A Machine Learning Approach to Classify Anti-social Bengali Comments on Social Media

Manash Sarker
Patuakhali Sc. and Tech. University
Patuakhali, Bangladesh
manash.sarker@pstu.ac.bd

Md. Forhad Hossain
Patuakhali Sc. and Tech. University
Patuakhali, Bangladesh
forhad14@cse.pstu.ac.bd

Fahmida Rahman Liza
Patuakhali Sc. and Tech. University
Patuakhali, Bangladesh
liza14@cse.pstu.ac.bd

Syed Nazmus Sakib
Patuakhali Sc. and Tech. University
Patuakhali, Bangladesh
sakib14@cse.pstu.ac.bd

Abdullah Al Farooq
Wentworth Institute of Technology
Boston, USA
farooqa@wit.edu

Abstract—**The growth of social media is causing the emergence of hate speech. Email extortion and cyberbullying are on the rise in Bangladesh, along with online sexual harassment of women. In order to prevent these crimes, studies on Bengali comments on social media have become progressively important. However, the requisite datasets are scarce for this kind of study. The motive of this research is to create a dataset of Bangla comments from social platforms and develop a classifier model as well as to detect whether the comments are social or anti-social quickly and efficiently. 2000 comments were gathered from Facebook and YouTube, two prominent platforms for social media. In our study, an artificial neural network model like Gated Recurrent Unit (GRU), and supervised machine learning classifiers like Logistic Regression (LR), Random Forest (RF), Multinomial Naive Bayes (MNB), and Support Vector Machine (SVM) were utilized in our study to distinguish between anti-social and socially acceptable comments. Finally, language models such as unigrams, bigrams, and trigrams have been implemented in our research. To the best of our knowledge, there are no studies regarding the anti-social classification in Bangla language. This work will help to prevent anti-social activities in Bangla community.**

Index Terms—**Classification, GRU, MNB, Supervised algorithms, SVM**

I. Introduction

Human desire is to live as a social being and be part of a social group. In July 2021, there were 4.48 billion social media users worldwide, equivalent to nearly 57 percent of the total global population [1]. In the year leading up to July 2021, 520 million new users joined social media, spending an average of about 2.5 hours each day on the medium [1]. Even two decades ago, a huge number of social media platforms would have been inconceivable. Facebook is without a doubt the most popular social networking platform. Facebooks monthly active users are estimated to be over 2.2 billion, with an 8 percent annual growth rate. Approximately 400 new users register per minute, and 1.59 billion users check in to Facebook each day. Per minute, 400 new users register on Facebook [2]. In Bangladesh, Facebook is used by 88% of people, whereas YouTube is used by 6.5% [3]. In July 2021, there were 50 million Facebook users, equivalent to 29.5 percent of the total population of the country. More than 71 percent of Facebook users are between the ages of 13 and 17 [4]. Dhaka is the second-most-active Facebook user city in the world [5]. As of 2020 and 2021, more than 9 million new users from Bangladesh have joined the social media network [6]. Recently, cybercrime, cyberbullying, harassment, and toxicity have become more prevalent on social media platforms [7]. Facebook addiction was 2.51 and 1.67 times higher among the students with a domestic violence history and depressive symptoms, respectively [8].A recent study shows that high school students who use social networking sites like Twitter, Facebook, and Instagram are more likely to have mental health issues [9]. According to an article from 2012, Facebook users create approximately 3.2 billion likes and comments every day [10]. Nearly 49% of pupils in Bangladesh have been victims of cyberbullying. Bengali is the worlds sixth most commonly spoken language [11]. It is also official language Bangladesh and West Bengal. As a result, we focused on Bengali Facebook and YouTube comments in order to detect anti-social material and develop methods to safeguard social media from anti-social behaviors. In a nutshell, in this paper, we have made the following contribution-

- A dataset of 2000 comments is developed from popular public pages and channels on Facebook and YouTube. A web tool named exportcomments.com was used to collect these comments. The sentiment is a persons perception of or attitude toward a circumstance or occurrence. Individual freedom of expression with respect to social views is frequently neglected in sentiment analysis. In the training dataset, binary classification was applied to separate social and anti-social comments. The anti-social comment was defined if it has any slang, abusive, hatred, toxic words towards any individual or community.
- To analyze the performance of machine learning algorithms such as LR, RF, MNB, GRU, etc. To compare the performance of the various models, evaluation metrics were implied. An accuracy of 80.51% and 78.89% was

978-1-6654-6944-9/22/$31.00 © 2022 IEEE

achieved respectively in MNB and GRU. In the near future, with the increase of resources in NLP for Bengali, ANN and other classifiers in machine learning hopefully will perform with better accuracy.

