2021 24th International Conference on Computer and Information
Technology (ICCIT), 18-20 December, 2021.

Social Media Opinion Mining Based on Bangla Public Post of Facebook

Shad Al Kaiser, Sudipta Mandal, Ashraful Kalam Abid, Ekhfa Hossain, Ferdous Bin Ali[1], Intisar Tahmid Naheen
ECE Department, North South University, Dhaka
[1]Statistics Department, Jahangirnagar University
{shad.kaiser, sudipta.mandal, ashraful.abid, ekhfa.hossain, intisar.naheen}@northsouth.edu
hridoyferdous@yahoo.com

***Abstract*—Social media holds the freedom to express anyone as they are. Still, people fail to follow community standards and cross the boundary of self-limit, hurting other people, sometimes leading to cyberbullying. Social media mining is the frontier where researchers contend with ensuring safe cyberspace with the help of robust information retrieval and data mining techniques. In this paper, we are aiming towards achieving such a goal for Bangla language-spoken people. We have created a corpus that contains 11006 Bangla comments from Facebook, analyzed them demographically, annotated them to create robust classifiers to classify these comments as positive, negative, and neutral polarity. We have decomposed these polarities to further sentiments based on contents of the text varying from wishful thinking to gender-based hate speech. Our multiclass classification algorithm, consisting of TF-IDF vectorizer alongside uni-gram, bi-gram, and tri-gram followed by MNB, MNB, and KNN, gives 82.60%, 82.33%, and 79.63% accuracy, respectively.**

***Index Terms*—Social Media Mining, Opinion Mining, Bangla Corpus, Online Harassment, Bully Detection, Sentiment Analysis, Natural Language Processing, Data mining**

I. Introduction

There are 4.48 billion people who actively use social media in 2021, with an increased rate of 13.13% from 2020 [11]. Facebook leads the social networks, having 2.9 billion active users monthly, followed by YouTube (2.3 billion), WhatsApp (2 billion), FB Messenger (1.3 billion), and WeChat (1.2 billion) [14]. The primary motivation behind this significant cyberspace is to make people feel connected and secure. But as the number is increasing, it's getting difficult to maintain the community standards. As a result, it is reported by Ditch the Label, one of the world's leading anti-cyberbullying organizations, that 37% of people are affected by different kinds of hate speech and cyberbullying, and 15% of people admitted that they have bullied and spread hate speech towards someone online [2]. Statistics show that 17% of people face this alarming situation due to their race, 15% for their sexuality, 15% for financial status, and 11% for religion [1].

Nowadays, 46 million people in Bangladesh use Facebook, and 60% of them have experienced harassment at least once in their social media lifetime [12]. Though it's a common research trend for English and other high resource languages to develop social media monitoring and analytical tools [24], it's rare for low resource languages like Bangla. It lacks necessary computational linguistic resources like corpus, language models, and powerful machine learning methods for performing various NLP tasks [13].

Our contribution: Our contribution: In our research, we have developed a properly annotated corpus and made robust machine learning classifiers to classify online hate speech as a stepping stone for social media monitoring tools. Our main contribution goes as follows

- We have created a carefully annotated dataset consisting of 11006 comments with their corresponding reactions to those comments.
- Our dataset is mainly focused on social media celebrities. We have explored data from different public figures based on profession like actor-actress, players, religious leaders, social media influencers, and so on.
- We have ensured the quality of annotation both manually and statistically.
- We have done exploratory demographic analysis to explore patterns in the nature of people who are involved in spreading hate speech and cyberbullying.
- We have annotated data into 3 polarities and 8 different sentiments.
- We have experimented with traditional machine learning classifiers to classify different sentiments.

