# A multi-label classification analysis for detecting different categories of bullies on social media in Bengali language

Maisha Maliha
*Department of Computing*
*Boise State University*
Boise, USA
Email:maishamaliha@u.boisestate.edu

***Abstract*—Since social media and online discussion platforms are becoming more and more ubiquitous, the popularity of analyzing their content increases. These environments provide versatile ways for users to express their opinions and thoughts. Alongside informative and witty remarks, there are comments that are far from being innocuous.They have also shown harmful side effects that can have serious consequences.The word toxic has become synonymous with online hate speech, internet trolling, and sometimes outrage culture. In this study, we build an efficient model to detect and classify toxicity in social media from user-generated content using Machine Learning algorithm and deep learning approach.The purpose of this research is to provide an overview of the current status of the research and theoretical perspectives on cyberbullying in hopes of encouraging improved methodologies. .**

***Index Terms*—Bengali social media,sentiment analysis, multilabel classification, cyberbullying, deep learning**

## I. Introduction

Revolutions have been created around the world to use the Internet for the extension of knowledge, scientific activities, education, and socialism. During the last decade, Bangladesh has witnessed significant growth on the Internet Like many developed and developing countries. Bangladesh's internet penetration rate stood at 31.5 percent of the total population at the start of 2022. The number of social media users in Bangladesh at the start of 2022 was equivalent to 29.7 percent of the total population [17].People have been utilizing Facebook, Twitter, and other social networking websites to share their opinions on numerous topics throughout the last decade, frequently in their native language, as the use of social media has increased. Since the introduction of various easy-to-use Bangla keyboard apps in recent years, the use of Bangla on social media has also increased. With the rapid growth of user-generated content in social media, abusive, toxic, bully comments can be found in online posts, messages, and comments across languages.

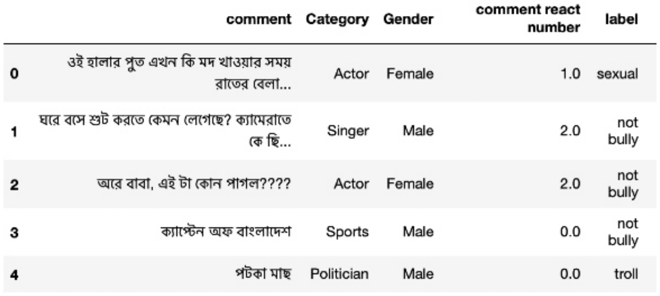
Bengali is the seventh most spoken language in the world and the fifth most widely used writing system in the world, until now, significant work hasn't addressed this issue [3]-[17]. Although a few papers tried to determine the offensive or hate speech in Bengali utilizing labeled data, few concentrated on this in a gender-specific way. Since social media platforms such as Facebook, Twitter, YouTube, and Instagram are popular in Bangladesh it is necessary to distinguish toxicity in the comments or reviews for various downstream tasks such as abusiveness, troll or hate speech detection, and understanding social behaviors. Furthermore, it is critical to investigate the relationship between toxicity and human sentiment.

## II. Background

Toxicity is frequently linked to abusive comments in the form of trolls, sexual, religious, thread, and other forms. It is required to identify the presence of toxicity on social media content [1].Moreover, liberal users tend to exercise vulgarity more on social media [1] [14].A text can be considered a threat or abusive if it contains sexist or racist slurs, attacks or criticizes any community or religious view, provokes criminal activities, etc. Bully victims are more likely to (between 2 and 9 times) commit suicide than non-victims.With the extensive growth of user interactions through prominent advances on the Web, sentiment analysis has obtained more focus from an academic and a commercial point of view. Recently, sentiment analysis in the Bangla language is progressively being considered an important task, for which previous approaches have attempted to detect the overall polarity of a Bangla document. Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text.The main reason that the Bangla Natural Language Processing sector is being held back, is the lack of sufficient data [16]