II. Related Works

A recent study has shown that anti-social activities such as cybercrime, cyberbullying, harassment, and toxicity spread through social media. In [12], the author introduced anti-social comment classification based ASB Corpus. One of the first papers [13] employed machine learning algorithms to classify attitudes in Twitter posts automatically. In [14], they used N-grams, Linguistic, and Syntactic characteristics to measure distinct parts of the online users comments. Another Paper [15] implemented machine learning and topic modeling approaches to identify profanity-related abusive posts on Twitter. They outperformed keyword-based techniques with a true positive rate of around 75%. In this field, some work has been done in the Bengali language. This paper [7], compared the accuracy of various machine learning algorithms, including GRU and Random Forest, and discovered that Random Forest reaches 52.2% accuracy. The GRU-based models accuracy (70.1%) has increased by about 18%. Linear SVC, LR, RF, ANN, and RNN with a Long Short Term Memory (LSTM) cell are the machine learning and deep learning algorithms implemented in this paper [16]. They gained the highest accuracy (82%) in RNN. The attention mechanism, LSTM, and GRU-based decoders are used to predict hate speech categories in this research [7]. The attention-based decoder was the most accurate of the three encoder-decoder algorithms (77%). There are several methods and techniques to analyze the social media Comments described in the research. However, relatively little study has been done in the Bengali language, which is continuously expanding. Considering all the limitations mentioned above, we compare and try to improve the performance of these methods and techniques in the Bengali language.

III. Methodology

In Bengali language, detecting anti-social, derogatory, or abusive text in social media is still a relatively new study subject in NLP [17]. There are two types of classifiers for text classification: binary and multi-label. Only binary classification can determine whether a comment is social or anti-social. The multi-label classifier categorizes hate speech based on ethnicity, gender, sexual orientation, and other factors. In Bengali comments, we implemented a binary classification. In Fig. 1, the workflow of the model is illustrated step by step through the flowchart.

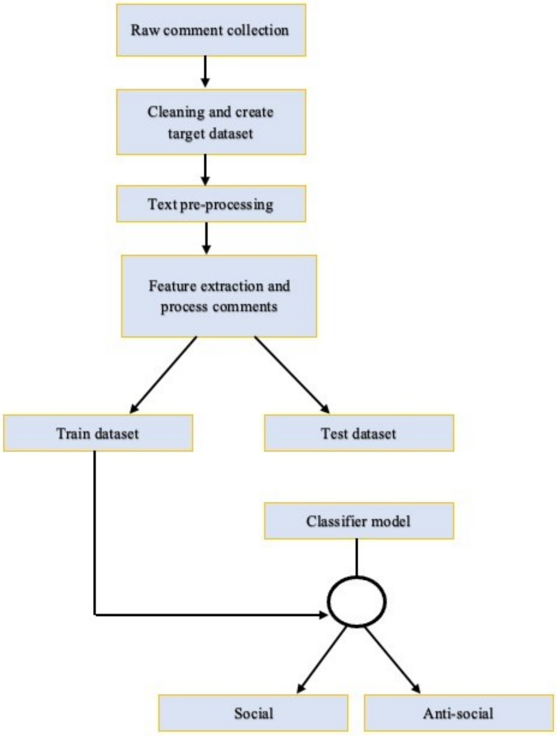

Fig. 1. Flowchart of the anti-social comment detection model

A. Dataset collection

Due to few available studies for anti-social comment identification on the Bengali language, a new dataset was created. The new dataset contains comments from different categories, e.g., sports, crime, religion, news, politics, entertainment, etc. from different Facebook pages and YouTube channels, e.g., Prothom Alo [18], Shakib Al Hasan [19], SalmoN TheBrownFish [20], TahseeNation [21], Dr. Muhammad Zafar Iqbal [22] and Solaiman Shukhon [23], etc. A website called ExportComment has extracted all the comments from the mentioned sources. After deleting unnecessary columns, the final selected are No., Category, Social acceptance, Comment. Finally, the data was stored in a CSV format. The dataset contains 2000 comments equally balanced between both classes. The train set and test set were respectively 90% and 10%. Label Encoder was applied to train and test sets to convert them into binary values given into the ML model.

