Bangla Emotion Detection Dataset With An Extended Taxonomy And Its Evaluation

Md Jahangir Alam[1], Md Shohanur Rahman Shohan[2], Ismail Hossain[1], Sai Puppala[3], and Sajedul Talukder[4]

[1] University of Texas at El Paso TX, USA
{malam10,ihossain}@miners.utep.edu
[2] Islamic University, Bangladesh
shohanur.rahman.iu@gmail.com
[3] Southern Illinois University Carbondale IL, USA
saimaniteja.puppala@siu.edu
[4] University of Texas at El Paso TX, USA
stalukder@utep.edu

Abstract. The rising interest in emotion detection in language is driven by the abundance of emotional expressions on Web 2.0 platforms. This paper explores the challenges in developing an automatic emotion detection system for Bengali, given the limited resources and lack of standard corpora. We describe the development of a comprehensive emotional dataset containing 20,247 texts categorized into 27 different emotional categories. Our process involved data collection, preprocessing, human and automatic labeling, and label verification. The dataset's evaluation, reflected by a Cohen's score of 0.89, indicates a high level of annotator agreement. Experiments conducted with machine learning, deep learning, and BERT-based models identified XLM-R as the best-performing model, achieving an F1 score of 0.79 and an accuracy of 0.82 in emotion classification tasks.

Keywords: social network · emotion detection · Bangla language · stylometric features · word vector.

1 Introduction

Serving as the primary language for 98% of the Bangladeshi populace, Bangla is formally recognized as the national language of Bangladesh [4]. Its influence is not restricted to national boundaries, with significant Bangla diaspora in regions such as the Middle East, Europe, and the USA [2]. Amidst the emergence of a Digital Bangladesh initiative [18], Bangla's digital imprint on platforms like Facebook, LinkedIn, and Twitter has grown. Facebook, emerging as Bangladesh's dominant social platform, claims 33.71 million active Bengali users [19]. Engaging in commerce, communication, and more via Facebook groups and Messenger, the urgency to address the proliferation of inauthentic Facebook profiles is evident. This manuscript delves into emotion detection from Bangla text on social

media. With exhaustive research in languages like English, Russian, and Arabic, the dearth of work in Bangla underscores the significance of our endeavor.

Facebook's textual content, inherently conversational, often conveys pivotal insights and viewpoints. Automating emotion detection from such content necessitates grasping both writing style and underlying context, going beyond rudimentary rule-based approaches. An author's writing style mirrors their psychological, sociological, and even physiological states. Stylometric features (SF) delineate unique aspects of this style, encompassing elements such as word frequency, sentence length, or special character utilization.

Emotion detection in language has recently gained significant attention from NLP researchers due to the widespread availability of expressions, opinions, and emotions shared through comments on Web 2.0 platforms. Developing an automatic emotion detection system in Bengali presents particular challenges due to resource scarcity and the lack of standard corpora. Therefore, creating a standard dataset is essential for analyzing emotional expressions in Bengali texts. This paper details the development process of a comprehensive emotional dataset, including data collection, pre-processing, labeling, and verification. The dataset comprises 20,247 texts, categorized into 27 emotional categories: admiration, amusement, anger, annoyance, approval, caring, confusion, curiosity, desire, disappointment, disapproval, disgust, embarrassment, excitement, fear, gratitude, grief, joy, love, nervousness, optimism, pride, realization, relief, remorse, sadness, and neutral. We perform human annotation with the help of a group of 5 graduate students who work on NLP. The evaluation of the dataset, with a Cohen's score of 0.89, indicates strong agreement among annotators. Additionally, we conducted experiments on emotion classification using various models, including machine learning (ML), deep learning (DL), and BERT-based models. Our experiments demonstrate that mBERT performs best on our dataset, achieving an F1 score of 0.78 and an accuracy of 0.81.

In this manuscript, we introduce a novel emotion dataset with 27 emotion categories and experiment with machine learning techniques tailored emotion detection derived from textual data. Leveraging a dataset user posts collated from various Facebook groups, pages and individual profiles, our primary objective is to develop a novel emotion dataset and evaluate the efficacy of different methods of emotion detection specifically within the context of the Bangla language. While numerous traditional machine learning algorithms (that eschew deep learning paradigms) have been tested for performance comparison, it's noteworthy to mention that, to our comprehension, our research will be a novel addition in the realm of bangla emotion detection by introducing larger number of emotion taxonomy. The salient contributions of this paper encapsulate:

- **Novel Dataset:** The formulation of a novel dataset comprised of 20,247 samples, 27 emotion taxonomy and comprehensive analysis of the dataset.
- **Larger Number of Emotion Taxonomy:** We introduce 27 emotion taxonomy inspired by GoEmotion dataset developed by Google which contains 27 emotions categories.

