Transformer-based Bengali Textual Emotion Recognition

Md. Atabuzzaman
Virginia Tech
Virginia, USA
atabuzzaman@vt.edu

Mst Maksuda Bilkis Baby
HSTU, Dinajpur
Bangladesh
maksuda.cse34@gmail.com

Md Shajalal
HSTU & Fraunhofer FIT
University of Siegen, Germany
shajalal@hstu.ac.bd

Abstract

Emotion recognition for high-resource languages has progressed significantly. However, resource-constrained languages such as Bengali have not advanced notably due to the lack of large benchmark datasets. Besides this, the need for more Bengali language processing tools makes the emotion recognition task more challenging and complicated. Therefore, we developed the largest dataset in this paper, consisting of almost 12k Bengali texts with six basic emotions. Then, we conducted experiments on our dataset to establish the baseline performance applying machine learning, deep learning, and transformer-based models as emotion classifiers. The experimental results demonstrate that the models achieved promising performance in Bengali emotion recognition.

Keywords: Bengali emotion; Benchmark Dataset; Deep Learning; Transformer model;

1 Introduction

Since learning emotions through automated systems was a decades-long objective (Picard, 2000), recognizing textual emotion is crucial for assessing the enormous volumes of user data on various virtual platforms. The generated texts contain different user emotions expressed in their own manner, making the emotion classification task more challenging (Das et al., 2021c). However, Ekman (1993) defined six primary emotions (joy, sadness, surprise, disgust, anger and fear) can be extracted from texts with scientific techniques and tools (Das et al., 2021a; Alswaidan and Menai, 2020) and recognizing them has major usages for online businesses and system enhancements such as removing unwanted content (Demszky et al., 2020) from the system.

The identification of emotions in high-resource languages such as English, Chinese, and others has advanced considerably (Alswaidan and Menai, 2020; Demszky et al., 2020; Plaza-del Arco et al., 2020) due to the availability of advanced computational processes and tools. However, emotion detection has not progressed significantly in resource-constrained languages such as Bengali due to the lack of large benchmark datasets and language processing tools (Das et al., 2021b; Zhang et al., 2021). There has been some research on emotion identification using machine learning (ML) and deep learning (DL) approaches. Tripto and Ali (2018) suggested an LSTM-based multi-label emotion recognition system for YouTube comments in Bengali and English. Another study by Azmin and Dhar (2019) identified joyful, sad, and angry emotions in Bengali texts. They applied the term frequency-inverse document frequency (TF-IDF) and multinomial naive Bayes for classification. Ruposh and Hoque (2019) used 1.2k texts to develop a framework for recognizing six raw emotions. They have proposed a method using bag-of-words (BoW) with support vector machine (SVM). Pal and Karn (2020) presented a method for identifying happiness, anger, sadness, and suspense in Bengali short stories using TF-IDF with logistic regression (LR). Das et al. (2021b) has proposed a Bengali emotion recognition dataset of 6.3k texts containing Ekman (1993) defined basic six emotions (anger, fear, disgust, sadness, joy, and surprise). Authors have applied transformer-based models m-BERT, Bangla-BERT (Sarker, 2021) and XLM-R to accomplish the task. These researchers utilized a small amount of data due to the scarcity of large datasets.

To address the lack of a large benchmark

dataset, we have constructed a benchmark Bengali emotions dataset comprised of almost 12k (11927) texts of six distinct emotions (joy, sadness, surprise, disgust, anger, and fear). We also conducted experiments to determine the performance of ML (Logistic Regression (LR), Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM) and ensemble method), DL (LSTM, BiLSTM, CNN-BiLSTM) and transformer-based (Bangla-BERT, mBERT, ELECTRA Bangla-BERT, XLM-R) methods on our dataset and to set the baseline of this dataset. As baseline methods, LR, SVM, and Bangla-BERT showed more promising performance on the developed dataset than other approaches. The contributions of this study are as follows:

- We developed the largest Bengali emotion classification dataset consisting of almost 12k texts of six distinct emotions (joy, sadness, surprise, disgust, anger, and fear).
- We provided the baseline performances of different ML, DL, and transformer-based approaches on the developed dataset.

