Sentiment Analysis with Bangla Language Social Media Texts

Jannatul Ferdous†
Electrical and Computer Engineering Department
North South University
Dhaka , Bangladesh

Md. Arifur Rahman Akib†
Electrical and Computer Engineering Department
North South University
Dhaka , Bangladesh

Shamlima Deb Treena†
Electrical and Computer Engineering Department
North South University
Dhaka , Bangladesh

Md Rezwanul Alam Sayem†
Electrical and Computer Engineering Department
North South University
Dhaka , Bangladesh

† authors contributed equally

***Abstract*—Anyone can express themselves freely on social media. Nevertheless, people occasionally harm others by disregarding social norms and going beyond their own boundaries resulting in online harassment. Using strong information retrieval and data mining techniques, academics are battling to provide a safe online environment in the field of social media mining. We want to accomplish this objective for those who speak Bangla in this paper. In order to develop strong classifiers that can distinguish between comments with positive, negative, and neutral polarity, we gathered a dataset of 11006 Bangla comments from Facebook, examined them demographically, and annotated them. 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. We also implemented DL model CNN. Overall, the model achieves an accuracy of 81%**

Keywords—Sentiment Analysis, Multiclass classification, Bangla Language, Deep learning.

I. INTRODUCTION

As of January 2024, 5.35 billion individuals used social media globally, demonstrating how prevalent it has become. The goal of WhatsApp, YouTube, and Facebook is to connect people, but. It's difficult to keep things safe when there are so many users. Hate speech and cyberbullying are serious issues that impact millions of users worldwide. With 52.9 million Facebook users in Bangladesh, more than 60% people face cyberbullying. The problem is that while tools to prevent these behaviors are effective in languages like English, they are less effective in languages like Bangla.Simply said, there aren't enough tools available to verify and comprehend what people are saying online. It is difficult to prevent hate speech, misinformation, and cyberbullying in languages like Bangla without effective tools. It is devoid of the computational linguistic resources required to carry out different NLP tasks, such as corpora, language models, and potent machine learning techniques. What We Did: As a precursor to social media surveillance tools, we have created strong machine learning classifiers and a well annotated corpus to categorize online hate speech.

II. LITERATURE REVIEW

In this part, we review some related materials that were relevant to our research. For hate speech on Bangla comments, Khan, proposed an SVM based model where they used TF-IDF to process data and multiple classification models to yield the accuracy, proposed a model based on encoder-decoder. Comments were categorized into 7 distinct categories of speech. All contents were divided into hateful and non-hateful categories. To extract and encode features 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, 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 crowdsourc- ing where 10k is hate comments. Das Bandyopadhyay 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 used TF-IDF feature extraction on Bangla text to generate results. We further review Countering online hate speech analyzing the dataset. It compiled a dataset with 6,898 counterspeech comments and 7,026 non-counterspeech comments. The psycholinguistic effects of counterspeech and non-counterspeech. The targeted communities for the dataset were Jews, African-Americans and LGBT communities. • Sentiment Analysis on Bengali Facebook Comments To Predict Fan's Emotions Towards a Celebrity: For hate speech on Bangla comments, they proposed an SVM based model where they used TF-IDF to process data and multiple classification models to yield the accuracy. • Hate Speech detection in the Bengali language: A dataset and its baseline evaluation: They 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(pretrained) embedding such as FastText, Word2Vector, and BengFastText were used on this

dataset. They have used 30k comments tagged by crowdsourc- ing where 10k is hate comments. • Phrase-level Polarity Identification for Bangla: They used a machine learning model called SVM to classify sentences as positive or negative. The model considered parts of speech (POS), function words (e.g. prepositions, conjunctions), and negative words from a sentiment lexicon. The model achieved 70 percent precision in identifying sentiment. • Thou Shalt Not Hate- Countering Online Hate Speech: This research explores ways to identify comments that challenge online hate speech on YouTube. They trained different machine learning models to analyze comment features like word usage, sentence structure, and overall vocabulary. The most successful model combined an XGBoost classifier with features from sentence vectors, TFIDF, and bag-of-words, achieving 71.6 percent accuracy in recognizing counterspeech.