– **Evaluation of the Dataset:** We perform experiment on emotion classification with the developed benchmark dataset utilizing ML, DL and BERT-based models.

2 Related Works

In this section we discuss the literature study for emotion detection from Bangla text.

In 2020, Rayhan et al. [14] utilized a public dataset from Kaggle, which was translated into Bengali using Google Translator. This dataset comprises 7214 sentences, with a vocabulary size of 57000 and an embedding dimension of 64. The maximum input length is 59. They employed two models for emotion classification: CNN-BiLSTM, which achieved an accuracy of 66.62%, and BiGRU, which reached 64.96%. The study focused on classifying six emotions: Happy, Fear, Sad, Angry, Surprise, and Love.

The same year, Nath et al. [9] collected song lyrics to create their dataset for emotion classification. They experimented with a variety of models such as Logistic Regression, SVM, Random Forest, Naive Bayes, LSVM, PSVM, KNN, and Decision Tree. Among these, Random Forest performed the best with an accuracy of 62%. This study aimed to classify lyrics into two emotions: Positive and Negative.

Pran et al. [11] also in 2020, used 1120 comments from coronavirus-related posts to study emotion classification. They utilized Word2Vec embeddings with CNN and LSTM models, with CNN achieving an impressive accuracy of 97.24%. This study classified comments into three emotions: Analytical, Angry, and Depressed.

Lora et al. [6] worked with datasets named Cricket and Restaurant, containing 4468 rows with four columns. They used Glove word embeddings and various models, including CNN, RNN, Stacked LSTM, and Stacked LSTM with 1D convolution. The RNN model achieved the highest accuracy at 98%. This study focused on binary classification of emotions into Positive and Negative.

In 2021, Purba et al. [12] gathered 27,731 Bangla documents from the internet, annotated 995 samples, and classified them into valid emotion classes. The models they used included Logistic Regression, MNB, ANN, and CNN, with MNB achieving the highest accuracy of 68.27%. The emotions classified in this study were Angry, Happy, and Sad.

Parvin et al. [10] accumulated 8458 texts from multiple social media platforms, including Facebook posts and comments. They tested models with both Bag of Words (BoW) and Tf-idf representations. The highest accuracy achieved was with the Tf-idf representation using the SVM model at 62%. This study classified texts into six emotions: Anger, Fear, Disgust, Sadness, Surprise, and Joy.

In 2022, Rahib et al. [13] explored emotion classification using a dataset consisting of 10,581 entries. These entries included comments from social media related to COVID-19, annotated by the authors. They tested multiple models,

including SVM, Random Forest, CNN, and LSTM, achieving the highest accuracy with LSTM at 84.92%. This study focused on three emotions: Insightful, Curious, and Gratitude.

From the literature study we find almost all research experiment on six basic emotion categories (happiness, sadness, anger, fear, surprise, disgust). We get inspired from categories of GoEmotion dataset [23] and try to fill the gap in this domain.

3 Dataset

3.1 Data Collection

For our dataset, we embarked on a meticulous data collection endeavor within the realm of Facebook. Specifically targeting public groups, pages and public user profiles to collect user posts.

We sought to capture a diverse array of perspectives and expressions. To ensure the authenticity and legitimacy of contributors, stringent criteria were applied during the screening process. Initial scrutiny involved assessing Facebook usernames for credibility, with any questionable identities flagged for further scrutiny. Upon identification, we meticulously anonymized personal and identifying information linked to the users through the application of cryptographic hash functions to their Facebook account names, thereby safeguarding individual privacy. The culmination of these efforts yielded a dataset comprising 20,247 meticulously curated posts. Each post underwent manual annotation by human annotators, with a singular focus on emotion labeling. we incorporated insights from the GoEmotions dataset, which contains 58,000 data points. We train several ML, DL and BERT-based models with the GoEmotions dataset to predict emotions within our curated dataset. We choose the best performing model to predict our dataset. Utilizing the trained model, we predicted emotions for each post in our dataset and subsequently compared these predictions with the human-annotated emotions. This comparative analysis allowed us to evaluate the efficacy of the BERT-based models in accurately predicting emotions within our dataset, providing valuable insights into the alignment between human perception and computational predictions of emotions in online discourse.

