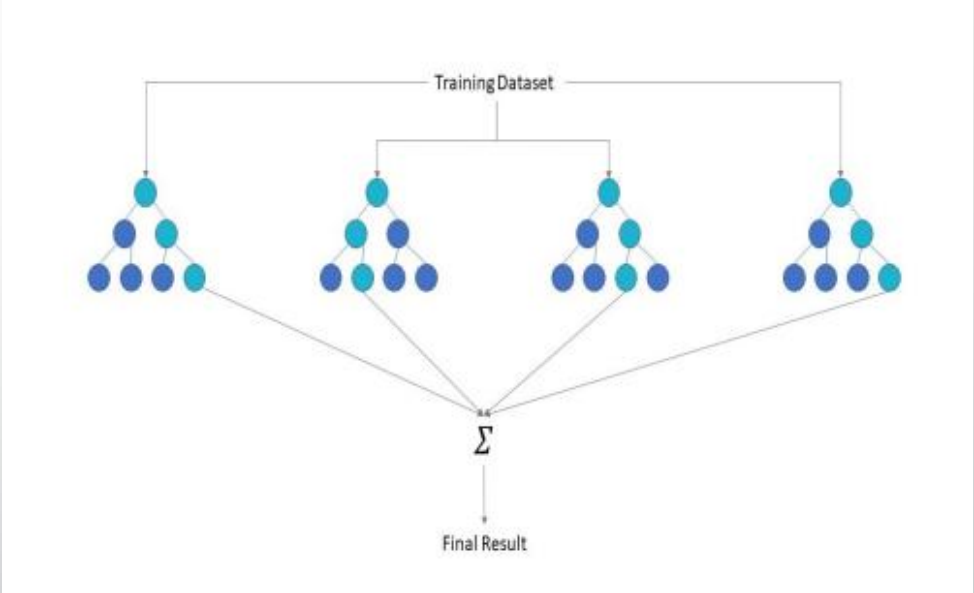
LOGISTIC REGRESSION

- It is a supervised learning algorithm based on regression algorithm used for classification problems.
- Logistic regression, in contrast to linear regression, changes its output using the logistic sigmoid function to deliver a probability value that may then be mapped to two or more discrete classes.



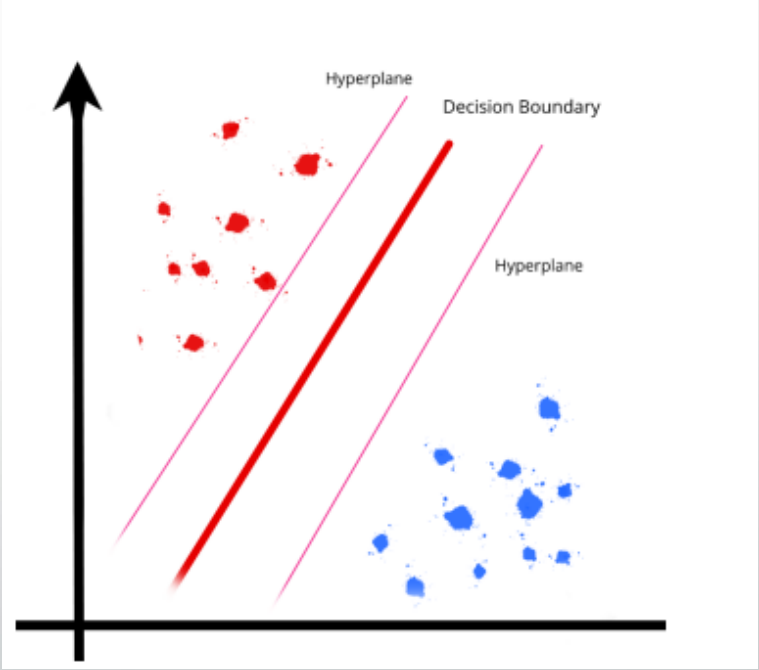
RANDOM FOREST CLASSIFIER

- It is a supervised learning method
- Similar to how a forest has many trees, it is made up of various decision trees.
- Based on a random selection of data samples, these algorithms create decision trees and obtain predictions from each tree. They then vote to determine which viable option is the best