This paper is organized as follows. Section II describes previous research literature. In Section III, we have discussed our research methodology and demographic exploratory data analysis. The experiment setup is presented in section VI. In section V, we have analyzed our results. Finally, Section VI concludes our research and focuses on our future research direction

II. Literature Review

In this part, we review some related materials that were relevant to our research. For hate speech on Bangla comments, Khan [15] et al. proposed an SVM based model where they used TF-IDF to process data and multiple classification models to yield the accuracy. Das [8] et al. proposed a model based on encoder-decoder. Comments were categorized into 7 distinct

978-1-6654-9435-9/21/$31.00 ©2021 IEEE

categories of speech. All contents were divided into hateful and non-hateful categories. To extract and encode feature from Bangla comments 1D convolutional layers were used. And at last, to predict hate speech the attention mechanism LSTM and GRU(Gated Recurrent Unit) - based decoders were used. Romim [21] et al. used many deep learning models which performed well. The approach they have conducted is a baseline experiment. Several deep learning models along with a lot of Bangla words(pre-trained) embedding such as FastText, Word2Vector, and BengFastText were used on this dataset. They have used 30k comments tagged by crowdsourcing where 10k is hate comments. Das & Bandyopadhyay [9] proposed an opinion polarity classification. The dataset was built from news text of Bangla sources and the authors used SVM to predict the accuracy. Nabi [19] et al. used TF-IDF feature extraction on Bangla text to generate results.

We further review Countering online hate speech, Mathew [17] et al. gave us an idea on how to analyze the dataset. The authors compiled a dataset with 6,898 counterspeech comments and 7,026 non-counterspeech comments. They analyzed the psycholinguistic effects of counterspeech and non-counterspeech. The targeted communities for the dataset were Jews, African-Americans and LGBT communities. Mathew [16] et al. also analyzed counter speech on Twitter. Table I compares the results for the models we have reviewed.

TABLE I: Different scores for different types of classifiers we reviewed in relevant literature studies.

Authors	**Approach**	**Accuracy**	**Recall**	**F1**	**Precision**
Khan [15] et al.	SVM	59	-	-	-
Das [8] et al.	CNN	77	-	-	-
Romim [21] et al.	SVM	87.5	-	91	-
Das & Bandy-opadhyay [9]	SVM	-	63	-	70
Nabi [19] et al.	SVM	83	-	-	-
Mathew [17] et al.	XGB+SV +TFIDF+ BOWV	71.5	715	71.5	71.6
Mathew [16] et al.(Twitter)	CB	78	78	77	83

III. Methodology

To start off the research we first collected data from public Facebook posts. We manually annotated those data according to our annotation index mentioned in Table II. After validating the annotation we cleaned the data and preprocessed it for classification purposes. At the same time, we performed various types of exploratory data analysis. These analysis provided us with various types of demographical information. Then we used traditional machine learning algorithms to build classification models based on our data. Fig. 1 illustrates the overall process from start to finish of our research.

