

# Bengali Slang detection using state-of-the-art supervised models from a given text

Md. Abdul Hamid, Eteka Sultana Tumpa, Johora Akter Polin, Jabir Al Nahian, Atiqur Rahman, Nurjahan Akther Mim

Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh

## Article Info

### Article history:

Received Sep 9, 2022

Revised Dec 16, 2022

Accepted Jan 20, 2023

### Keywords:

Bengali Slang

Classifier

Machine learning

Social media

Supervised models

## ABSTRACT

Almost all Bengalis who own smartphones also have social media accounts. People from different regions occasionally employ regional Slang that is unfamiliar to outsiders and confuses the meaning of the sentence. Nearly all languages can now be translated thanks to modern technology, but only in very basic ways, which is a concern. Bengali Slang terms are difficult to translate due to a dearth of rich corpora and frequently occurring new Slang terms developed by people, making it impossible for speakers of other languages to understand the context of a sentence in which Slang is used. We developed a solution to this issue. To create models that can detect Bengali Slang terms from social media, we gather various Slang phrases from various regions and develop a modest corpus. Our suggested method nearly always succeeds in extracting Bengali Slang terms from fresh material. We create a total of 7 supervised models and assess which is the most effective for our study. One of them has a 70% accuracy and 86% recall rate for successful identification. Our models may be linked to the social media platform's backend to restrict the use of Bengali Slang in posts, blogs, comments, and other areas.

*This is an open access article under the [CC BY-SA](#) license.*



## Corresponding Author:

Md. Abdul Hamid

Department of Computer Science and Engineering, Daffodil International University

Abdullahpur, South Keranigonj, Dhaka, Bangladesh

Email: ahmadhamid040@gmail.com

## 1. INTRODUCTION

Language is the most prevalent means for people to communicate with one another. Slang is a local language spoken by a certain group of individuals [1]. Every language has its own set of Slang phrases. Slang is difficult to describe but it is a term that people can easily understand within a group [2]. People of a younger age use Slang terminology with their friends and on social media. Slang words are often utilized in movies and songs. People increasingly utilize a variety of social media channels to express themselves in their native language. In comparison to other languages, English has already had enough research done to distinguish Slang phrases. Because these languages lack adequate study to distinguish Slang words, users in other languages can easily publish Slang phrases on social media [3]. For this limitation, one can easily use any kind of Slang words in the media which is not a good thing for society. Our younger generations are used to utilizing social media platforms, and they commonly use Slang terminology in their posts, texts, and messages. Teenagers are immature, and they lack the understanding necessary to apply proper context on social media. On social media and other platforms, Slang, jargon, and hostile context will have a long-term unfavorable impact [4]. Detecting informal words on social media is another significant task since it frequently leads to language abuse. It is necessary to create a model in which Slang terms are easily identifiable while keeping in mind that no one is allowed to use unnecessary

Slang, offensive, or jargon phrases improperly way. Due to the limited collection of Bengali Slang words, there is not enough study done on Slang terms in the Bengali language. Slang phrases are now commonly used on social networking sites. According to 2022, Bengali is the 5<sup>th</sup> worldwide spoken language [5]. These huge-spoken people use Slang words continuously on social media if it continues to happen then it will negatively affect both language and society. Due to the limited size of the Bengali Slang word dataset, it's very hard to build a proper model which will detect Slang words from social media.