3.2 Pre-processing

Despite the richness of these posts, they posed a challenge due to the presence of extraneous elements including URLs, images, tags, links, and emojis. To ensure the purity and efficacy of our dataset, a meticulous array of techniques was employed. We selectively retained Bangla alphanumeric characters and punctuation, discarding all other distractions during the preprocessing phase. This entailed the systematic removal of images, links, hashtags, emojis, and user tags, thus distilling the essence of textual content for analysis. Moreover, to refine the dataset further, stop words and meaningless sentences were diligently filtered

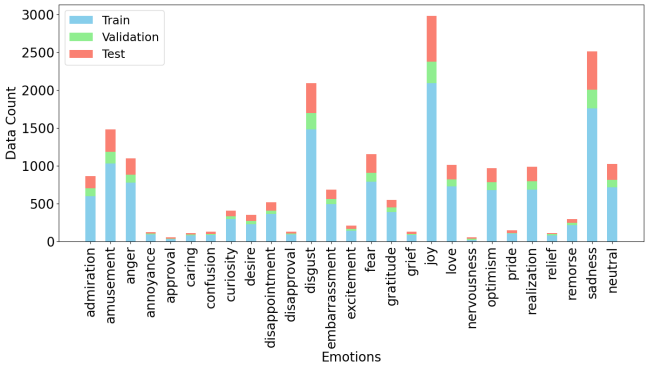

Fig. 1. Bengali Emotion Dataset Distribution

out. This rigorous pre-processing regimen not only streamlined the data but also laid the groundwork for the robust training of our machine learning models. In conjunction with these efforts, we employed the Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer. TF-IDF is a numerical statistic that reflects the importance of a word in a document relative to a collection of documents. By utilizing TF-IDF vectorization, we were able to represent the textual content of our dataset in a format conducive to machine learning analysis. This approach facilitated the extraction of meaningful features from the text, enhancing the effectiveness of our subsequent machine learning algorithms in capturing the nuanced emotional nuances embedded within the posts.

3.3 Data Statistics

Our dataset contains 20,247 user posts. Figure 1 shows the distribution of our dataset for each emotion categories. We see four emotion categories (amusement, disgust, sadness, joy) contain larger amount of samples.

3.4 Ethical Considerations

Our data collection adhered to ethical guidelines, obtaining permissions and consent from Facebook users and groups. We stored only anonymized data and took steps to de-identify sensitive information. Approval from the Institutional Review Board was obtained to ensure compliance. Measures like de-identifying username were in place to protect user privacy and uphold ethical standards throughout the study.

4 Methodology

In the following, we discuss the detailed methodology of our research. We first create a dataset consisting of Facebook posts collected from different Facebook

groups, Facebook pages, individual Facebook profiles and annotate the data before performing the feature extraction. We perform semi-automated annotation process as shown in Figure 2. Then, to reduce noise from the data, we use a variety of pre-processing algorithms on the text. To evaluate our dataset we train different machine learning models with our dataset. To convert each text to a particular input format, we implement the TF-IDF, stylometric and word embedding feature approach for feature extraction. Finally, we present the architectures of our models.

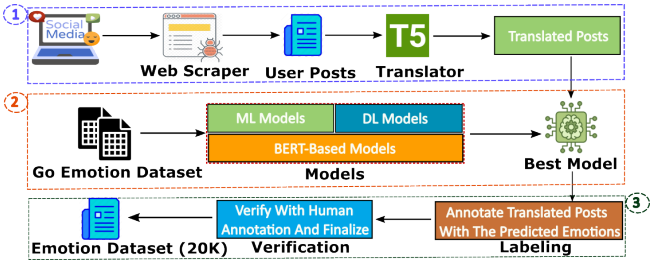

Fig. 2. Semi-automated dataset annotation process flow.

4.1 Dataset Development Process

The prime objective of our work is to develop an emotion dataset that can be used to classify emotion expression, usually written in Bengali, into one of the 27 emotion categories. We introduce 27 emotion categories inspired by categories from GoEmotions dataset [23] developed by researchers at Google. To develop dataset in Bengali is a critical challenge as a language processing task. One of the notable challenges is the scarcity of appropriate emotion text expression. Some links, misspelled sentences, and "Benglish" sentences were obtained during data collection. Moreover, emotion detection from plain text is challenging than detecting emotion from facial expression. Because sometimes people pretend to be alright through text messages with having lots of emotional problem in his day to day life. Figure 2 shows an overview of the development process of our dataset, which consists of five major phases: data collection, preprocessing, data annotation by human annotator, BERT-based automatic label prediction and label verification respectively. We extended the method described by [3] to develop the dataset.