### *A. Related Work*

Nafis Irtiza Tripto and Mohammed Eunus Ali(2018) [8] build deep learning-based models to classify a Bangla sentence with a three-class (positive, negative, neutral) and a five-class (strongly positive, positive, neutral, negative, strongly negative) sentiment label. They also build models to extract the emotion of a Bangla sentence as anyone of the six basic emotions (anger, disgust, fear, joy, sadness and surprise).



| | comment | Category | Gender | comment react number | label |
|---|---|---|---|---|---|
| 0 | ওই হালার পুত এখন কি মদ খাওয়ার সময় রাতের বেলা... | Actor | Female | 1.0 | sexual |
| 1 | ঘরে বসে শুট করতে কেমন লেগেছে? ক্যামেরাতে কে ছি... | Singer | Male | 2.0 | not bully |
| 2 | আরে বাবা, এই টা কোন পাগল???? | Actor | Female | 2.0 | not bully |
| 3 | ক্যাপ্টেন অফ বাংলাদেশ | Sports | Male | 0.0 | not bully |
| 4 | পটকা মাছ | Politician | Male | 0.0 | troll |

Fig. 1. Raw Data

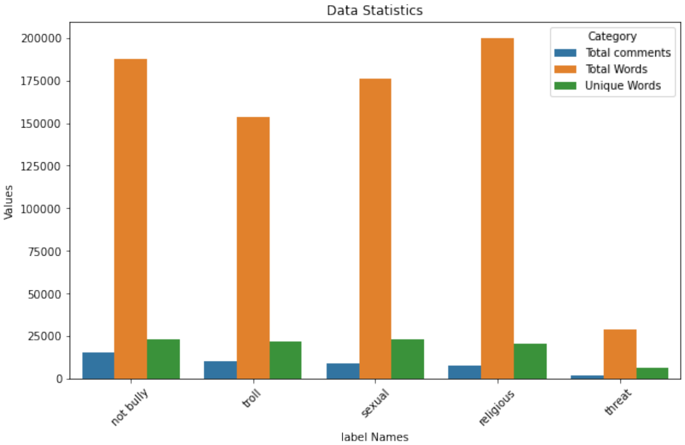


Fig. 2. Statistics of the corpus

Davidson et al. (2017) [17] studied how hate speech is different from other instances of offensive language. They used a crowd-sourced lexicon of hate language to collect tweets containing hate speech keywords. Using crowd-sourcing, they labeled tweets into three categories: those containing hate speech, only offensive language, and those with neither. We train a multi-class classifier to distinguish between these different categories. They analyzed when hate speech can be reliably separate from other offensive language and when this differentiation is very challenging. Plaban Kr. Bhowmick, Anupam Basu, Pabitra Mitra & Abhishek Prasad (2009) [18] presented a novel method for classifying news sentences into multiple emotion categories using the MultiLabel K Nearest Neighbor classification technique.Their emotion data consists of 1305 news sentences and the emotion classes considered are disgust, fear, happiness and sadness.Experiments have been performed on feature comparison and feature selection.

Another system, proposed by Shahnoor Chowdhury Eshan and Mohammad Shahidul Hasan(2017) [19]is based on Bengali Unicode text implemented both count vector and TF-IDF for unigram, bi-gram and tri-gram. Then MNB and SVM (linear and RBF) classifiers were used. This system only dealt with Unicode Bengali words while we considered both words and emoticons. Besides, consecutive exclamation marks and question marks are considered single words in our system.