There have been enormous amounts of research work going on in natural language processing (NLP) [6]. Some authors have previously worked on Slang detection and their work has greatly helped research society. Pal and Saha [4] developed a system that can detect jargon words from e-text with the help of semi-supervised learning. From a text file, every word picks and checks whether it is a jargon word or not. Some drawbacks that should be fixed here such as they considered some words that used in the medical fields and judiciary are as jargon, it should be exceptional here. In another paper, Haq *et al.* [3] detect Slang or jargon words from social media in Urdu. To implement their work, firstly they convert their data into UCS transformation format-8-bit (UTF-8) encoding and then apply their algorithm to find results. Though they successfully find out their accuracy up to 72.6% and 55.21% consequently offensive and non-abusive, their dataset only carries 1,200 words which are very few words. Another study was done to explain the meaning of the Chinese language words Slang automatically by the machine [7]. For this particular purpose, dual character-level encoding was used using the seDCEAnn model, a focus-based neural network. To reduce human comment bias [8] a dual-class dislike speech (HS) dataset (HS-BAN) in Bengali contains more than 50,000 tagged comments, of which 40.17% dislike speech and the rest are non-dislike speech. In their dataset, some Slang words have also been included and Benchmark has achieved an 86.78% F1 score using the Bi-LSTM model on top of short text informal word embedding. By analyzing the paper of Emon *et al.* [9], it can be seen that they use different types of machine learning and deep learning-based algorithm in their paper to detect abusive Bengali text. Here the accuracy of the deep learning-based algorithm can be observed best. By analyzing the paper of Hossain *et al.* [10] about "discovering political Slang in readers' comments", it can be seen that they have used machine learning approaches. Specially they used an unsupervised algorithm for detecting creative Slang. They also developed poliSlang, an algorithm used to extract creative Slang words from data sets. Asghar *et al.* [11] primary contribution is a system for recognizing and scoring internet Slang (DSIS) utilizing SentiWordNet and other lexical resources. The results of the comparison reveal that the suggested system outperforms the existing approaches. They suggest that their approach may be utilized to create an opinion lexicon for Slang terms. Alhumoud and Wazrah [12] collect data from Facebook comments and use data mining techniques. They also provide an algorithm that can detect cyberbullying in a remark. A fair 77% accuracy in recognizing one of the following cyberbullying categories: sexual, physical sexual, religious, political, appearance, racism, cultural, psychological, adversely praying for a person, and general cyberbullying. The support vector machine classifier produced the greatest results, while the adaptive boosting technique earned the highest precision rate of 94%. Jiamthapthaksin *et al.* [13] proposed a method for extracting popular Thai Slang by comparing social media posts and using tokenization, as well as a dictionary-based method for extracting unknown words, before expanding it by using the n-gram method to determine what is currently trending and popular Slang words. The rest of the paper is as follows: in section two methodology section discussed details, the result and discussion are discussed in section three and finally, in section four we conclude our research.

## 2. METHOD

The method section is divided into three sections, each showing how we achieved the specific aim. 4 subsections are as follows: a description of the data, preprocessing of data, a brief discussion about the classifiers, and the model evaluation. Figure 1 shows the working procedure of our work.

### 2.1. Description of the data

A fresh Slang phrase from a certain group or culture appears every day. However, humans don't write it on a piece of paper or in a database. It's well-liked in some exclusive areas and on social media. This is why it's hard to translate the Bengali Slang word collect. The internet, social media, private groups, literature, and many different Slang communities were some of the places where these Slang terminologies were gathered. We have a binary class in our research, Slang, and non-Slang. We maintain proportions when we collect for each class. We collect around 8,110 data and store it in our corpus. The Slang class has 4,056 and the non-Slang class has 4,054 data.

### 2.2. Data preprocessing

Preprocessing is an important activity in text mining, NLP, and information retrieval (IR) [14]. After collecting raw data, data annotation is a must case as we apply the supervised model. Supervised training

necessitates a massive quantity of labeled data [15]. A model gives the worst outcome when we give messy data to a model. To find out a better result from a model, clean data is necessary. Sometimes some unnecessary character takes place in a sentence or a phrase that is unnecessary often. Such as “[”, “/”, “@”, “[”, “)”, and so on. These characters are meaningless in this case, so they must be clean from every sentence or phrase. Table 1 demonstrates the removing unnecessary character.