The rest of the paper is organized as follows: Section 2 details some related works of Bengali emotions classification. Then, we describe our proposed method for Bengali emotion classification in Section 3. The experimental results and the details of the developed dataset are presented in Section 4. Finally, Section 5 discusses the concluding statements of this paper.

2 Related Works

Researchers have proposed different approaches to emotion identification tasks using emotional texts for various languages, including English, Chinese, and Arabic (Liu et al., 2019). Hybrid and learning-based approaches that utilize traditional text representation with distributed word representation outperformed some existing approaches on benchmark corpora (Alswaidan and Menai, 2020). Graph convolution network architecture also showed good performance (Lai et al., 2020). Recently, the transformer-based model (i.e., BERT) has gained popularity among researchers in analyzing emotional texts. A new method based on ELECTRA and a hybrid neural network can avoid the inconsistency of the mask training and fine-tuning process of the traditional pre-training model proposed by Zhang et al. (2022). Huang et al. (2019) proposed a novel Hierarchical LSTMs for Contextual Emotion Detection (HRLCE) model. They combined the results generated by BERT and HRCLE to achieve the overall performance.

Although emotion identification tasks on high-resource languages have attracted many researchers for further investigation, low-resource languages like Bengali are still preliminary in research progress. Few researchers proposed different methods employing ML and DL approaches. Zamsuri et al. (2023) proposed K-nearest neighbour-based method on Indonesian texts and showed 58% accuracy for the six-label classification and 79% accuracy for the two-label classification. Azmin and Dhar (2019) identified multi-class emotions from Bangla texts using Multinomial Naïve Bayes (NB) classifier along with various features such as stemmer, parts-of-speech (POS) tagger, n-grams, tf-idf. They used three emotion classes (happy, sad, and angry) and reported an overall accuracy of 78.6%. DL models classified Bangla sentences with three classes (positive, negative, neutral) and five (strongly positive, positive, neutral, negative, strongly negative) sentiment labels with 65.97% and 54.24% accuracy, respectively (Tripto and Ali, 2018). Ruposh and Hoque (2019) presented an emotion recognition technique by developing a corpus consisting of 1200 emotive words that are used to train the SVM classifier with 73% accuracy which is higher than the Naive based approach (60%). However, our proposed dataset comprises almost 12k Bangali emotional texts.

3 Methodology

This section discusses the strategies we explored for developing baseline models for identifying Bengali emotions on our developed dataset. We collected emotional texts from different community websites like Facebook, YouTube, and News portal comments. Then, we applied different ML, DL, and Transformer-based models to predict the type of emotional text.

3.1 ML Approaches

We did experiments on the emotion dataset by applying different ML models. The **LR**, **RF**, **NB**, and **SVM** techniques are applied using TF-IDF text vectorizer. Finally, we employed the ensemble method of these ML models with a majority voting technique to classify Bengali emotions.

3.2 DNN Approaches

We employed Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), Bidirectional LSTM (BiLSTM) (Hochreiter and Schmidhuber, 1997), and a combined model of Convolutional Neural Network (CNN) and BiLSTM to investigate the performance of emotion identification tasks. To obtain the semantic representation of the emotional text data, we used the embedding layer of keras[1], which provides a 100-dimensional sequence matrix for each word. After that, the sequence matrices are fed to the DL model's layer. To train the models, we employed 'categorical_crossentropy' as a loss function, *'Adam'* optimizer with a learning rate of *0.001* and batch size of 128, and *30* epochs with early stopping when the validation accuracy did not improve for 7 epochs. We also employed batch normalization and 'l2' regularizer to reduce the overfitting.

3.3 Transformer-based method

The transformer-based BERT model has achieved state-of-the-art performance in almost all text classification tasks (i.e., sentiment analysis (Chen et al., 2021; Das et al., 2021b)). Inspired by their methods, we have utilized four transformer-based pre-trained models: Bangla-BERT (Sarker, 2020), mBERT (Devlin et al., 2018), E-Bangla-BERT (ELECTRA) (Bhattacharjee et al., 2021), and XLM-R (Liu et al., 2019). These models are pre-trained using corpora of Bengali texts and many other languages. We have fine-tuned the parameters of the transformer models before applying them to our dataset. We used a learning rate of *1e-5* with *CrossEntropyLoss* as a loss function, batch size of 16, and *AdamW* as optimizer with PyTorch framework.