B. Dataset Annotation

Social perception regarding anti-social norms and behaviors varies to individuals and communities. Thus, defining what constitutes hate speech is a challenging task. Strict guidelines are established according to the community standard of Facebook [24] and YouTube [25] which are followed throughout the data annotation process.

- An anti-social comment is a phrase that dehumanizes someone or even more individuals or a group [24], [25]. Comparing a group or person to an insect, an item, or a criminal can be used to dehumanize them. Targeting a person's ethnicity, gender, or physical or mental impairment is another way to do it.
- Slang or negative word may be used in a sentence. However, slang is not considered anti-social unless it dehumanizes a person or community.

- Positive or negative sentiment is not focused on comments. As long as a negative sentiment comment is in a good manner, do not use slang, do not spread hatred towards a community or region, then it is considered a socially accepted comment. Some examples are given below.

"মা হিসেবে আপনি জাতির গর্ব কারণ আপনার সন্তানকে আপনি অনেকগুলা বাবা উপহার দিয়েছেন"
(As a mother you are the pride of the nation because you have given many fathers to your child.)
-In this scenario, an individual character such as a mother was questioned, which is not socially acceptable. Hence this is termed as anti-social.

“আর ভারতে যাইয়া পুজোর উদ্বোধন করে আসো"
(Travel to India to inaugurate Puja.)
-Although no obscene words are used in this sentence, it offends a community's religious emotions. So, it is an anti-social comment.

“হিরো কনডম সবসময় এক নাম্বার"
(Hero condom is always number one.)
-Here no slang word is used or no one is dehumanized. So, it is not an anti-social comment.

“এটা ই যেন শেষ আন্দোলন হয়"
(Hope it will be the last movement.)
-It yearned for the end of a movement indicating peace. which is socially acceptable.

All annotators were instructed to follow the above-mentioned guidelines.

C. Data Preprocessing

After completing the annotation of our dataset, we focus on data cleaning and preprocessing.

1) Correction of typos, punctuation, etc.: Texts were collected and then sanitized by eliminating non-letters such as punctuation like comma (,), dot (.), semicolon(;) etc. The system was improved by removing the noisy and undesirable characters. This was necessary to ensure accurate Unicode encoding.

2) Tokenization of the comments string: String tokenization is the process of dividing a string into multiple pieces. Each component is referred to as a token. Tokens contribute to comprehending the context or facilitating the process of developing models for the (NLP). Fig 3 illustrates tokenization.

3) Removal of Bengali stop words: Removing stop words may increase performance because there are fewer and only relevant tokens remaining. 500 Bengali stop words were used for Removing stop words. Stop words contain common Bangla words like "অথচ","অথবা ", "অনুযায়ী" etc. Fig. 2 represents the length-frequency distribution of comments after removing the stop words.

D. Feature Extraction

That step comprises obtaining appropriate feature characterizations, depending on the type of NLP function and the type of model individuals plan to use.