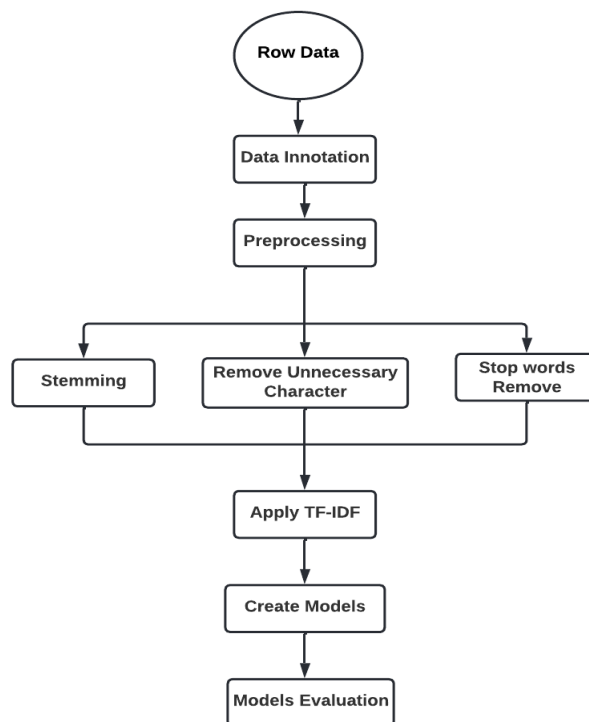


Figure 1. Working procedure of our work

Table 1. Removing unnecessary character

Original text	Removing unnecessary character
তুমি কে?	তুমি কে

Stopwords, often known as noise words, are words that transmit just a little quantity of information that is usually unnecessary [16]. They must be removed before the model is trained. Some of the Bengali stop words are “অই”, “অতএব”, “অবধি”, and “হলে”. We develop a stop words corpus and when there are stop words found in the dataset it removes those stop words. As stop words have no meaning in the context, Table 2 represents removing stop words.

Table 2. Removing stop words

Text	After removing stops words
অতএব এখন	এখন

Stemming means getting the root word of a word. It is highly effective while we work on the NLP task because of model learns better by getting the root words. Some of the stemming rules are being applied to get the root word. In the concluding state of preprocessing, sentences or words are string-type data that can't read a model. To overcome this problem, some techniques are available and one of them is the TF-IDF vectorizer. Here, the model will carry each word as a unique value by assigning each word some weights depending on frequency. By dividing a term's frequency in a given text by the proportion of papers in which it appears, the TF-IDF calculates values for each word in the document. The document in which a word appears is strongly associated with words with high TF-IDF values [17]. The IF-IDF vectorizer formulas are shown in (1):

$$TF = \frac{\text{Number of times the term appears in a document}}{\text{Total number of words in the document}} \tag{1}$$

$$IDF = \log \left( \frac{\text{Number of the documents in the corpus}}{\text{Number of the documents in the corpus contain the term}} \right) \quad (2)$$

$$TF - IDF = TF * IDF \quad (3)$$

We divided our dataset into two sections, using 80% of the data for training and the remaining 20% for testing.

### 2.3. Discussion of classifiers

Support vector machine (SVM) is an algorithmic implementation of principles from statistical learning theory [18] that addresses the challenge of constructing consistent approximation from data. The k-nearest neighbor technique differs in that it uses the data directly for classification rather than first developing a model [19], [20]. A tree-like structure is produced by the decision tree method, which repeatedly divides the data set into subsets using a criterion that optimizes data separation [21], [22].

### 2.4. Model evaluation

The confusion matrix is generally used to evaluate the model performance of multi-class or binary-class classification models [23]. We evaluate our models through different evaluation metrics. They are precision, recall, and F1 score. Model accuracy is also given for the model evaluation.

## 3. RESULTS AND DISCUSSION

This work is mainly a binary classification problem. Hence, it produced a 2×2 confusion matrix. To evaluate the 7 classifier models, different performance metrics uses such as model accuracy, precision, recall, and F1-score. Below is the result of the models.