[1]https://keras.io/api/layers/core_layers/embedding/

4 Experiments and Results

This section details the dataset collection and annotation techniques along with the experimental results and analysis.

4.1 Data Collection and Annotation

To address the absence of a normative Bengali emotion dataset, we developed a dataset containing texts of six basic distinct emotions (joy, sadness, surprise, disgust, anger, and fear). Three undergraduate students interested in natural language processing research worked on data preparation by collecting texts from Facebook, YouTube, and online News Portal comments over 6 months. Initially, they collected 12635 comments from Facebook (6451), YouTube (3897), and News portal (2287). Then, each of them annotated all the collected texts (joy (0): 3076; sadness (1): 2546; surprise (2): 1299; disgust (3): 2267; anger (4): 2183; fear (5): 1264), and the annotations were cross-checked to eliminate multi-label inconsistency by the two experts (one has research experience of five years with Bangla language processing (BLP) and other has two years). The experts removed 247 joy (neutral emotion annotated as joy by the annotators), 23 disgust (confusing disgust or anger), and 37 anger (confusing disgust or anger) texts from the dataset because of inconsistency in labels. The processed dataset contains 12328 texts with their corresponding label and some samples of texts with their labels. Then, the dataset went through the prepossessing phase, where non-Bengali words and texts of less than four words and greater than 300 words were removed. To ensure the quality of the annotation, we measured the inter-annotator agreement using the coding reliability (Krippendorff, 2011) and Cohen's kappa (Cohen, 1960) scores. We found an inter-coder reliability of 89.3% with Cohen's Kappa score of 0.88. These scores indicate the quality of the corpus. Table 1 illustrates the statistics of the prepossessed dataset (11927 texts). We also assessed the Jaccard similarity score among different emotional texts using each emotion's 200 most frequent words. Table 2 shows that the high-

Classes	Labels	Instances	Total words	Unique words	Words/Instance
Joy	0	2654	50035	12691	18.85
Sadness	1	2477	56130	13718	22.66
Surprise	2	2229	27218	9045	21.50
Disgust	3	2047	48109	10548	21.58
Anger	4	1266	37321	10957	18.23
Fear	5	1254	26215	8092	20.91
Total:	6	11927	245028	65091	aver.:20.54

Table 1: Statistical measures of our benchmark dataset.

est Jaccard similarity score is obtained for disgust and anger texts. Then, the second highest is for disgust and sadness classes. This highest Jaccard similarity score means that there are more common words between these classes. On the other hand, there is the lowest number of common words between joy and fear classes. Additionally, some texts with their corresponding labels are demonstrated in the Appendix (table 4).

	joy	sad	sur.	dis.	ang.	fear
joy	1.0	0.54	0.51	0.53	0.49	0.47
sad		1.0	0.55	**0.61**	0.58	0.56
sur.			1.0	0.58	0.54	0.58
dis.				1.0	**0.66**	0.54
ang.					1.0	0.51
fear						1.0

Table 2: Jaccard similarity score among the emotional texts.

model	precision	recall	f1-score	acc.
LR	0.54	0.54	0.54	0.54
RF	0.52	0.50	0.49	0.50
NB	0.59	0.50	0.47	0.50
SVM	0.57	0.54	0.54	0.54
Ensemble	0.56	0.54	0.54	0.54
LSTM	0.48	0.48	0.48	0.48
BiLSTM	0.49	0.48	0.48	0.48
CNN-BiLSTM	0.50	0.46	0.46	0.46
Bangla-BERT	0.54	0.54	0.54	0.54
mBERT	0.51	0.51	0.49	0.51
E-Bangla-BERT	0.36	0.38	0.35	0.38
XLM-R	0.46	0.44	0.44	0.44

Table 3: Our proposed baseline performance with ML, DL, and Transformer models on the proposed dataset.