Data Collection We apply a semi-automatic data annotation process to a dataset of 20,247 posts, extending the work initiated in Ahmed et al. [1], with strict adherence to ethical and legal considerations. Targeting Facebook groups, profiles, and pages ranging from 1,000 to 5,000 members or followers. We ensure compliance with Facebook's data usage policies and user privacy guidelines. Our approach encompassed approximately 500 unique users, ensuring diverse representation of Bangla language users and capturing the language's

nuanced expressions comprehensively. Throughout our data collection, we prioritized anonymity and respected user privacy, aligning with GDPR and local regulations in Bangladesh. We extensively de-identified personal information using cryptographic hash functions to anonymize Facebook account names as needed.

Preprocessing We addressed unique challenges specific to the Bangla language in social media, particularly within Facebook group discussions. These posts are often informal and include non-textual elements such as URLs, images, tags, and links. Our preprocessing focused on refining these texts by meticulously removing non-Bangla alphanumeric characters, punctuation, URLs, images, links, hashtags, and user tags. This process, consistent with our previous work [29, 5], we prepare our dataset of size 20,247 with 27 emotion categories (excluding neutral category).

A crucial aspect of our preprocessing involved handling stop words, which are common words that add little semantic value and can introduce dataset noise. We tokenized the texts and systematically removed Bangla stop words using a comprehensive list available in a GitHub repository [20], referenced in the work of Tripto and Ali [21]. This approach ensures transparency and reproducibility in our research methodology.

Through these tailored preprocessing steps, our aim was to refine the dataset to accurately reflect the linguistic patterns essential for emotion detection from social media texts.

Human Annotation The whole corpus was labeled manually, followed by majority voting to assign a suitable label. The labeling or annotation tasks were performed by a group of graduate students from Computer Science background, all of them are doing research on NLP. The labels assigned by the group were finalized by following majority voting.

Label Prediction On top of human annotation we perform automatic label prediction with the best performing model from an array of models trained with a benchmark emotion dataset GoEmotions. We train several ML, DL, BERT-based models for this purpose and mBERT performs the best. Before performing prediction we translate Bengali to English since GoEmotions is an English dataset. We perform several preprocessing steps before performing prediction on the translated dataset.

Label Verification We compare the predicted labels and the human annotated labels. If we find mismatch then we consider the post for relabeling by the human annotators. We find the human annotated labels of the mismatched data points by applying majority voting mechanism. We find less than 50% match with the predicted labels. For this we perform human annotation again on more than half of the data.

4.2 Feature Extraction Methods

We use different feature extraction methods which includes TF-IDF, stylometric, word embedding. To evaluate our dataset we train and compare performance of different ML, DL and BERT-based models. ML and DL models are unable to process raw text. That is why we need feature extraction to train these models.

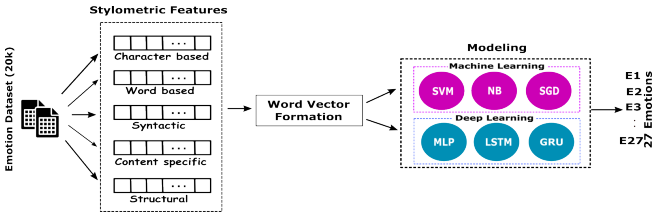

Fig. 3. Model Architecture for Stylometric Feature.

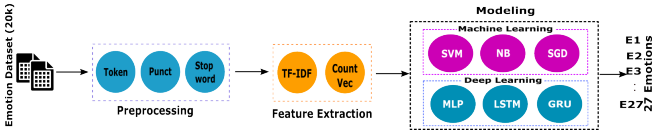

Fig. 4. Model Architecture for TF-IDF Feature.

Stylometric Features Approach Stylometric features are the features that capture the writing style of different authors by calculating some statistical values. We compute a large set of stylometric features based on existing work of [1] These features are categorized into five types: character based features, word based lexical features, syntactic features, structural features, functional words.