III. METHODOLOGY

A. Dataset

1) Dataset Description: Our dataset consists of the comments, along with the num- ber of reactions and the number of replies each comment has on Facebook public posts. Our dataset divides the comments into three main categories: positive, negative, and neutral. Each of these categories further includes its own subcategories, each with an assigned index number. It contains 11,006 comments from various public posts. The dataset consists of 47.6% negative comments, 17.3 %positive comments, and 35.1 % neutral comments.

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

Figure : Dataset Table 1

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

Figure : Dataset Table 2

2) Data Cleaning: For data cleaning process we used BNLP and BLTK 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. 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.

3. 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.

4. 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'.

B. Classification

After cleaning, preprocessing and vectorizing the data using TF-IDF, we created the train and test split. We used 90 percent of the data for our train feature vector and the other 10 percent for the test feature vector. We used LogisticRegression, De- cisionTreeClassifier, 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.

With start with cleaning the data and preprocess 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 classical machine learning algorithms to build classification models based on our data.

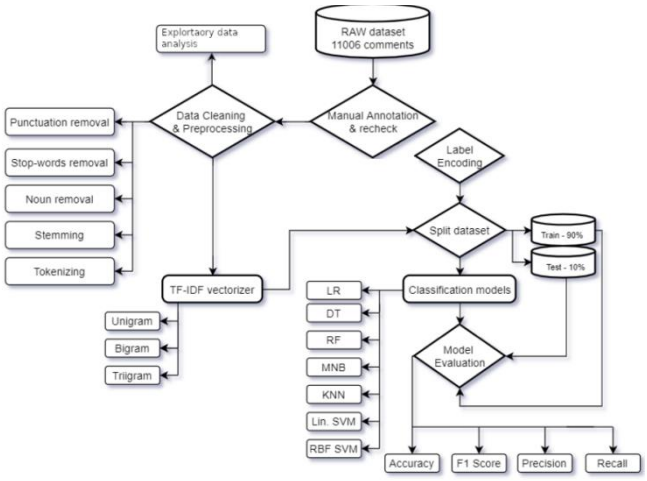

Figure : Process Flowchart

C. Package Description

1. NumPy: NumPy is a fundamental package for scientific computing with Python, providing support for arrays and matrices along with a collection of mathematical functions to operate on these data structures.

2. Pandas: Pandas offers data structures and operations for manipulating numerical tables and time series. It is a critical tool for data wrangling and preparation in this project.

3. Scikit-learn: This library is used for implementing machine learning algorithms and tools for data mining and data analysis.

4. Matplotlib and Seaborn: For data visualization, we employed Matplotlib and Seaborn. Matplotlib is a plot- ting library that allows for the creation of static, animated, and interactive visualizations, while Seaborn builds on Matplotlib to provide a high-level interface for drawing attractive and informative statistical graphics.

5. TensorFlow: TensorFlow is an open-source library for numerical computation and large-scale machine learn- ing. It provides a comprehensive, flexible ecosystem of tools, libraries, and community resources that lets researchers push the state-of-the-art in ML, and developers easily build and deploy ML-powered applications.

D. Model Explanation:

1. LogisticRegression(LR): Logistic Regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable. It is used to estimate the probability of a binary outcome.

2. Decision Trees (DT): Decision Trees are a non-parametric supervised learning method used for classifi- cation and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

3. Random Forest (RF): Random Forest is an ensemble learning method for classification and regression that operates by constructing multiple decision trees during training and outputting the mode of the classes (classifi- cation) or mean prediction (regression) of the individual trees.

4. Multinomial Naive Bayes (MNB): Multinomial Naive Bayes is a probabilistic learning method used for classi- fication. It is based on Bayes' Theorem with an assump- tion of independence among predictors. It is particularly suited for text classification problems.

5. K-Nearest Neighbors (KNN): KNN is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training exam- ples in the feature space.