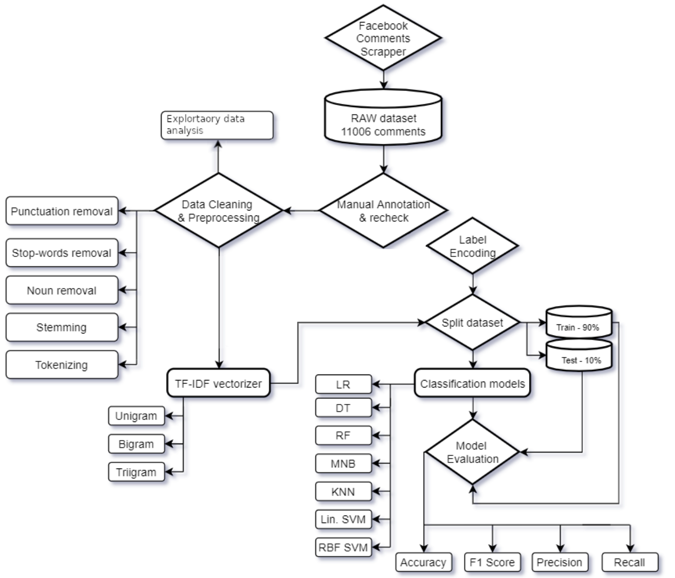


Fig. 1: Process flowchart

A. Data collection from Facebook

To make a dataset with different types of Bangla texts, we identified Facebook pages of Bangla celebrities, politicians, religious figures, actors, actresses, cricketers, social media influencers, singers and various other professions of people. We collected data from public posts with the help of "Instant Data Scrapper" [20]. We extracted the comments and compiled them in an excel file. For privacy, we didn't include any names or profile links. Our dataset consists of the comments, the number of reactions and the number of replies each comment has.

B. Data annotation

For annotation purposes, four annotators(all native Bangla speakers) went through the comments of the dataset. We divided the comments into three main categories, positive, negative and neutral. Positive and negative have their own subcategories with an assigned index number for each of those subcategories. Table II shows the details of the annotation types. We identify the hate comments based on the community standards of Facebook [4]. If any comment didn't follow the

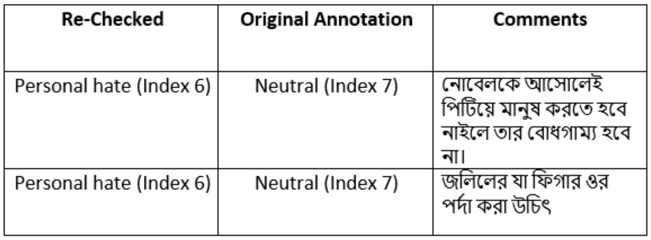
TABLE II: Annotation categories

Category	**Sub category**	**# of Comments**	**Percentage**	**Ann. Index**
Positive	Wishful thinking	967	8.8%	1
	Appreciation	942	8.6%	2
Negative	Gender-based hate	525	4.8%	3
	Religious hate	731	6.6%	4
	Political hate	572	5.2%	5
	Personal hate	1995	18.1%	6
	Sarcasm	1414	12.8%	8
Neutral	N/A	3860	35.1%	7

rules we put that comment in the negative category and after further analysis, we decided on a subcategory for the comment.

C. Annotation validation:

After completing the annotations we assigned all four of our annotators to cross-validate the annotation done by their counterparts. The validation process corrected the original annotator's annotation when needed. We give an example on Fig. 2.



Re-Checked	Original Annotation	Comments
Personal hate (Index 6)	Neutral (Index 7)	নোবেলকে আসোলেই পিটিয়ে মানুষ করতে হবে নাইলে তার বোধগাম্য হবে না।
Personal hate (Index 6)	Neutral (Index 7)	জলিলের যা ফিগার ওর পর্দা করা উচিৎ

Fig. 2: Annotation validation

If we translate the first comment it becomes "We have to beat up Noble to make him human so that he understands". Originally this comment was tagged as neutral which is the wrong tag for this particular type of comment. So the cross checker revised the comments and annotated them as personal hate. The same goes for the second comment, where the comment was a body-shaming remark.

We used Cohen's Kappa statistic [7] to evaluate the quality of annotation. If $m(\geq 2)$ annotators annotated n comments into mutually exclusive categories such that $k(\geq 2)$, the proportion of score $\bar{p}_j$ and the kappa $\hat{k}_j$ for category j are computed as,

$$\bar{p}_j = \frac{\sum_{i=1}^{n} x_{ij}}{nm} \tag{1}$$

$$\hat{k}_j = 1 - \frac{\sum_{i=1}^{n} x_{ij}\left(m - x_{ij}\right)}{nm(m-1)\bar{p}_j\left(1 - \bar{p}_j\right)} \tag{2}$$

x_{ij} is the score for i into category j. The kappa $\hat{\bar{k}}$ is:

$$\hat{\bar{k}} = \frac{\sum_{j=1}^{k} \bar{p}_j\left(1 - \bar{p}_j\right)\hat{k}_j}{\sum_{j=1}^{k} \bar{p}_j\left(1 - \bar{p}_j\right)} \tag{3}$$

We achieved a kappa value of 0.92 which indicates that the level of agreement was almost perfect [10].