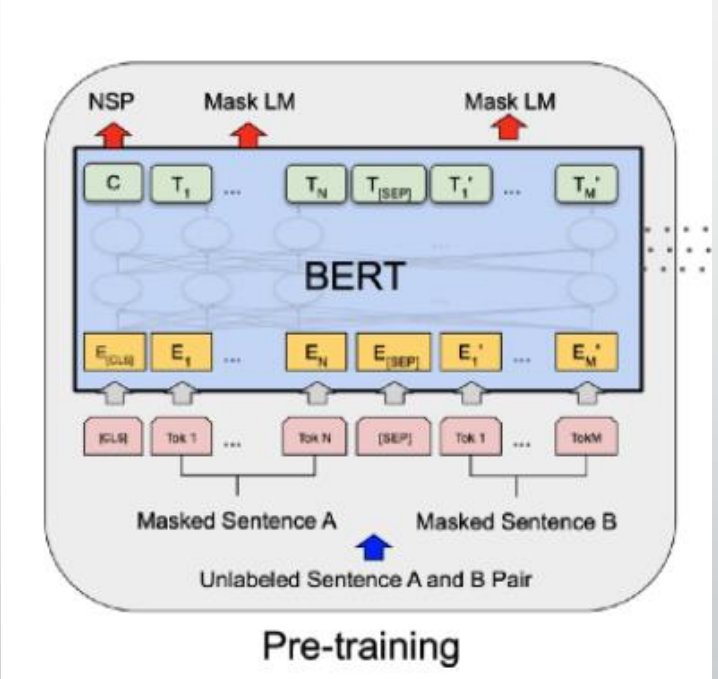
LINEAR SVM

- SVM stands for Support Vector Machine. It is a simple supervised learning method which is used for regression and classification.
- In order to efficiently classify new data points in the future, the SVM algorithm aims to determine the optimum line or decision boundary that can split n-dimensional space into classes. The name of this optimal decision boundary is a hyperplane.
- SVM selects the extreme points that aid in the creation of the hyperplane.



BERT

- Bidirectional Encoder Representations from Transformers
- It is a machine learning framework for handling natural language that is free and open-source
- BERT is truly bidirectional.



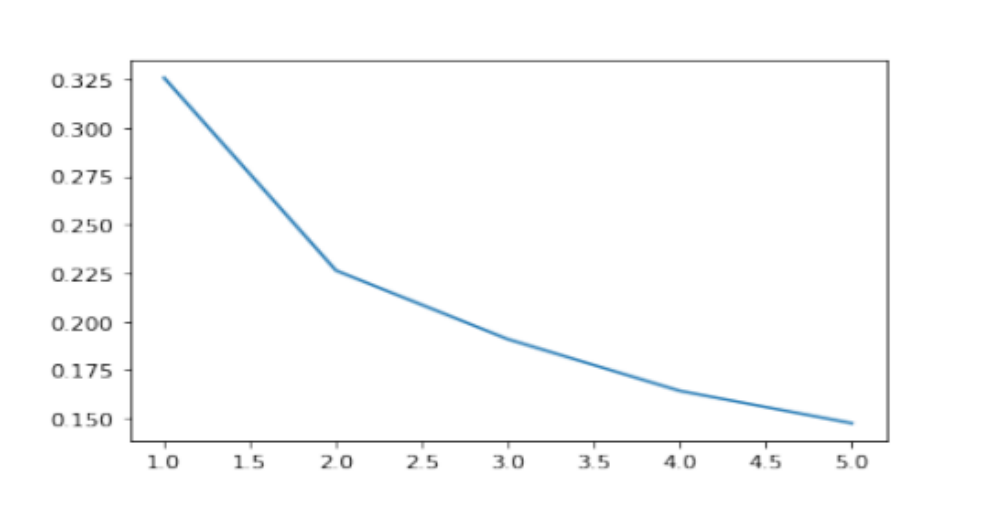
Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1 Score
- Confusion Matrix