### 3.1. Result of models

Table 3 summarizes the results from the models and shows performance metrics measurement. Overall, our proposed models perform substantially well. Logistic regression gives the best model accuracy compared to others and XG-Boost gives a slightly lower percentage compared to others. On the other hand, the best Precision can be found in the XG-Boost Slang class which is 0.90, and the lowest in also the same model with 0.56 non-Slang class. By observing 5 models such as logistic regression, KNN, random forest, decision tree, and SVM they have a lower difference between the two classes in terms of precision, recall, and F1-score and that's why they have high model accuracy and low variance. The best model accuracy gives 70% with 0.72 precision, 0.86 recall, and 0.69 F1-score.

Table 3. Classification report and accuracy

Classification algorithms	Class	Precision	Recall	F1-score	Average precision	Accuracy (%)
Naive Bayes	Slang	0.59	0.86	0.70	0.67	64
	Non-Slang	0.75	0.41	0.53		
Logistic Regression	Slang	0.79	0.53	0.63	0.72	70
	Non-Slang	0.65	0.86	0.74		
KNN	Slang	0.79	0.44	0.57	0.74	66
	Non-Slang	0.61	0.89	0.73		
Random Forest	Slang	0.82	0.49	0.61	0.73	69
	Non-Slang	0.64	0.89	0.74		
Decision Tree	Slang	0.79	0.51	0.62	0.71	69
	Non-Slang	0.64	0.86	0.73		
SVM	Slang	0.78	0.52	0.63	0.71	69
	Non-Slang	0.64	0.86	0.74		
XG-Boost	Slang	0.90	0.23	0.37	0.73	60
	Non-Slang	0.56	0.98	0.71		

### 3.2. Discussion on models performance

The receiver operating characteristic (ROC) curve was used to assess the performance of the classification method. ROC curve is the best way to show a classifier's performance to select a suitable model from different models [24]. A ROC curve is a plot that gives an abstract of the performance of a binary classification model on the positive class. Figure 2 shows the ROC curve of our 7 classifiers model. Here, the x-axis indicates the false positive rate and the y-axis indicates the true positive rate. From The ROC curve, it is seen that the yellow line is the highest peak line compared to others. The yellow line is the logistics regression model line and its model accuracy is highest compared to others. On the other side, XG-Boost is the lowest positive case and lowest curve compared to the other 6 models. Every model in our case gives a higher true

positive rate, which is good for a model and thus we conclude that every model is balanced in this case and the model learns better from our dataset and gives a much higher model performance. Figure 2, describes the ROC curve of the 7 classifiers model.

The area under the ROC curve (AUC) is a very used performance indicator for classification [25]. ROC AUC is the measurement of whether a model is balanced or imbalanced against a given dataset. When the positive case is more compared to the negative case, we called it a balanced model. The more positive case better the model is. The negative case with a greater number of examples and a positive case with a minority of examples is the imbalanced model. ROC AUC value stands between 0.0 to 1.0. More the percentage better the model is and at the same time learning rate is also good. Every model has a decent value number which stands perfect machine learning model in the NLP case. Figure 3, demonstrate the classification report.