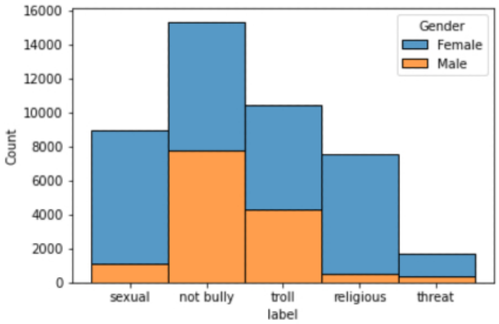


Fig. 3. Number of comment corresponds to gender of celebrity

## III. Methodology

### A. Data Collection

A labeled data set that has been collected from github.This data set has been labeled and annotated from the remark association area under public posts by celebrities, government officials and athletes on the Facebook stage will be used. The total amount of collected comments is 44,001 [23]. To eliminate noise from data and convert it to a certain input format, we use various pre-processing and word embedding algorithms on texts. The architecture of our model is then provided.Figure1 shows a portion of raw data.The dataset is a great source for developing the ability of machines to differentiate whether a word is a bully expression or not with the help of Natural Language Processing and to what extent it is improper if it is an inappropriate comment. The comments are labeled with different category bullies with the help of experts and consensus.Figure 3 shows female celebrities are getting more sexual and thread comment comparing to male celebrities. Due to the scarcity of data collection of categorized Bengali language comments, this dataset can have a significant role for research in detecting bully words, identifying inappropriate comments, detecting different categories of Bengali bullies, etc.

### B. Pre-processing

The comments obtained from Facebook posts is noisy and often contain errors,unnecessary information and duplication. Although a lot of pre-processing steps are present in rule-based sentiment analyzer, not all these steps are required in our approach. We tokenize each sentence and remove stopwords from them. We used a Bangla stopwords removal technique for this research also Bangla tokenizer [21]. Elongated words,punctuation marks and emoticons often contain sentiment information for multi class categorization. However, we remove links,URLs,user tags, mentions of special character and emojis from comments.

### C. Word Embedding

For vector representation and feature extraction, Tf-IdfVectorizer and Count Vectorizer were used . The CountVectorizer creates a vocabulary of the words in the corpus and

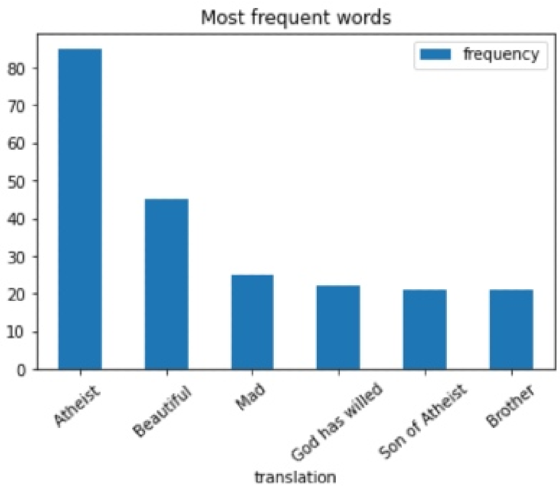
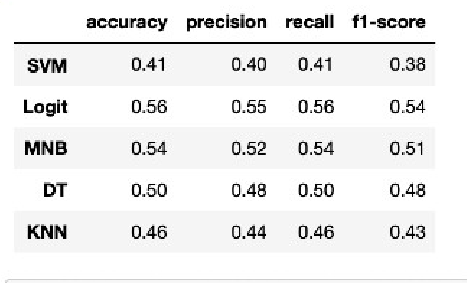


Fig. 4. Most frequent words translated to English



| | accuracy | precision | recall | f1-score |
|---|---|---|---|---|
| **SVM** | 0.41 | 0.40 | 0.41 | 0.38 |
| **Logit** | 0.56 | 0.55 | 0.56 | 0.54 |
| **MNB** | 0.54 | 0.52 | 0.54 | 0.51 |
| **DT** | 0.50 | 0.48 | 0.50 | 0.48 |
| **KNN** | 0.46 | 0.44 | 0.46 | 0.43 |

Fig. 5. ML Algorithms Accuracy report

counts the frequencies of the words. The TfidfVectorizer function from scikit-learn is used for Training machines Learning models.we do not apply lemmatization but Stemming has been applied to text contents to facilitate feature extraction.We used Bangla stemmer for this.Text-based features such as TF-IDF and gender-specific information related to the online activity and connectivity network were separated for later usage.