D. Exploratory data analysis

Dataset description: We collected 11006 comments from various public posts and after annotation process, we summed up the number of comments by each subcategory as shown in Table II. The dateset consists of 47.6% negative comments, 17.3% positive and 35.1% neutral comments.

Gender-based comments analysis: We analyzed the data by gender of our samples and inspect the proportion of hate comments. Fig. 3 shows the percentage of comments on each subcategory and category for male samples.

TABLE III: Data description

Category	# of Sentence	Max Length of Sentence	Min Length of Sentence	Avg length of sentence
Positive	1869	65	2	14
Negative	5201	153	2	14
Neutral	3803	66	2	13

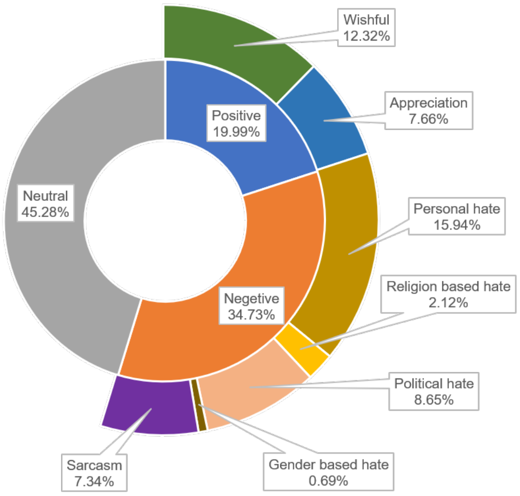


Fig. 3: Comment analysis for male samples

For male samples, almost half the comments are neutral. Out of the negative category, the majority of the comments are personal hate, 15.94%. On the other hand, Fig. 4 shows that the number of neutral comments is low in our female samples, only 19.99%. Almost 2/3 of the comments on female pages are negative, where personal hate occupies 21.36%, 10.81% gender related hate which is far more than the 0.69% of gender

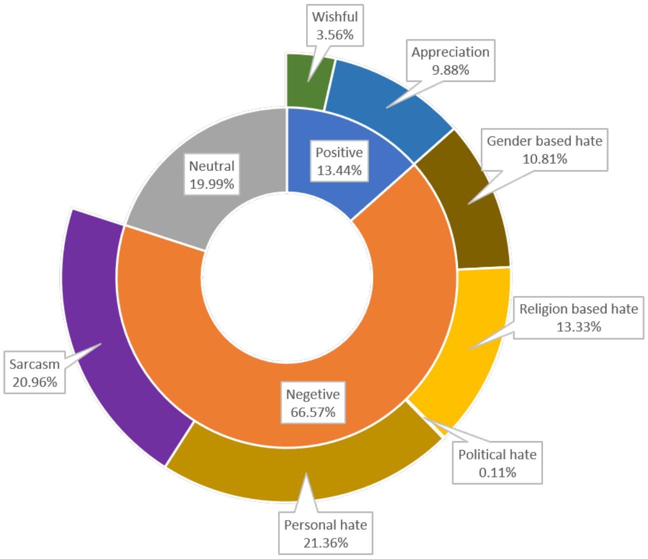


Fig. 4: Comment analysis for female samples

based hate comments on male samples. We also see more religious hate speech on female samples, 13.33% while male samples had 2.12%.

Reaction based analysis: From our dataset, we analyzed the total number of reactions on comments on different criteria. Fig. 5 shows us the reactions count for each of the subcategories. Out of all the subcategories, neutral types comments had the most reaction from people.