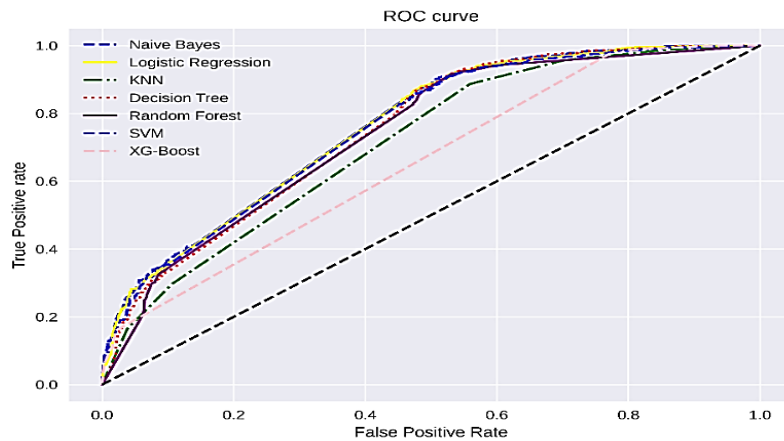


Figure 2. ROC curve of 7 classifiers model

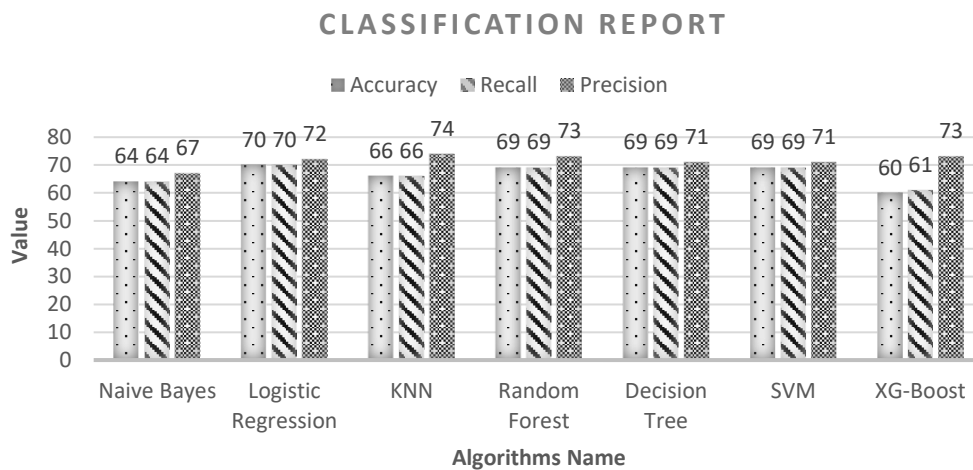


Figure 3. Comparison of accuracy, recall, and precision of different classifiers

#### 4. CONCLUSION

Everyone knows that Slang is a colloquial language that can have a detrimental impact on a community. To readily recognize Bengali Slang terms, a proper model is required. Since here, classification-based research has been conducted, with the model being able to detect Slang phrases automatically. The provided model can recognize Slang terms with 70% accuracy and 72% precision, according to the analysis of the results. Even though most Slang phrases are short, it is observed that extended Slang terms influence the model. Bert and XL-net are two models that can be used in future work. People are being alerted to an issue that is nothing more than a Bengali internet Slang dictionary. A perfect model for Bengali Slang detection must be created, with additional Slang terms being added to the dictionary so that individuals may contribute new Slang words with suitable meanings. Every day, new Slang terms are used, making it difficult to compile a list





of them all. A significant amount of Slang phrases may be gathered, and then effective algorithms like deep learning models and transformer-based models like Bert or XL-net can be used.