### *D. Baseline Evaluation:*

To evaluate a machine learning model that ends up with an accuracy number or other metric, we need to know if it is meaningful. Particularly in imbalanced classification models, one way of evaluating it is comparing with the baseline value. For a machine learning algorithm to demonstrate that it has skill on a problem, it must achieve accuracy better than this Baseline value. This is a multiclass problem we choose to go with a theoretical baseline strategy.As our dataset comments has five label with a ratio of sexual 0.20,not bully 0.348,troll 0.237,religious 0.172,threat 0.038 so the theoretical baseline is

$$0.2^2 + 0.35^2 + 0.23^2 + 0.17^2 0.038^2 = 0.249$$

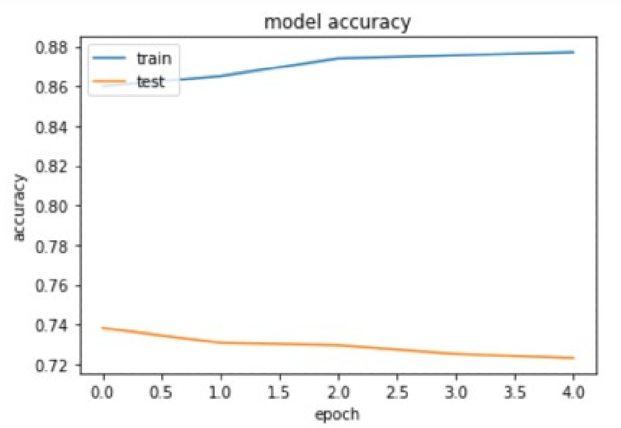


Fig. 6. Accuracy over epoch

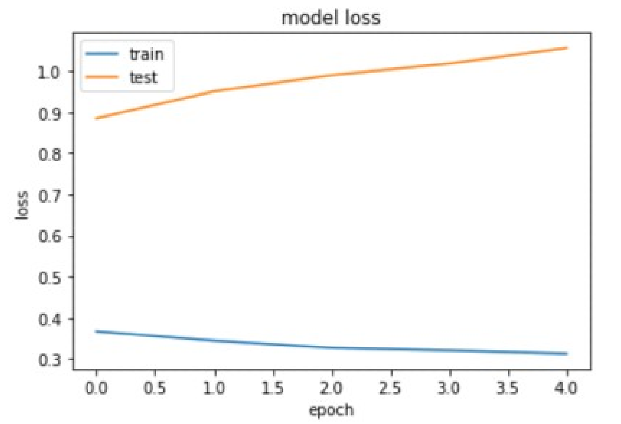


Fig. 7. Loss over epoch

## IV. Model Architecture

An exhaustive exploration has been performed to identify a suitable machine learning algorithm for Bangla text categorization. Available machine learning models from the literature have been considered (i.e. NB, LogisticRegression ,SVM, Decision Tree and KNN).Deep learning based methods like Feed-Forward Neural Network were implemented. The workflow of the model construction process is illustrated.

### *A. ML Algorithms:*

In the training phase, extracted feature sets were trained by the popular supervised machine learning algorithms. Because this was a multi-label classification problem, we trained our models by setting up multi-label output. We used linear SVC in the support vector machine (SVM) implementation. The following machine learning algorithms were used:

- Support Vector Machine (SVM)
- logistic regression
- K-nearest neighbor (KNN)
- Decision tree
- Multinomial Naive Bayes

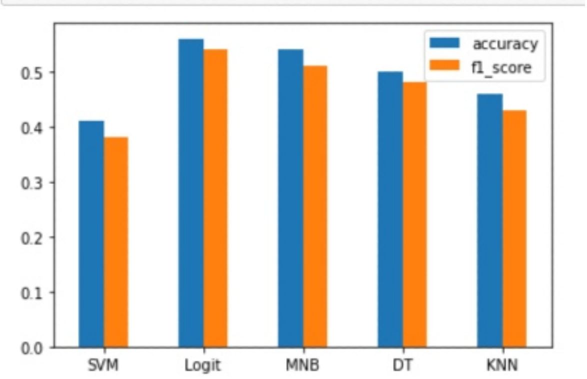


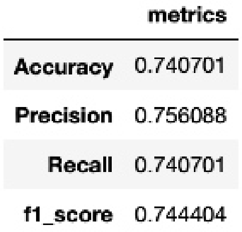
Fig. 8. Accuracy and F1 Score comparison between ML classifier