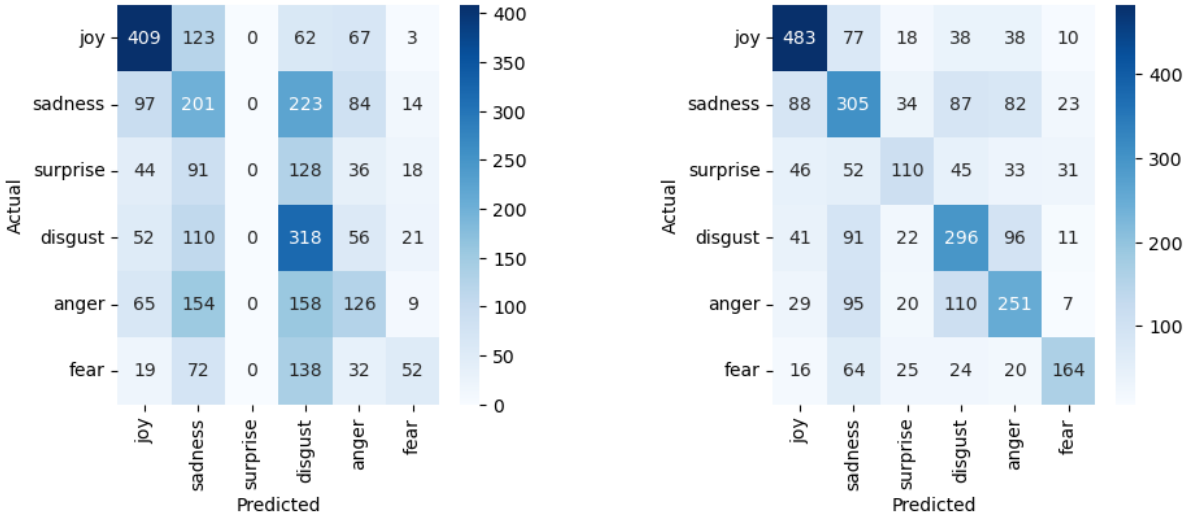
4.2 Experimental Results and Analysis

We evaluated the performance of Bengali textual emotion classification methods using four metrics: precision, recall, f1-score, and accuracy (acc.). Table 3 presents the attained scores of different evaluation metrics of the employed models in various settings.

Both LR and SVM achieved the highest score (accuracy: 0.54) among ML models (RF: 0.50 and NB: 0.50) in terms of all the evaluation matrices. Moreover, SVM outperforms LR by 3% higher score in precision (0.57). The ensemble of the ML models also performed well, with an accuracy of 0.54. However, we found that ML models face challenges in classifying the surprise emotional texts. The ensemble model reported the lowest recall score of 0.24 to detect surprise emotion and could correctly classify only 82 among 344 texts (table 5).

Among DL approaches, CNN-LSTM reported the lowest accuracy of 0.46, and both LSTM and BiLSTM achieved an accuracy of 0.48. Though we expect better performance from DL approaches than ML approaches, ML outperformed DL by achieving 6-8% higher accuracy. Again, like ML approaches, DL models also found it more challenging to classify surprise texts than other types of emotional sentences. Table 5 in the Appendix detailed the class-by-class performances of ML, DL, and transformer-based approaches.

Bangla-BERT performs well in classifying



(a) Confusion matrix of E-Bangla-BERT

(b) confusion matrix of Bangla-BERT

Figure 1: Confusion matrices of E-Bangla-BERT and Bangla-BERT.

six types of emotional texts and achieves the highest accuracy of 0.54 among the transformer-based approaches (mBERT: 0.51, E-Bangla-BERT: 0.38, and XLM-R: 0.44). On the other hand, E-Bangla-BERT obtained the lowest accuracy of 0.38. From the confusion matrix 1a, we can see that E-Bangla-BERT fails to detect any surprise emotional texts and has not predicted any other emotional sentences as a surprise. It classified the surprise emotion mostly as sadness (91) and disgust (128). In Bengali, many words are being used multi-purposely (Das et al., 2021a), which might be a reason for E-Bangla-BERT's failure to predict surprise emotion texts. Besides this, table 2 shows a moderate Jaccard similarity score (0.54-0.58) for surprise emotional text, which might be another possible reason for E-Bangla-BERT's poor performance. As the highest Jaccard similarity scores are obtained for sadness & disgust and disgust & anger emotions, confusion matrix 1a depicts that the highest miss-classification rate is obtained for both of these classes by E-Bangla-BERT. However, Bangla-BERT overcomes these limitations and shows promising performance in detecting all six types of emotions. Thus, developing a sizeable precious dataset with more diverse sentences might reduce the incorrect prediction of different approaches.