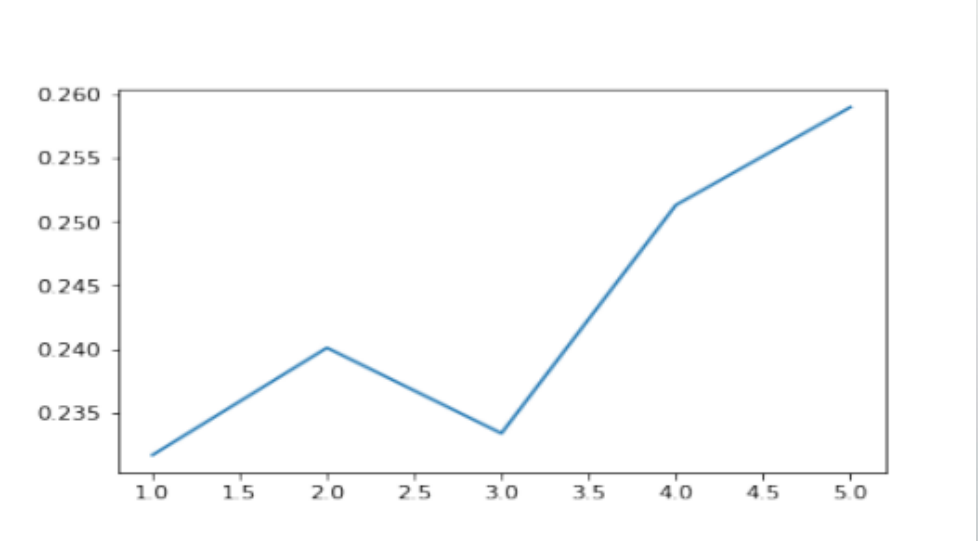
RESULT

Models / Metrics	Accuracy	F1-score	Precision	Recall
Naive Bayes	0.64	0.70	0.79	0.65
Logistic Regression	0.89	0.88	0.88	0.90
Random Forest	0.90	0.89	0.89	0.91
LinearSVM	0.89	0.89	0.88	0.89
BERT	0.92	0.92	0.91	0.92

RESULT CONTINUED



Training loss versus epochs



Validation loss versus epochs

FUTURE WORK

- Data Augmentation on our data in order to counter the imbalances present in the dataset.
- Heading into the vast area of Deep Learning in order to build a deep learning model best suited for our project.

REFERENCES

[1] Aizawa, A.: 2003, An information-theoretic perspective of tf–idf measures, Information Processing & Management 39(1), 45–65.

[2] Anand, M. and Eswari, R.: 2019, Classification of abusive comments in social media using deep learning, 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 974–977

[3] Banik, N. and Rahman, M. H. H.: 2019, Toxicity detection on bengali social media comments using supervised models, 2019 2nd International Conference on Innovation in Engineering and Technology (ICIET), pp. 1–5.

[4] Eshan, S. C. and Hasan, M. S.: 2017, An application of machine learning to detect abusive bengali text, 2017 20th International Conference of Computer and Information Technology (ICCIT), pp. 1–6.

[5] Grandini, M., Bagli, E. and Visani, G.: 2020, Metrics for multi-class classification: an overview

REFERENCES CONTINUED

[6] Hussain, M. G., Mahmud, T. A. and Akthar, W.: 2018, An approach to detect abusive bangla text, 2018 International Conference on Innovation in Engineering and Technology (ICIET), pp. 1–5

[7] Kannan, S., Gurusamy, V., Vijayarani, S., Ilamathi, J., Nithya, M., Kannan, S. and Gurusamy, V.: 2014, Preprocessing techniques for text mining, International Journal of Computer Science & Communication Networks 5(1), 7–16.

[8] Koroteev, M.: 2021, Bert: A review of applications in natural language processing and understanding, arXiv preprint arXiv:2103.11943 .

[9] Rahul, Kajla, H., Hooda, J. and Saini, G.: 2020, Classification of online toxic comments using machine learning algorithms, 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1119–1123.

[10] Saif, M. A., Medvedev, A. N., Medvedev, M. A. and Atanasova, T.: 2018, Classification of online toxic comments using the logistic regression and neural networks models, AIP conference proceedings, Vol. 2048, AIP Publishing LLC, p. 060011.

THANKYOU