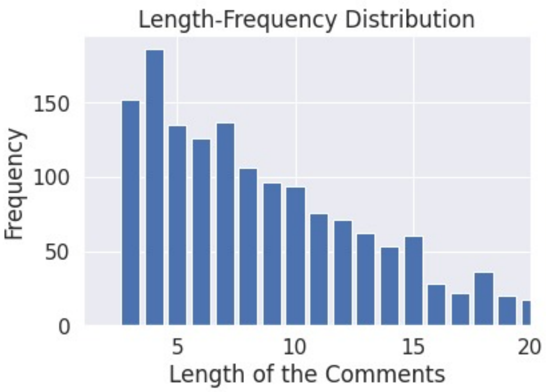

Fig. 2. Length of frequency

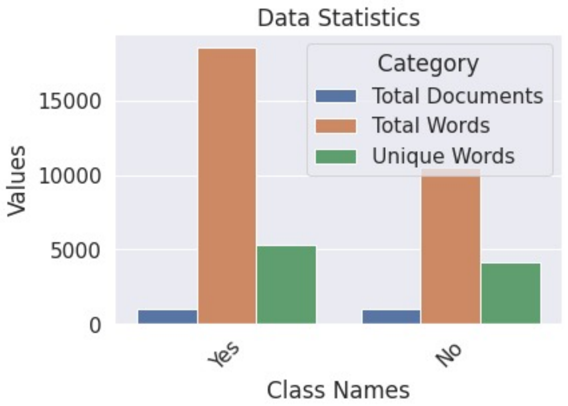

Fig. 3. Tokenization

Using TF-IDF to model N-grams:
The TF-IDF evaluates the significance of words in a text. The TF-IDF value grows in lockstep with the number of times a word occurs in the document. Some words appear higher than others, and step helps to compensate for that fact [26].

$$TF\left(t\right) = \frac{number\ of\ times\ appearance\ of\ term\ t}{total\ number\ of\ terms}$$

$$IDF\left(t\right) = log\frac{total\ number\ of\ documents}{number\ of\ term\ t\ with\ documents}$$

$$TF - IDF\left(t\right) = TF(t) \times IDF(t)$$

For example, the word “রাজাকার” itself is not a slang and does not dehumanize directly. So when n = 1, it does not specify as anti social. But for bigram or trigrams(n=2,3) like “সাংবাদিকরা রাজাকার”, it can be classified as anti social. To determine the anti-social comment with TF-IDF weights we added N-gram modeling [27].

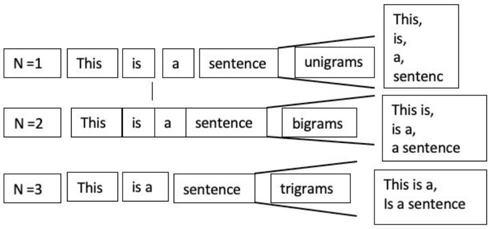

Fig. 4. N-grams modeling

E. Model selection

We must feed our data and features into appropriate models after the pre-processing stages and feature selections. For our study, we applied both ANN like GRU and supervised machine learning classifiers like Logistic Regression, Random Forest, Multi-nominal Naive Bayes (MNB), SVM. Finally, MNB and GRU performed better than other classifiers in this study. Table I represents the experimental settings for the GRU model. A bidirectional GRU layer was used in the model architecture. For the final dense layer, softmax activation function was used. The embedding dimension was set to 64.

TABLE I
EXPERIMENTAL SETTINGS OF THE GRU MODEL

Settings	*Values*
batch size	30
epoch	15
drop out	0.2
optimizer	‘adam’
embedding dimension	64

F. Evaluation metrics

The models were compared using the following evaluation metrics:

Accuracy: Total amount of correct classified data divided by the total amount of input data [28].

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP}$$

Precision: Determine what fraction of positive predictions is right [28].

$$Precision = \frac{TP}{TP + FP}$$

Recall: Recall is defined as the proportion of accurately predicted observations to all observations in the actual class [28].

$$Recall = TP/TP + FN$$

Here TP is true positive and FN is False Negative. F1 score: Precision and Recall are weighted into the F1 Score [28].

$$F1Score = 2 * (Recall * Precision)/(Recall + Precision)$$

IV. RESULT ANALYSIS AND DISCUSSION

After data cleaning and removal, in feature extraction TF-IDF with N-grams model was applied. At initial step, supervised classifiers such as LR, RF, MNB, Linear SVM and Gated Recurrent Unit (GRU) were used. MNB has the highest accuracy of 80.51% where recall, precision are 81.55%. GRU has the second highest accuracy of 78.89%. LR, RF and Linear SVM has the accuracy of 74.36%, 71.28%, 70.26% respectively.