### B. Deep learning,NN:

**DL** is a subset of multiple layers of neurons that perceives from nonlinear neural network with matrix representations and converts the output at one level into an intense and abstract peak. Feed forward Neural Networks, also known as Deep feed forward Networks or Multi-layer Perceptrons, are the focus of this article.The data enters the input nodes, travels through the hidden layers, and eventually exits the output nodes. The network is devoid of links that would allow the information exiting the output node to be sent back into the network.the intermediate layer, which is concealed between the input and output layers. This layer has a large number of neurons that perform alterations on the inputs. They then communicate with the output layer.A drop-out rate of 0.5 is applied to the dropout layer; ReLU activation is used in the intermediate layers. In the final layer, softmax activation is applied. As an optimization function, Adam optimizer, and as a loss function, Categorical-cross entropy are utilized.

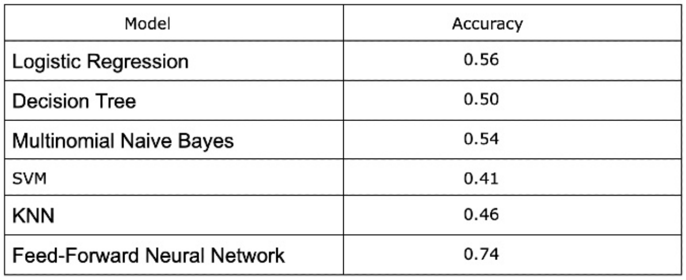
## V. Performance Evaluation

ML algorithm models performed better with TF-IDF vectorizer and achieved quite a similar accuracy as Figure 5 shows.The machine learning added least 15% to our theoretical baseline.That's significant. More importantly, our accuracy is also well above the theoretical baselines value.Figure 8 Overall logistic Regression classifier performed among the classifier. We applied the NN with Keras API and the model is Sequential,we set out loss to categorical cross entropy and the optimizer is adam.After several experiments with activation function,test-train split ,batch,epoch size we figure out as Fig 3 shows the accuracy increases for training set as epoch number increases but loss in validation set increase also. Therefore,the problem of overfitting prevails in our approach.Figure 7 and Figure 6 displaying details. As result, we find the epoch number to five and batch size 32 showing the best performance with an accuracy of 74% at the test data.NN added 49% to our theoretical baseline which is significant for imbalanced classification problems.Figure 10 shows details.



| | metrics |
|---|---|
| **Accuracy** | 0.740701 |
| **Precision** | 0.756088 |
| **Recall** | 0.740701 |
| **f1_score** | 0.744404 |

Fig. 9. Accuracy report of Feed Forward Neural Network



| Model | Accuracy |
|---|---|
| Logistic Regression | 0.56 |
| Decision Tree | 0.50 |
| Multinomial Naive Bayes | 0.54 |
| SVM | 0.41 |
| KNN | 0.46 |
| Feed-Forward Neural Network | 0.74 |

Fig. 10. comparison between all models

## VI. Conclusion

The growing severity level of cyberbullying in Bangladesh calls for serious consideration to our educators, researchers, administrators and authorities.Figure3 shows female celebrities are getting more sexual and thread comment comparing to male celebrities which indicates the attitude of social media user toward women.Researchers are trying to develop sophisticated automated support systems for a long time. Though a notable amount of work has been performed for cyber bullying detection on English text, very few work have been done on Bangla text. In this paper revisited four state-ofthe-art supervised machine learning algorithms and feed-forward neural network.Our research shows that deep learning based model performed well.Our models for multilabel sentiment achieved at least 15% more accuracy the theoretical baseline. The analysis reveals the strengths and weaknesses of varied approaches and provides the directions for future research.

## References

[1] Sazzed S. Identifying vulgarity in Bengali social media textual content. PeerJ Computer Science. 2021 Oct 19;7:e665.

[2] Lashkarashvili N, Tsintsadze M. Toxicity detection in online Georgian discussions. International Journal of Information Management Data Insights. 2022 Apr 1;2(1):100062.