6. Linear Support Vector Machines (Linear SVM): Linear SVM is a supervised learning model used for classification and regression analysis. It constructs a hyperplane or set of hyperplanes in a high-dimensional space, which can be used for classification, regression, or other tasks.

7. Radial Basis Function Support Vector Machines (RBF SVM): RBF SVM is a type of SVM that uses a radial basis function as the kernel. It is effective in high- dimensional spaces and when the number of dimensions exceeds the number of samples.

8. 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 files and stored it in Drive to conclude our evaluation.

IV. ANALYSIS

We used four measures to evaluate our classification models: accuracy, F1-score, precision and recall. Here is a brief explanation of these measures :

1. Accuracy: Accuracy is defined by the proportion between correctly predicted data (TP, TN) and total number of data (TP, TN, FP, FN).

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

2. 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 the 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.

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

3. 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.

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

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

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

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

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

Figure : Uniram, Bigram and Trigram Results Analysis

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, 79.63, 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 in the below figure

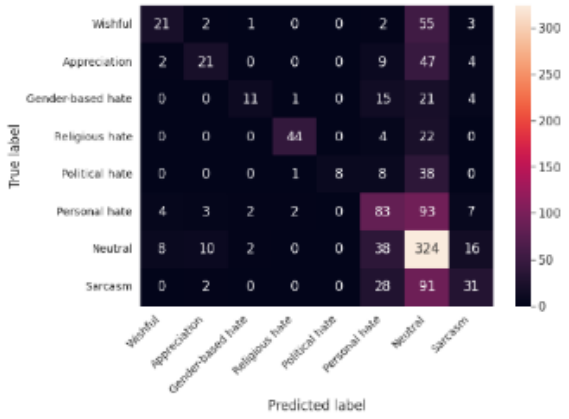

Figure : Confusion Matrix

Since we had the best accuracy with the MNB classifier with unigram features, we plotted the confusion matrix for that model. If we look at religious hate, for example, 62.85% data were predicted correctly. Amongst the misclassification, 5.72% of data was tagged as personal hate and 31.43% of data was tagged as neutral.

V. CONCLUSION

Cyberbullying and hate speech on the internet negatively impact both socioeconomic stability and personal psyche.

Uncontrolled social media activity and hostile cyberspace were the sole causes of multiple riots that occurred globally, including in Bangladesh. There is also a pressing concern for the psychological trauma caused by such harmful remarks and actions. Together with digital literacy, we anticipate that this analytical and predictive study will be essential to guaranteeing everyone's online safety.

Future Direction: We are open to carrying out additional study. Enhancing the dataset is our top focus. We intend to gather additional information and annotate it with a range of sentiments. Creating a language model for this type of noisy data will be the second stage.

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] Bnltk: The bengali natural language processing toolkit. https://pyropi/project/bnltk/description. Accessed: 2021-08-14.

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

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

[6] Harikrishna R N B. Confusion matrix, accuracy, precision, recall, f1 score. https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-aed299cf63d.

[7] Bin Chen, Dennis Zebastz, and Lynn Seal. A roadmap to calculate kappa statistics for categories by multiple raters. In Proceedings 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(2):169–182, 2010.

[10] Louis de Bruijn. Inter-annotator agreement (iaa). https://www.academicasecure.com/inter-annotator-agreement-24fc6637bf3.

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

[15] Md Serajus Salehin Khan, Sanjida Reza Rafi, 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 on twitter. arXiv preprint arXiv:1812.02176, 2018.

[17] Binny Mathew, Punyajoy Saha, Hardik Tharad, Subham Rajgaria, Pawan 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://tedmei.medium.com/demystify-tf-idf-in-indexing-and-ranking-5ca38ce8456c.

[19] Muhammad Mahmudul Nabi, Md Tanzir Altaf, and Sahir 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] Nursos Romin, Moshaed Ahmed, Hrienshwar 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:2102.04045, 2021.

[23] Sm Taher, Kazi Akhter, and K. M. Hasan. N-gram based sentiment mining for bangla text using support vector machine. pages 5–9, 2010.

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