For each user post, we generate a feature vector of a dimension of 141, which represents the values of the 141 stylometric features. As these feature sets contain information on the writing style of a user measured by various methods, the feature values can range from 0 to any positive value. As we want to ensure all features are treated equally in the classification process, we normalize the feature values using the min-max normalization method to ensure all feature values are between 0 and 1. We normalize the feature values using the equation below:

$$x'_{ij} = \frac{x_{ij} - min(x_j)}{max(x_j) - min(x_j)}$$

where x_{ij} is the jth feature in the ith example, min (x_j) and max (x_j) are the minimum and maximum feature values of the jth feature respectively.

TF-IDF Vectorizer Approach TF-IDF (term frequency-inverse document frequency) is a text vectorizer that converts the text into a feature vector. After tokenizing a text we get a list of tokens or words. Each token is referred to as a term. In any document, the term frequency represents the number of occurrences of a term. On the other hand, document frequency represents the number of documents containing that term. Term frequency indicates the importance of a specific term in a document. Document frequency indicates how common the term is [7]. We implement TF-IDF vectorizer from `scikit-learn python` library. We take the most frequent 1000 words (tokens) to limit the number of features to be extracted from each document. As part of pre-processing, we remove letters other than the Bangla alphabet. For example, let us consider we

have some texts, which have to be converted to feature vectors using a TF-IDF vectorizer. For converting these texts into feature vectors, we first identify unique words and count how many times these words occur in each text. Then we compute inverse document frequency (IDF) using the following formula:

$$idf_i = log\frac{n}{df_i}$$

where df_i represents how many documents contain the term i and n is the total number of documents. We calculate the inverse document frequency for each word. Then TF matrix is multiplied by the IDF score to get a vectorized form of each text. We convert all texts into vectors. These vectors can be fed into any machine-learning algorithm.

Word Embedding Representation Approach In this approach, individual words are represented as vectors in a predefined vector space, capturing semantic and contextual information efficiently. The word2vec algorithm, introduced by Google [8], facilitates this process through two main architectures: Continuous Bag of Words (CBOW) and Skip-Gram (SG).

CBOW predicts a target word based on its surrounding context words, while SG predicts surrounding context words given a central word. Both models excel in learning word embeddings that encapsulate meaningful relationships in language without being overly influenced by word order.

To apply word2vec to our dataset, we leverage these CBOW and Skip-Gram models to generate word embeddings from user posts on Facebook. These embeddings encode semantic, conceptual, and contextual information, enhancing our understanding of the language used in the dataset.

Finally, we utilize these word embeddings to classify our dataset, enabling us to infer emotions from textual data more effectively.

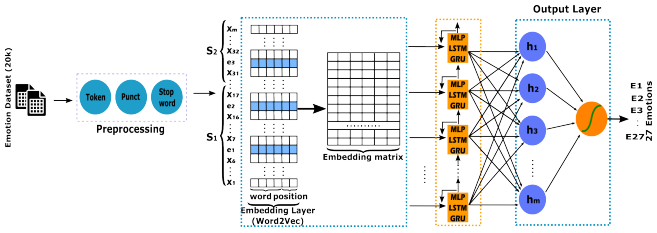

Fig. 5. Model Architecture for Word Embedding Feature.

5 Model Architecture

We utilize ML, DL, and BERT-based models for emotion classification. Our ML model training includes Support Vector Machine, Naive Bayes and SGD. For DL

models, we train Multi-layer Perceptron (MLP), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM). Additionally, we fine-tune BERT-based models such as mBERT, indicBERT, BanglaBERT, and XLM-R.

5.1 Models with Stylometric Features

We prepare word vectors using the stylometric feature approach. These vectors are passed to the LSTM layer having 300 nodes. The output of the LSTM layer is passed to a dense layer. We use sigmoid [28] as an activation function. The optimizer is RMSprop and binary cross entropy [16] is used as a loss function. RMSprop is a gradient-based optimization technique used to change weights and the learning rate of the neural network to reduce the losses. We repeat the same process for MLP, GRU, and other traditional machine learning models. For each model, we note down each model's accuracy and F1-Score. Figure 3 shows the proposed architecture of the stylometric features for the traditional and deep learning models.

5.2 Models with TF-IDF Features

We implement traditional machine learning and deep learning models for emotion detection from the text using Term Frequency Inverse Document Frequency (TF-IDF) features. We initialize the TF-IDF vectorizer with n-gram range of (1, 5) and max_feature set to 1000. Then we transform our data using the TF-IDF vectorizer. Figure 4 shows the architecture of traditional machine learning and deep learning models with TF-IDF features.