5 Conclusion

This paper investigates the performance of different ML, DL, and Transformer-based methods to classify the emotion of resource-constrained Bengali language. Due to the scarcity of a large dataset, Bengali emotion classification research has not achieved remarkable progress compared to other high-resourced languages. Therefore, we developed the largest Bengali emotion recognition dataset comprising almost 12k texts containing six basic human emotions. Then, we employed different ML, DL, and transformer-based approaches to set the baseline performances for the developed dataset. The experimental results showed that the ML and Transformer-based methods better classify emotions than DL models. Moreover, the results also indicate the challenges in identifying emotions in low-resource languages like Bengali. Therefore, in the future, we would like to increase the size of this dataset and develop a large corpus for different resource-constrained languages with more emotions for low-resourced languages.

6 Limitations

The performances of ML, DL, and Transformer-based methods to classify Bengali emotion differ from the standard compared to high-resourced languages like English. The possible reason might be that

the sentences need to be better structured, and there might be noise in the dataset as the data are collected from social media platforms and online news portals.

References

Nourah Alswaidan and Mohamed El Bachir Menai. 2020. A survey of state-of-the-art approaches for emotion recognition in text. *Knowledge and Information Systems*, 62(8):2937–2987.

Sara Azmin and Kingshuk Dhar. 2019. Emotion detection from bangla text corpus using naive bayes classifier. In *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, pages 1–5. IEEE.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Kazi Samin, Md Saiful Islam, Anindya Iqbal, M Sohel Rahman, and Rifat Shahriyar. 2021. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. *arXiv preprint arXiv:2101.00204*.

Ben Chen, Bin Chen, Dehong Gao, Qijin Chen, Chengfu Huo, Xiaonan Meng, Weijun Ren, and Yang Zhou. 2021. Transformer-based language model fine-tuning methods for covid-19 fake news detection. In *International Workshop on Combating On line Ho st ile Posts in Regional Languages dur ing Emerge ncy Si tuation*, pages 83–92. Springer.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H. Sarker. 2021a. Emotion classification in a resource constrained language using transformer-based approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 150–158, Online. Association for Computational Linguistics.

Avishek Das, Omar Sharif, Mohammed Moshiul Hoque, and Iqbal H Sarker. 2021b. Emotion classification in a resource constrained language using transformer-based approach. *arXiv preprint arXiv:2104.08613*.

Rajarshi Das, Manzil Zaheer, Dung Thai, Ameya Godbole, Ethan Perez, Jay Yoon Lee, Lizhen Tan, Lazaros Polymenakos, and Andrew McCallum. 2021c. Case-based reasoning for natural language queries over knowledge bases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9594–9611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *arXiv preprint arXiv:2005.00547*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paul Ekman. 1993. Facial expression and emotion. *American psychologist*, 48(4):384.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Chenyang Huang, Amine Trabelsi, and Osmar R Zaïane. 2019. Ana at semeval-2019 task 3: Contextual emotion detection in conversations through hierarchical lstms and bert. *arXiv preprint arXiv:1904.00132*.

Klaus Krippendorff. 2011. Agreement and information in the reliability of coding. *Communication methods and measures*, 5(2):93–112.

Yuni Lai, Linfeng Zhang, Donghong Han, Rui Zhou, and Guoren Wang. 2020. Fine-grained emotion classification of chinese microblogs based on graph convolution networks. *World Wide Web*, 23:2771–2787.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Aditya Pal and Bhaskar Karn. 2020. Anubhuti–an annotated dataset for emotional analysis of bengali short stories. *arXiv preprint arXiv:2010.03065*.

Rosalind W Picard. 2000. *Affective computing*. MIT press.

Flor Miriam Plaza-del Arco, Carlo Strapparava, L Alfonso Urena Lopez, and M Teresa Martín-Valdivia. 2020. Emoevent: A multilingual emotion corpus based on different events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498.

Hasan Abid Ruposh and Mohammed Moshiul Hoque. 2019. A computational approach of recognizing emotion from bengali texts. In *2019 5th International Conference on Advances in Electrical Engineering (ICAEE)*, pages 570–574. IEEE.