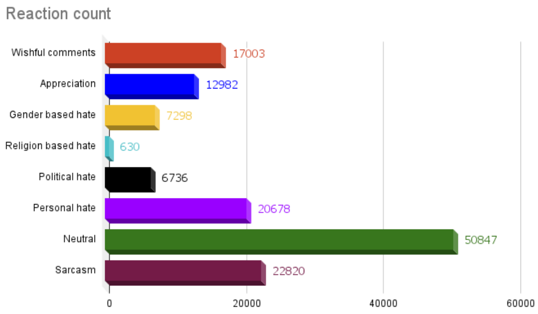


Fig. 5: Reaction count

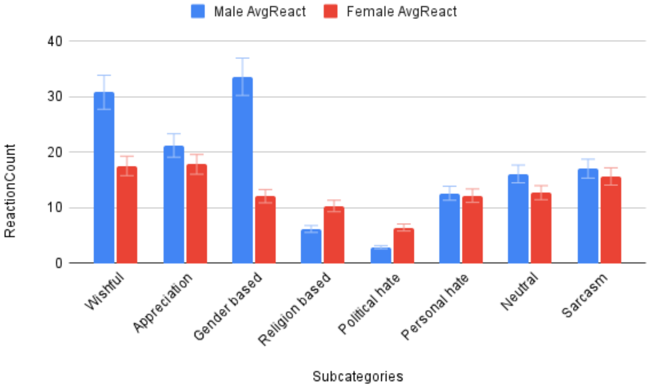


Fig. 6: Average reaction on male vs female samples

For further analysis, we filtered the reaction count based on the gender of the samples. Fig. 6 shows the average reaction on male and female sample pages. Male pages had more reactions on comments that contained gender based hate and positive wishful comments than female pages. Other subcategories were pretty close to each other in terms of average number of reactions.

Profession based analysis: As mentioned, we collected data from actors to politicians to social influencers, we have tried to include people from all walks of life. Fig. 7 is a summary of our dataset based on the professions of our samples.

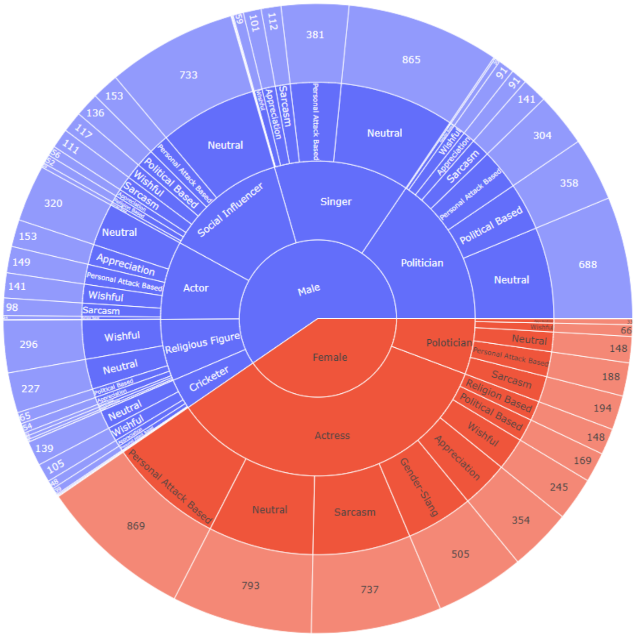


Fig. 7: Sunburst chart of the dataset

Actresses got 896 personal hate comments compared to actors who received 149 personal hate comments. Actresses also received 505 gender based hate comments, compared to that actors' gender based hate comments are negligible. It shows that the people in the comments are more hostile towards women than men.

E. Data cleaning:

For data cleaning process we used BNLP [22] and BLTK [3] toolkits. The BNLP Bangla Corpus class to remove Bangla stop words and punctuation from our data. Then we tokenized our data to word form using the BLTK word tokenizer. From these tokens, we removed the names of any sample pages. Stemming is also an important part of data cleaning that is why we used the Fatick Stemmer [5]. Using the stemmer we turned the tokenized words into their root form. Bangla is a very rich language so the process of stemming words is very complicated, that is why we used this lightweight stemmer to determine the stem or identical words as the stem.