5.3 Models with Word Embedding Features

After preprocessing each sentence, we employ a tokenizer that generates one-hot encoding vectors of length 100. Only the top 1000 most frequent words are included in our vocabulary, with longer sentences truncated to 100 words and shorter ones padded with zeros. These vectors are then passed through an embedding layer initialized with weights from a word2vec model. The embedding layer can be configured as trainable or non-trainable; the latter freezes the pre-trained weights to prevent updates during training. Each word vector in the embedding layer has a dimensionality of 300, consistent with the word2vec model. The sequence of 100 words is subsequently processed through an LSTM layer, and the output is fed into a dense layer for emotion detection. In the dense layer, a Sigmoid activation function is applied [28]. The optimizer RMSprop is utilized, with binary cross-entropy serving as the loss function [16]. This architecture is replicated across all other deep-learning models used in our experiments. Figure 5 shows the architecture of our models with the word embedding features.

5.4 BERT-based Models

We perform experiment on four BERT-based models: m-BERT, Bangla-BERT, XLM-R, and Indic-BERT on our dataset. In recent years transformer-based BERT models are being used extensively for classification tasks to achieve state-of-the-art results. We use Huggingface transformers library and fine-tune on our dataset by using PyTorch package.

mBERT mBERT, based on the transformer architecture introduced in [24], is a model pre-trained on a corpus spanning 104 languages, featuring over 110 million parameters. For our study, we utilized the 'bert-base-multilingual-cased' variant and fine-tuned it using a batch size of 16 on our dataset.

IndicBERT IndicBERT [26] is a transformer model designed specifically for languages of the Indian subcontinent, leveraging techniques similar to BERT (Bidirectional Encoder Representations from Transformers). It is pre-trained on a diverse corpus covering multiple Indic languages, including but not limited to Hindi, Bengali, Tamil, and Telugu, among others. IndicBERT is based on the 'bert-base-multilingual-cased' architecture, which incorporates cased representations suitable for differentiating capitalizations in languages like Hindi and others that require case distinctions. We adapted IndicBERT with a batch size of 16.

Bangla BERT Bangla BERT, as described in [15], is a BERT-based model pre-trained on a substantial Bengali corpus. In our work, we employ the 'sagorsarkar/Bangla-bert-base' model and conducted fine-tuning tailored to our dataset. A batch size of 16 was utilized to enhance the model's performance on our specific tasks.

XLM-R XLM-R, developed by [22], is a large-scale multilingual language model trained across 100 diverse languages. Our study utilizes the 'xlm-Roberta-base' model, applying it to our dataset with a batch size of 16.

Each of the BERT-based models underwent training for 5 epochs using a learning rate set at 2e-5. The best-performing model based on validation results was selected for final predictions on the test dataset.

6 Results

In this section, we assess the effectiveness of our proposed approaches for emotion detection using data sourced from Facebook. We evaluate the performance across traditional machine learning models, deep learning models, and BERT-based architectures.

6.1 Experimental Setup

We use Scikit-learn for traditional machine learning models, PyTorch for deep learning, and Huggingface for BERT-based models. Experiments were conducted on a machine with an Intel Core i7 1.8GHz processor, 8GB RAM, and an NVIDIA GeForce MX150 GPU with 2GB memory, which was fully utilized in PyTorch-based experiments. Adding a GPU significantly speeds up PyTorch experiments, as shown in [27].

6.2 Result Analysis

In this section, we examine the performance of the ML, DL and BERT-based models using accuracy and F1 score.

Performance of ML models We measure the machine learning models performance by observing the F1-score and accuracy. We list the F1-score and accuracy of each model in Table 1. SGD model for TF-IDF feature approach achieves the highest F1 score 68% and accuracy 72% among the models evaluated. SVM and NB models follow closely behind SGD, with slightly lower scores across all metrics. Overall, SGD performs the best in this evaluation, suggesting its effectiveness in this particular classification task compared to SVM and NB models.

Model	**F1(%)**	**Acc(%)**
NB (t)	63	68
NB (s)	64.50	69
SVM (t)	62.50	67.50
SVM (s)	63.50	68.50
SGD (t)	**68**	**72**
SGD (s)	67.50	71.50

Table 1. Performance measure of ML models. Here s = 'stylometric', t = 'TF-IDF'.