Sagor Sarker. 2020. Banglabert: Bengali mask language model for bengali language understading. *textsIGitHub*.

Sagor Sarker. 2021. Banglabert: Bengali mask language model for bengali language understanding.

Nafis Irtiza Tripto and Mohammed Eunus Ali. 2018. Detecting multilabel sentiment and emotions from bangla youtube comments. In *2018 International Conference on Bangla Speech and Language Processing (ICBSLP)*, pages 1–6. IEEE.

Ahmad Zamsuri, Sarjon Defit, and Gunadi Widi Nurcahyo. 2023. Classification of multiple emotions in indonesian text using the k-nearest neighbor method. *Journal of Applied Engineering and Technological Science (JAETS)*, 4(2):1012–1021.

Chen Zhang, Yu Xie, Hang Bai, Bin Yu, Weihong Li, and Yuan Gao. 2021. A survey on federated learning. *Knowledge-Based Systems*, 216:106775.

Shunxiang Zhang, Hongbin Yu, and Guangli Zhu. 2022. An emotional classification method of chinese short comment text based on electra. *Connection Science*, 34(1):254–273.

Appendix

Table 4: Some texts with corresponding labels from the introduced dataset.

Bengali Texts with English translation	Emotion
বিশ্ববিদ্যালয়ে বৃষ্টিতে ভিজে অনুষ্ঠান দেখার মধ্যেও একটা ফিল আছে! (There is a feeling in watching the event in the rain in the university!)	joy (0)
ইসলাম শব্দের অর্থ শান্তি কেন তারা ইসলামের নামে বিশৃঙ্খলা সৃষ্টি করে (The word Islam means peace why they create chaos in the name of Islam)	sadness (1)
আমি সত্যিই আশ্চর্যজনক মনে করি যে লোকেরা কীভাবে কারও সাথে বা বিপক্ষে এত দৃঢ়ভাবে নিজেদের সারিবদ্ধ করে (I find it really amazing how people align themselves so strongly with or against anyone)	surprise (2)
আপনার লেখাটা কপি পেস্ট মনে হয় (Your writing seems like copy paste)	disgust (3)
আমরা অসভ্য ইউএনওর দৃষ্টান্তমূলক শাস্তি চাই। (We want exemplary punishment of the lewd UNO.)	anger (4)
বাংলাদেশের খেলা দেখে আমরা অধিকাংশ বাঙালিরা হাটা টাক করার মতো অবস্থা হয়ে যায়।আর আর্জেন্টিনার মানুষ তো মনে হচ্ছে নগদে মারা যাবে (Most of us Bengali's get goosebumps after watching Bangladesh's game. And the people of Argentina seem to die of it.)	fear (5)

Table 5: Performance of Ensemble, LSTM, E-Bangla-BERT, and Bangla-BERT on the proposed dataset.

Model	class	precision	recall	f1-score	support
Ensemble	0	0.61	0.76	0.68	682
	1	0.43	0.61	0.51	606
	2	0.53	0.24	0.33	344
	3	0.53	0.54	0.54	527
	4	0.54	0.43	0.48	503
	5	0.78	0.47	0.59	320
	weighted avg	0.56	0.54	0.54	2982
LSTM	0	0.57	0.62	0.59	682
	1	0.40	0.45	0.42	630
	2	0.38	0.28	0.33	299
	3	0.55	0.50	0.52	542
	4	0.51	0.38	0.43	511
	5	0.40	0.56	0.47	318
	weighted avg	0.48	0.48	0.48	2982
E-Bangla-BERT	0	0.57	0.64	0.60	664
	1	0.27	0.42	0.33	619
	2	0.00	0.00	0.00	317
	3	0.31	0.45	0.37	557
	4	0.34	0.18	0.24	512
	5	0.61	0.32	0.42	313
	weighted avg	0.36	0.38	0.35	2982
Bangla-BERT	0	0.69	0.73	0.71	664
	1	0.45	0.49	0.47	619
	2	0.48	0.35	0.40	317
	3	0.49	0.53	0.51	557
	4	0.48	0.49	0.49	512
	5	0.67	0.52	0.59	313
	weighted avg	0.54	0.54	0.54	2982