TABLE II
ACCURACY OF MODELS

Model	*Accuracy*	*Precision*	*Recall*	*F1 Score*
LR	74.36	78.49	70.87	74.49
RF	71.28	80.52	60.19	68.89
MNB	80.51	81.55	81.55	81.55
L. SVM	70.26	89.47	49.51	63.75
GRU	78.89	78.89	90.00	87.66

From Table I, there is a 70-80 percent accuracy range for different models. Random Forest and Linear SVM have the accuracy of 71.28% and 70.26%. LR and MNB have the accuracy 74.36% and 80.51%. 78.89% accuracy is found on ANN based GRU. Fig. 5 represents the confusion matrix of the GRU model where out of 92 positive instances, 74 were true positives and out of 88 negative instances, 68 were true negatives.

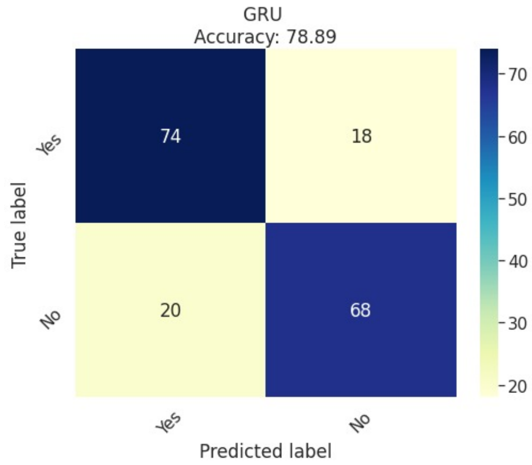

Fig. 5. Confusion matrix of Gated Recurrent Units(GRU) on our dataset

Table I also presents that GRU has the best f1 score of 87.66% where Naive Bayes model has the f1 score of 81.55%.

Our study suggests that, MNB performs well on accuracy compared to other models. Though, the accuracy metrics of GRU and MNB are pretty close to each other. Among the ANN models, GRU works well with small datasets and it is generalized faster [29]. Besides, the purpose of the ANN architecture is to imitate sequential occurrences such as word sequences. ANN can retrieve relatively high information and make judgments depending on every word’s context which results in better performance. On the other hand, MNB is also better with shorter documents than other non-NN models. Besides, MNB assumes that every word of a sentence as independent [30].

One of the limitations of the study is that the size of the dataset is small. By increasing the size of the dataset, the performance of the ANN models can be increased.

A. Comparison of study

[19] and [28] have used GRU based model respectively with word2vec and word embedding modeling. In our study TF-IDF with N-grams modeling approached for MNB have been compared. In the same approach, GRU has 78.89% accuracy.

TABLE III
ACCURACY OF RELATIVE STUDY

Paper	*Accuracy*
A. K. Das et. al [16]	83%
A. M. Ishmam et. al [31]	70.21%
Our approach	80.51%

While the above mentioned papers from Table II were all about sentiment analysis, our study took constructive criticism into account and labeled those as socially accepted comments which was mentioned in the dataset annotation.

V. CONCLUSION

There has been a widespread adoption of anti-social comments on social media platforms now-a-days. To preserve user freedom, it has become a significant problem to prevent the widespread usage of anti-social comments in Bengali and other languages. In this research, a dataset is developed and machine learning models are used to automatically defend social media against anti-social comments. Comments were analyzed by binary classification based on socially acceptable or not. Finally, the performance of various machine learning and deep neural network models was compared to discover that GRU and MNB perform better than other models. Antisocial comments or activities have significant effects on mental health, can cause conflict between communities and break the harmony of society. This work will help to prevent these by detecting such type of comments. In future, a Bangla corpus will be developed from the dataset. Additionally, the dataset will be enriched to allow for more in-depth research and implementation of other deep learning algorithms to increase accuracy.

REFERENCES

[1] "Datareportal," https://datareportal.com/social-media-users., accessed: August 2021.

[2] E. Stănculescu and M. D. Griffiths, "Anxious attachment and facebook addiction: The mediating role of need to belong, self-esteem, and facebook use to meet romantic partners," *International Journal of Mental Health and Addiction*, pp. 1–17, 2021.

[3] "statcounter," https://gs.statcounter.com/social-media-stats/all/bangladesh, accessed: 24 August 2021.

[4] A. Kaye, "Facebook use and negative behavioral and mental health outcomes: A literature review," *Journal of Addiction Research & Therapy*, vol. 10, no. 1, pp. 1–10, 2019.