Performance of DL models We measure the DL models performance by observing the F1-score and accuracy. We list the F1-score and accuracy for each DL model in Table 2. Across the models, LSTM for stylometric feature approach achieves the highest F1 score (64%) and accuracy (68%), followed closely by MLP, GRU. GRU and MLP models generally exhibit lower performance with F1 scores ranging from 44.5% to 47.5%.

Performance of BERT-based models We measure the BERT-based model performance by observing the F1-score and accuracy. Table 3 presents the performance metrics of various pre-trained language models evaluated on emotion detection. XLM-R emerges as the top-performing model across all metrics,

Model	F1(%)	Acc(%)
MLP (s)	44.50	47.50
MLP (w)	44	47
GRU (s)	44	47
GRU (w)	43.50	46.50
LSTM (s)	**64**	**68**
LSTM (w)	63.50	67.50

Table 2. Performance measures for Deep Learning models. Here s = 'stylometric', w = 'word2vec'.

achieving an F1 score of 79% and an accuracy of 82%. This performance signifies XLM-R's superior capability in the evaluated tasks compared to the other models. mBERT and indicBERT follow with slightly lower scores across all metrics, demonstrating solid performance but trailing behind XLM-R. BanglaBERT shows the lowest performance, with an F1 score of 67% and an accuracy of 72%, indicating room for improvement in handling emotion detection tasks.

Model	F1(%)	Acc(%)
mBERT	75	78
indicBERT	73	76
XLM-R	**79**	**82**
BanglaBERT	67	72

Table 3. Performance measures for BERT-based models.

6.3 Comparison with Existing Work

In recent years, there has been significant progress in emotion detection research. Studies have explored various methods, ranging from traditional machine learning techniques to advanced deep learning approaches. Table 4 shows a comparison among existing work on emotion detection dataset, evaluation methods. From the list we see almost all the work are related to six basic emotion categories whereas our paper introduces 27 categories, a larger number of emotion taxonomy.

7 Conclusion & Future Work

Since no other research has been done to introduce larger number of emotion taxonomy for emotion detection in Bangla language, our dataset represents a novel introduction of emotion dataset with larger number of emotion taxonomy. In this work, we have gathered a user post dataset from different Facebook groups, Facebook pages and individual profiles. Then we perform a semi-automated data

Ref.	Acc. (%)	No. of Emotions	Data Size
[14]	66.62	6	7k
[13]	84.92	3	10k
[9]	62	2	1500
[12]	68.27	3	27k
[11]	97.24	3	1120
[6]	98	2	4k
[10]	62	6	8k
[14]	62.62	6	7k
[3]	69.61	6	6k
[25]	-	6	7k
This paper	82	**27**	20k

Table 4. Comparison with existing work.

annotation mechanism where automatic emotion prediction was performed by BERT-based model trained on GoEmotion dataset. Then, demonstrate some statistical analysis of our dataset. Finally, we evaluated the efficacy of ML, DL, BERT-based models in emotion detection on our dataset. We measure F1 score, accuracy to measure performance of the models. The BERT based model XLM-R performs best among all the models. To improve the deep learning model's performance, the data set can be improved, parameter tuning can be performed. The dataset is not currently available publicly, it can be provided upon request through email. As a future work we can increase dataset size, compare deep learning models performance with GPT models.

8 Acknowledgment

This research was supported by NSF grant CNS-2153482.

References

1. Ahmed, S., Alam, Md J., Talukder, S., Hossain, I.: Towards Addressing Identity Deception in Social Media using Bangla Text-Based Gender Identification. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, pp. 72–76 (2023)
2. Chung Hwan Kwak.: New World Encyclopedia. (2020). https://www.newworldencyclopedia.org/entry/Bengali_language. Accessed 2 August 2024
3. Das, A., Sharif, O., Hoque, M. M., Sarker, I. H.: Emotion classification in a resource constrained language using transformer-based approach. arXiv preprint arXiv:2104.08613 (2021)
4. Eglitis-media.: worlddata.info. (2024). https://www.worlddata.info/languages/bengali.php. Accessed 2 August 2024
5. Hossain, I., Puppala, S., Alam, Md J., Talukder, S.: A Visual Approach to Tracking Emotional Sentiment Dynamics in Social Network Commentaries (2024)