F. Data preprocessing:

Preprocessing the data means removing punctuation, Bangla stop-words, any names of the sample since we are committed to protect the privacy of our samples. We also drop the data that are smaller than 2 in length. Then the cleaned data is passed on to the tokenizer and vectorizer to handle the rest of the classification process.

G. Label encoding and features extraction:

We used label encoding to encode the target with a value between 0 and n_classes-1. Next, we vectorize the data using Tfidf vectorizer. Tfidf stands for Term Frequency Inverse Document Frequency. The idea behind Tfidf is to compare the number of times a word appears in a doc to the number of docs the word appears in. It essentially converts our data into a matrix of 'features'.

We can represent the mechanism as [18], for word i in comments j :

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$w_{i,j}$ = feature matrix
$tf_{i,j}$ = number of occurrences of i in j
df_i = number of comments containing i
N = total number of comments
For the ngram_ranges in the vectorizer, we used unigram (1,1), bigram (1,2) and trigram (1,3). For unigram, our vector dimension or feature size is 22873, for bigram, it is 124897 and for trigram, it is 242286. Fig. 8 illustrates how the different ngram_ranges are used to extract features for models for each of the ngram we used for Bangla text [23].

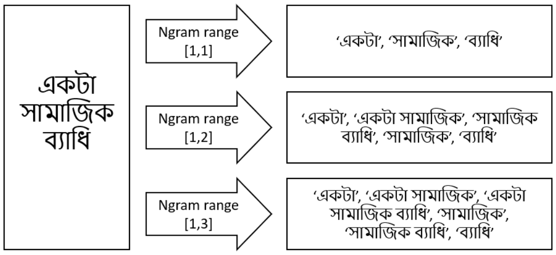

একটা সামাজিক ব্যাধি
Ngram range [1,1]
‘একটা’, ‘সামাজিক’, ‘ব্যাধি’
Ngram range [1,2]
‘একটা’, ‘একটা সামাজিক’, ‘সামাজিক ব্যাধি’, ‘সামাজিক’, ‘ব্যাধি’
Ngram range [1,3]
‘একটা’, ‘একটা সামাজিক’, ‘একটা সামাজিক ব্যাধি’, ‘সামাজিক’, ‘সামাজিক ব্যাধি’, ‘ব্যাধি’

Fig. 8: ngram_range example

H. Classification:

After cleaning, preprocessing and vectorizing the data using TF-IDF, we created the train and test split. We used 90% of the data for our train feature vector and the other 10% for the test feature vector. We used LogisticRegression, DecisionTreeClassifier, RandomForestClassifier, MultinomialNB, KNeighborsClassifier, SVM with Linear kernel and SVM with RBF kernel for classification. We define these models every time for each of the TF-IDF gram feature vectors.

IV. Experiment Setup

To evaluate the model we used Google Colaboratory (GPU runtime), a free online based jupyter notebook. We defined the TF-IDF vectorizer based on our dataset. We also used Google Drive for storage purposes. We saved our performance parameter into json file and stored it in Drive to conclude our evaluation.

V. Result Analysis

To evaluate our classification models we used four measures, accuracy, F1-score, precision and recall. Here is a brief explanation of these measures: [6]

Accuracy: Accuracy is defined by the proportion between correctly predicted data(TP, TN) and total number of data(TP, TN, FP, FN).
TP=True positive, TN=True negative, FP=False positive, FN=False negative

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{4}$$

Precision: In an ideal situation precision should be 1 because it means how accurate the classifier is. It is defined by the proportion of *True Positive* with respect to total number of positive(*TP*+*FP*). So when the precision is 1 it means the *False Positive* is 0 making the numerator and denominator equal.

$$precision = \frac{TP}{TP+FP} \tag{5}$$

Recall: Like precision, recall is ideally 1 for a good classification model. It is defined by the proportion of *True positive* and total number of all positive instances.

$$recall = \frac{TP}{TP+FN} \tag{6}$$

F1-Score: F1-Score is defined as the *Harmonic Mean* of *precision* and *recall*.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{7}$$

Performance: Table IV demonstrates the result we obtain for classifications models on unigram, bigram and trigram feature for all the classification models we used.