## REFERENCES

- [1] Z. Pei, Z. Sun, and Y. Xu, "Slang Detection and Identification," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 2019, doi: 10.18653/v1/k19-1082.
- [2] M. Hafiza and R. N. Rosa, "An Analysis of Word Formation of English Slang Used in Straight Outta Compton Movie," *English Language and Literature*, vol. 9, no. 1, p. 38, Feb. 2020, doi: 10.24036/ell.v9i1.108045.
- [3] N. U. Haq *et al.*, "USAD: An Intelligent System for Slang and Abusive Text Detection in PERSO-Arabic-Scripted Urdu," *Complexity*, vol. 2020, pp. 1–7, Nov. 2020, doi: 10.1155/2020/6684995.
- [4] A. R. Pal and D. Saha, "Detection of Slang Words in e-Data using semi-Supervised Learning," *International Journal of Artificial Intelligence and Applications*, vol. 4, no. 5, pp. 49–61, Sep. 2013, doi: 10.5121/ijaia.2013.4504.
- [5] K. K. Podder *et al.*, "Bangla Sign Language (BdSL) Alphabets and Numerals Classification Using a Deep Learning Model," *Sensors*, vol. 22, no. 2, p. 574, Jan. 2022, doi: 10.3390/s22020574.
- [6] B. Alarie, A. Niblett, and A. H. Yoon, "How artificial intelligence will affect the practice of law," *University of Toronto Law Journal*, vol. 68, pp. 106–124, Jan. 2018, doi: 10.3138/utlj.2017-0052.
- [7] C. Yi, D. Wang, C. He, and Y. Sha, "Learning to Explain Chinese Slang Words," in *Artificial Neural Networks and Machine Learning–ICANN 2019: Text and Time Series*, Springer International Publishing, 2019, pp. 22–33, doi: 10.1007/978-3-030-30490-4\_3.
- [8] N. Romim, M. Ahmed, M. S. Islam, A. S. Sharma, H. Talukder, and M. R. Amin, "HS-BAN: A Benchmark Dataset of Social Media Comments for Hate Speech Detection in Bangla," *arXiv*, Dec. 03, 2021, doi: 10.48550/arXiv.2112.01902.
- [9] E. A. Emon, S. Rahman, J. Banarjee, A. K. Das, and T. Mitra, "A Deep Learning Approach to Detect Abusive Bengali Text," in *2019 7th International Conference on Smart Computing and Communications (ICSCC)*, Jun. 2019, doi: 10.1109/icsc.2019.8843606.
- [10] N. Hossain, T. T. T. Tran, and H. Kautz, "Discovering Political Slang in Readers' Comments," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 12, no. 1, Jun. 2018, doi: 10.1609/icwsm.v12i1.15074.
- [11] Dr. M. Asghar, F. M. Kundi, S. Ahmad, and A. Khan, "Detection and Scoring of Internet Slangs for Sentiment Analysis Using SentiWordNet," *Life Science Journal*, vol. 11, no. 9, pp. 66–72, May 2014, doi: 10.6084/M9.FIGSHARE.1609621.
- [12] S. O. Alhumoud and A. A. A. Wazrah, "Arabic sentiment analysis using recurrent neural networks: a review," *Artificial Intelligence Review*, vol. 55, no. 1, pp. 707–748, Apr. 2021, doi: 10.1007/s10462-021-09989-9.
- [13] R. Jiamthaphaksin, P. Setthawong, and N. Ratanasawetwad, "A system for popular Thai Slang extraction from social media content with n-gram based tokenization," in *2016 8th International Conference on Knowledge and Smart Technology (KST)*, Feb. 2016, doi: 10.1109/kst.2016.7440478.
- [14] R. Rex, "Preprocessing Techniques for Text Mining", Accessed: Jan. 01, 2023. [Online]. Available: [https://www.academia.edu/35015140/Preprocessing\\_Techniques\\_for\\_Text\\_Mining](https://www.academia.edu/35015140/Preprocessing_Techniques_for_Text_Mining).
- [15] Y. Chali and S. A. Hasan, "Query-focused multi-document summarization: automatic data annotations and supervised learning approaches," *Natural Language Engineering*, vol. 18, no. 1, pp. 109–145, Apr. 2011, doi: 10.1017/s1351324911000167.
- [16] J. Kaur and P. K. Buttar, "A Systematic Review on Stopword Removal Algorithms," *International Journal on Future Revolution in Computer Science and Communication Engineering*, vol. 4, no. 4, Apr. 2018.
- [17] J. E. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," 2003.
- [18] S. Dreiseitl and L. O-Machado, "Logistic regression and artificial neural network classification models: a methodology review," *Journal of Biomedical Informatics*, vol. 35, no. 5–6, pp. 352–359, Oct. 2002, doi: 10.1016/s1532-0464(03)00034-0.
- [19] B. V. Dasarathy, *Nearest Neighbor (NN) Norms: Nn Pattern Classification Techniques*. IEEE Computer Society Press, 1991. Accessed: Jan. 01, 2023. [Online]. Available: <https://cir.nii.ac.jp/crid/1572261550010307072>.
- [20] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996, doi: 10.1017/cbo9780511812651.
- [21] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, "Regression Trees," in *Classification And Regression Trees*, Routledge, 2017, pp. 216–265, doi: 10.1201/9781315139470-8.
- [22] S. L. Salzberg, "C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993," *Machine Learning*, vol. 16, no. 3, pp. 235–240, Sep. 1994, doi: 10.1007/bf00993309.
- [23] S. Visa, B. Ramsay, A. Ralescu, and E. Knaap, "Confusion Matrix-based Feature Selection.," USA, Jan. 2011, vol. 710, pp. 120–127.
- [24] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997, doi: 10.1016/s0031-3203(96)00142-2.
- [25] D. J. Hand, "Measuring classifier performance: a coherent alternative to the area under the ROC curve," *Machine Learning*, vol. 77, no. 1, pp. 103–123, Jun. 2009, doi: 10.1007/s10994-009-5119-5.