6. Lora, S. K., Jahan, N., Antora, S. A., Sakib, N.: Detecting emotion of users' analyzing social media bengali comments using deep learning techniques. In: 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT), pp. 88–93. IEEE (2020)
7. Luthfi Ramadhan.: TF-IDF Simplified. (2021). https://towardsdatascience.com/tf-idf-simplified-aba19d5f5530. Accessed 2 August 2024
8. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781 (2013)
9. Nath, D., Roy, A., Shaw, S. K., Ghorai, A., Phani, S.: Textual lyrics based emotion analysis of bengali songs. In: 2020 International Conference on Data Mining Workshops (ICDMW), pp. 39–44. IEEE (2020)
10. Parvin, T., Hoque, M. M.: An ensemble technique to classify multi-class textual emotion. Procedia Computer Science, vol. 193, pp. 72–81. Elsevier (2021)
11. Pran, Md S. A., Bhuiyan, Md R., Hossain, S. A., Abujar, S.: Analysis of Bangladeshi people's emotion during COVID-19 in social media using deep learning. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–6. IEEE (2020)
12. Purba, S. A., Tasnim, S., Jabin, M., Hossen, T., Hasan, Md K.: Document level emotion detection from bangla text using machine learning techniques. In: 2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD), pp. 406–411. IEEE (2021)
13. Rahib, Md R. H. K., Tamim, A. H., Tahmeed, M. Z., Hossain, M. J.: Emotion detection based on Bangladeshi people's social media response on COVID-19. SN Computer Science, vol. 3, no. 2, p. 180. Springer (2022)
14. Rayhan, Md M., Al Musabe, T., Islam, Md A.: Multilabel emotion detection from bangla text using bigru and cnn-bilstm. In: 2020 23rd International Conference on Computer and Information Technology (ICCIT), pp. 1–6. IEEE (2020)
15. Sagor Sarker.: BanglaBERT: Bengali Mask Language Model for Bengali Language Understanding. (2020). https://github.com/sagorbrur/bangla-bert
16. Saxsena, S.: Binary Cross Entropy/Log Loss for Binary Classification. Log Loss for Binary Classification, pp. 02–08 (2021)
17. Sharma, O.: A New Activation Function for Deep Neural Network. In: International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. Feb 14-16, Faridabad, India (2019)
18. Simon Kemp.: DIGITAL 2021: BANGLADESH. (2021). https://datareportal.com/reports/digital-2021-bangladesh. Accessed 2 August 2024
19. StatCounter Global Stats.: Social Media Stats in Bangladesh. (2024). https://gs.statcounter.com/social-media-stats/all/bangladesh. Accessed 2 August 2024
20. stopwords-iso.: Stopwords Bengali. (2024). https://github.com/stopwords-iso/stopwords-bn. Accessed 2 August 2024
21. Tripto, N. I., Ali, M. E.: Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments. In: International Conference on Bangla Speech and Language Processing (ICBSLP), pp. Sept 21-22, Sylhet, Bangladesh (2018)
22. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised Cross-Lingual Representation Learning at Scale. In: arXiv preprint arXiv:1911.02116 (2019)
23. Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., Ravi, S.: GoEmotions: A Dataset of Fine-Grained Emotions. In: arXiv preprint arXiv:2005.00547 (2020)

24. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. In: arXiv preprint arXiv:1810.04805 (2018)
25. Iqbal, M. A., Das, A., Sharif, O., Hoque, M. M., Sarker, I. H.: BeMoc: A Corpus for Identifying Emotion in Bengali Texts. In: SN Computer Science, vol. 3, no. 2, pp. 135 (2022)
26. Kakwani, D., Kunchukuttan, A., Golla, S., Gokul, N. C., Bhattacharyya, A., Khapra, M. M., Kumar, P.: IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-Trained Multilingual Language Models for Indian Languages. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 4948–4961 (2020)
27. Matthews, A. G. de G., Van Der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., Le, P., Ghahramani, Z., Hensman, J., et al.: GPflow: A Gaussian Process Library Using TensorFlow. In: Journal of Machine Learning Research, vol. 18, no. 40, pp. 1–6 (2017)
28. Sharma, O.: A new activation function for deep neural network. In: 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), pp. 84–86 (2019)
29. Hossain, I., Puppala, S., Alam, Md J., Talukder, S.: Monitoring Dynamics of Emotional Sentiment in Social Network Commentaries. In: Proceedings of the International Conference on Advances in Social Networks Analysis and Mining, pp. 51–55 (2023)