For **unigram** highest accuracy, f1-score, precision and recall were achieved by MNB at 82.6, 81.56, 81.74, 82.6 respectively.

For **bigram** highest accuracy, f1-score and recall were achieved by MNB at 81.33, 82.36 and 81.33 respectively and RBF SVM had the best precision at 82.18.

For **trigram** highest accuracy, f1-score and recall were achieved by KNN at 79.63, 79.80, 79.63 respectively and LR had the best precision at 82.18

Confusion matrix: We analyze the number of correct classification number and miss-classification number with the help of confusion matrix on Fig. 9

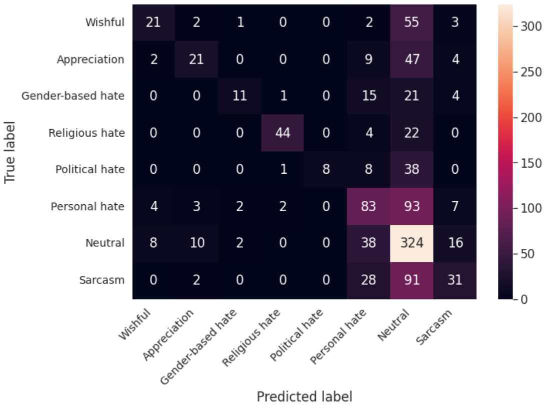


Fig. 9: Confusion matrix

Since we had the best accuracy with MNB classifier with unigram features, we plot the confusion matrix for that model. If we look at religious hate for example, 62.85% data were predicted correctly. Amongst the miss-classification, 5.72%

TABLE IV: Performance table for unigram, bigram and trigram

Model Name	Unigram feature				Bigram feature				Trigram feature			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
LR	80.2	81.44	80.2	76.22	77.37	80.93	77.37	70.51	76.52	**82.18**	76.52	68.48
DT	76.8	75.7	76.8	76.1	79.07	77.77	79.07	78.01	78.5	77.31	78.5	77.64
RF	78.93	78.28	78.93	75.18	79.77	80.98	79.77	75.54	78.93	80.32	78.93	74.02
MNB	**82.6**	**81.74**	**82.6**	**81.56**	**81.33**	81.4	**81.33**	**81.36**	77.79	79.55	77.79	78.43
KNN	80.48	80.69	80.48	80.58	80.91	80.81	80.91	80.85	**79.63**	80.01	**79.63**	**79.8**
Linear SVM	76.8	79.51	76.8	69.6	75.25	81.45	75.25	65.8	74.82	81.22	74.82	64.86
RBF SVM	78.36	80.39	78.36	72.78	76.52	**82.18**	76.52	68.48	75.53	81.61	75.53	66.41
XGBoost	76.94	76.94	76.94	70.85	76.66	76.16	76.66	70.5	76.52	75.9	76.52	70.24

data was tagged as personal hate and 31.43% data was tagged as neutral.

VI. Conclusion & Future work

Online hate speech and cyberbullying have adverse effects on individual psychology as well as socio-economic stability. Several riots happened worldwide, even in Bangladesh, only for uncontrolled social media activity and offensive cyberspace. And also, the traumatic effect of this kind of toxic comments and behavior on victim's psychology is a burning question. We hope, alongside digital literacy, this analytical and predictive research will play a vital role in ensuring online safety for everyone.