[3] Akhter S. Social media bullying detection using machine learning on Bangla text. In2018 10th International Conference on Electrical and Computer Engineering (ICECE) 2018 Dec 20 (pp. 385-388). IEEE.

[4] Ahmed MT, Rahman M, Nur S, Islam A, Das D. Deployment of machine learning and deep learning algorithms in detecting cyberbullying in bangla and romanized bangla text: A comparative study. In2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) 2021 Feb 19 (pp. 1-10). IEEE.

[5] Hossen S. Nature and Aftermath of Cyberbullying with Female University Students in Bangladesh.

[6] Hossain MR, Hoque MM, Siddique N, Sarker IH. Bengali text document categorization based on very deep convolution neural network. Expert Systems with Applications. 2021 Dec 1;184:115394.

[7] Mohammed HH, Dogdu E, Görür AK, Choupani R. Multi-Label Classification of Text Documents Using Deep Learning. In2020 IEEE International Conference on Big Data (Big Data) 2020 Dec 10 (pp. 4681-4689). IEEE.

[8] Tripto NI, Ali ME. Detecting multilabel sentiment and emotions from bangla youtube comments. In2018 International Conference on Bangla Speech and Language Processing (ICBSLP) 2018 Sep 21 (pp. 1-6). IEEE.

[9] Sharfuddin AA, Tihami MN, Islam MS. A deep recurrent neural network with bilstm model for sentiment classification. In2018 International conference on Bangla speech and language processing (ICBSLP) 2018 Sep 21 (pp. 1-4). IEEE.

[10] S. Chowdhury and W. Chowdhury, "Performing sentiment anal-ysis in bangla microblog posts," in Informatics, Electronics Vision (ICIEV), 2014 International Conference on. IEEE,2014, pp. 1–6

[11] M. S. Islam, M. A. Islam, M. A. Hossain, and J. J. Dey,"Supervised approach of sentimentality extraction from bengalifacebook status," in Computer and Information Technology(ICCIT), 2016 19th International Conference on. IEEE, 2016,pp. 383–387

[12] A. K. Paul and P. C. Shill, "Sentiment mining from bangla datausing mutual information," in Electrical, Computer Telecom-munication Engineering (ICECTE), International Conferenceon. IEEE, 2016, pp. 1–4.

[13] Romim, N., Ahmed, M., Talukder, H., Islam, M. S. (2020). Hate Speech Detection in the Bengali language: A dataset and its baseline evaluation. arXiv preprint arXiv:2012.09686

[14] Cachola I, Holgate E, Preoţiuc-Pietro D, Li JJ. Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. InProceedings of the 27th International Conference on Computational Linguistics 2018 Aug (pp. 2927-2938).

[15] Preoţiuc-Pietro D, Liu Y, Hopkins D, Ungar L. Beyond binary labels: political ideology prediction of twitter users. InProceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 2017 Jul (pp. 729-740)

[16] Chowdhury RR, Hossain MS, Hossain S, Andersson K. Analyzing sentiment of movie reviews in bangla by applying machine learning techniques. In2019 International Conference on Bangla Speech and Language Processing (ICBSLP) 2019 Sep 27 (pp. 1-6). IEEE.

[17] Waseem Z, Davidson T, Warmsley D, Weber I. Understanding abuse: A typology of abusive language detection subtasks. arXiv preprint arXiv:1705.09899. 2017 May 28
https://github.com/cypher-07/Bangla-Text-Dataset

[18] Bhowmick PK, Basu A, Mitra P, Prasad A. Multi-label text classification approach for sentence level news emotion analysis. InInternational conference on pattern recognition and machine intelligence 2009 Dec 16 (pp. 261-266). Springer, Berlin, Heidelberg.

[19] Chakraborty P, Seddiqui MH. Threat and abusive language detection on social media in bengali language. In2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) 2019 May 3 (pp. 1-6). IEEE.

[20] https://datareportal.com/reports/digital-2022-bangladesh

[21] https://bnlp.readthedocs.io/en/latest/

[22] https://pglsinc.com/languages/bengali

[23] https://github.com/cypher-07/Bangla-Text-Dataset