[5] M. Murad, "bdnews24," https://bdnews24.com/bangladesh/2017/04/15/dhaka-ranked-second-in-number-of-active-facebook-users, accessed: 25 August 2021.

[6] "Dhakatribune," https://www.dhakatribune.com/bangladesh/2021/04/26/bangladesh-charts-9m-new-social-media-users, accessed: 25 August 2021.

[7] E. A. Emon, S. Rahman, J. Banarjee, A. K. Das, and T. Mittra, "A deep learning approach to detect abusive bengali text," in *2019 7th International Conference on Smart Computing & Communications (ICSCC)*. IEEE, 2019, pp. 1–5.

[8] A. Sayeed, M. N. Hassan, M. H. Rahman, S. El Hayek, M. H. Al Banna, T. Mallick, A.-R. Hasan, A. E. Meem, and S. Kundu, "Facebook addiction associated with internet activity, depression and behavioral factors among university students of bangladesh: a cross-sectional study," *Children and Youth Services Review*, vol. 118, p. 105424, 2020.

[9] H. Sampasa-Kanyinga and H. Hamilton, "Social networking sites and mental health problems in adolescents: The mediating role of cyberbullying victimization," *European psychiatry*, vol. 30, no. 8, pp. 1021–1027, 2015.

[10] M. McGee, "Martech," https://martech.org/facebook-3-2-billion-likes-comments-every-day/, accessed: 25 August 2021.

[11] "wikipedia," https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers, accessed: 25 August 2021.

[12] N. Chandra, S. K. Khatri, and S. Som, "Anti social comment classification based on knn algorithm," in *2017 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*. IEEE, 2017, pp. 348–354.

[13] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N project report, Stanford*, vol. 1, no. 12, p. 2009, 2009.

[14] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th international conference on world wide web*, 2016, pp. 145–153.

[15] G. Xiang, B. Fan, L. Wang, J. Hong, and C. Rose, "Detecting offensive tweets via topical feature discovery over a large scale twitter corpus," in *Proceedings of the 21st ACM international conference on Information and knowledge management*, 2012, pp. 1980–1984.

[16] A. K. Das, A. Al Asif, A. Paul, and M. N. Hossain, "Bangla hate speech detection on social media using attention-based recurrent neural network," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 578–591, 2021.

[17] K. Kumari and J. P. Singh, "Ai ml nit patna at hasoc 2019: Deep learning approach for identification of abusive content." in *FIRE (Working Notes)*, 2019, pp. 328–335.

[18] "Facebook," https://www.facebook.com/DailyProthomAlo, accessed: 25 August 2021.

[19] "Facebook," https://www.facebook.com/Shakib.Al.Hasan, accessed: 25 August 2021.

[20] "Youtube," https://www.youtube.com/user/salmanmuqtadir, accessed: 25 August 2021.

[21] "Youtube," https://www.youtube.com/user/TahseeNation, accessed: 25 August 2021.

[22] "Facebook," https://www.facebook.com/our.zafar.sir, accessed: 25 August 2021.

[23] "Facebook," https://www.facebook.com/Muhammad.Solaiman, accessed: 25 August 2021.

[24] "Hate speech policy - facebook," https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/, accessed: 9 September 2021.

[25] "Hate speech policy - youtube help," https://www.youtube.com/howyoutubeworks/policies/community-guidelines/, accessed: 9 September 2021.

[26] T. Roelleke and J. Wang, "Tf-idf uncovered: a study of theories and probabilities," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 435–442.

[27] "Deepai," https://deepai.org/machine-learning-glossary-and-terms/n-gram, accessed: 25 August 2021.

[28] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, p. 1, 2015.

[29] Ł. Kaiser and I. Sutskever, “Neural gpus learn algorithms,” *arXiv preprint arXiv:1511.08228*, 2015.

[30] A. Y. Ng and M. I. Jordan, “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes,” in *Advances in neural information processing systems*, 2002, pp. 841–848.

[31] A. M. Ishmam and S. Sharmin, “Hateful speech detection in public facebook pages for the bengali language,” in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 2019, pp. 555–560.