We are willing to continue this research further. Our priority is enriching the dataset. We want to collect more data and annotate with the diverse sentiment. The second step will be developing a language model for this kind of noisy data. And finally, we are going to democratize and give a public release of datasets and models so that anyone willing to carry this research can contribute to ensuring a safer cyber world.

References

[1] All the latest cyber bullying statistics and what they mean in 2021. https://www.broadbandsearch.net/blog/cyber-bullying-statistics.

[2] The annual bullying survey 2017. https://www.ditchthelabel.org/wp-content/uploads/2017/07/The-Annual-Bullying-Survey-2017-1.pdf.

[3] Bltk: The bengali natural language processing toolkit. https://pypi.org/project/bltk/description. Accessed: 2021-08-14.

[4] Facebook community standards and guidelines. https://www.facebook.com/communitystandards/introduction. Accessed: 2021-09-04.

[5] Fatick stemmer. https://github.com/MIProtick/Bangla-stemmer. Accessed: 2021-08-14.

[6] Harikrishnan N B. Confusion matrix, accuracy, precision, recall, f1 score. https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd.

[7] Bin Chen, Dennis Zaebst, and Lynn Seel. A macro to calculate kappa statistics for categorizations by multiple raters. In *Proceeding of the 30th Annual SAS Users Group International Conference*, pages 155–30. Citeseer, 2005.

[8] Amit Kumar Das, Abdullah Al Asif, Anik Paul, and Md Nur Hossain. Bangla hate speech detection on social media using attention-based recurrent neural network. *Journal of Intelligent Systems*, 30(1):578–591, 2021.

[9] Amitava Das and Sivaji Bandyopadhyay. Phrase-level polarity identification for bangla. *Int. J. Comput. Linguist. Appl.(IJCLA)*, 1(1-2):169–182, 2010.

[10] Louis de Bruijn. Inter-annotator agreement (iaa). https://towardsdatascience.com/inter-annotator-agreement-2f46c6d37bf3.

[11] Brian Dean. Social network usage growth statistics: How many people use social media in 2021? https://backlinko.com/social-media-userssocial-media-usage-stats.

[12] Statista Research Department. Leading countries based on facebook audience size as of july 2021. https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/.

[13] Md Karim, Sumon Kanti Dey, Bharathi Raja Chakravarthi, et al. Deephateexplainer: Explainable hate speech detection in under-resourced bengali language. *arXiv preprint arXiv:2012.14353*, 2020.

[14] Simon Kemp. Digital 2020: Global digital overview. https://datareportal.com/reports/digital-2020-global-digital-overview.

[15] Md Serajus Salekin Khan, Sanjida Reza Rafa, Amit Kumar Das, et al. Sentiment analysis on bengali facebook comments to predict fan's emotions towards a celebrity. *Journal of Engineering Advancements*, pages 118–124, 2021.

[16] Binny Mathew, Navish Kumar, Pawan Goyal, Animesh Mukherjee, et al. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*, 2018.

[17] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Prajwal Singhania, Suman Kalyan Maity, Pawan Goyal, and Animesh Mukherjee. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380, 2019.

[18] Ted Mei. Demystify tf-idf in indexing and ranking. https://ted-mei.medium.com/demystify-tf-idf-in-indexing-and-ranking-5c3ae88c3fa0/.

[19] Muhammad Mahmudun Nabi, Md Tanzir Altaf, and Sabir Ismail. Detecting sentiment from bangla text using machine learning technique and feature analysis. *International Journal of Computer Applications*, 153(11):28–34, 2016.

[20] Web Robots. Instant data scraping extension. https://webrobots.io/instantdata/.

[21] Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 457–468. Springer, 2021.

[22] Sagor Sarker. Bnlp: Natural language processing toolkit for bengali language. *arXiv preprint arXiv:2102.00405*, 2021.

[23] Sm Taher, Kazi Akhter, and K. M. Hasan. N-gram based sentiment mining for bangla text using support vector machine. pages 1–5, 09 2018.

[24] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer, 2018.