## BIOGRAPHIES OF AUTHORS






**Md. Abdul Hamid**     is a student of BSc in Computer Science and Engineering at the Daffodil International University in Bangladesh. In a few days, he will receive the Bachelor's degree in Computer Science and Engineering from this university. His current research interest lies in NLP, machine learning and deep learning. He is an active researcher and involve in various sectors such as DIU NLP and ML lab and DIU CPC research wing. He can be contacted at email: [abdul15-12387@diu.edu.bd](mailto:abdul15-12387@diu.edu.bd).








**Eteka Sultana Tumpa**    is a student in the department of computer science and engineering (CSE) of Daffodil International University (DIU), Bangladesh. Her research interest is NLP and ML. She is currently working on different areas of research. She can be contacted at email: eteka15-12121@diu.edu.bd.






**Johora Akter Polin**    is a lecturer for the Computer Sciences and Engineering Department at the Daffodil International University in Bangladesh, and also working as a judge in Social Business Creation Competition which is conducted by HEC Montreal, Canada. She holds a bachelor's degree in computer sciences and engineering. Being passionate about research, social business, networking, and data analytics, Johora is especially interested in machine learning, deep learning, human computer interaction, and social media analysis. She is also actively involved in the community as a public speaker and a mentor to support and encourage young researchers. She can be contacted at email: polin.cse@diu.edu.bd.






**Jabir Al Nahian**    received the Bachelor's degree in Computer Science and Engineering from Daffodil International University, Bangladesh. He is currently Research Scientist in the Computational Intelligence Lab, Bangladesh specializing in providing machine learning solutions for expertise. His current research interests lie in the area of data science, natural language processing, machine learning, deep learning and particularly in areas pertaining to their application for the Bangla language. He is an active researcher and reviewer at several international conferences and journals. He can be contacted at email: jabir15-10414@diu.edu.bd and jabirnahian009@gmail.com.



**Atiqur Rahman**    is a Computer Science and Engineering student at Daffodil International University in Bangladesh. Being enthusiastic about networking, research, and co-curricular activities. Machine learning, image processing, deep learning, and human-computer interaction are areas in which Atiqur is particularly interested. He can be contacted at email: atiqur15-13604@diu.edu.bd.



**Nurjahan Akther Mim**    is a student at the Daffodil International University in Bangladesh studying computer science and engineering being interested in research, networking, and extracurricular activities. Nurjahan is interested in machine learning, image processing, deep learning, and human-computer interaction. She can be contacted at email: mim15-4782@diu.edu.bd.