2023 26th International Conference on Computer and Information Technology (ICCIT)
13-15 December 2023, Cox's Bazar, Bangladesh

BEmoFusionNet: A Deep Learning Approach For Multimodal Emotion Classification in Bangla Social Media Posts

Zaima Sartaj Taheri[1], Animesh Chandra Roy[2], Ahasan Kabir[3]
Department of Computer Science & Engineering,
Chittagong University of Engineering and Technology, Chattogram-4349, Bangladesh
{Email-zstaheri1999, animeshroycse, ahasankabir146}@gmail.com

***Abstract*—Multimodal emotion classification, incorporating both image and text modalities, has gained significant attention due to the exponential growth of multimedia data. This research aims to develop a robust system for multimodal emotion classification in Bangla social media content, utilizing both image and text data. Despite limited resources, understanding emotions in Bangla is crucial for mental health interventions. Unlike previous studies focused on well-resourced languages, this research addresses the gap by specifically targeting regional languages like Bangla, going beyond the traditional positive, negative, and neutral classes. This work presents a multimodal Bangla social post dataset containing 4660 samples. Transfer learning techniques, utilizing pre-trained models like ResNet50, VGG16, and InceptionV3, extract visual features, while deep learning architectures such as BiLSTM and CNN are employed for textual content analysis. Multimodal learning techniques, including feature fusion and decision fusion, are explored to combine visual and textual representations. We evaluated the feature fusion of InceptionV3 and BiLSTM features on our Bangla social media post dataset. Our approach achieved a weighted f1-score of 77.50%.**

***Index Terms*—Natural language processing, Multimodal, Feature Fusion, Decision Fusion**

I. Introduction

Social media platforms have become popular venues for people to express their opinions on a wide range of topics. Automatic emotion analysis can be used to identify a user's sentiments and perspectives on specific events or topics and has applications in a variety of fields, including cognitive psychology and automated identification. Emotions can be expressed through a variety of channels, including images, speech, and text, and fusing these modalities can significantly improve the accuracy of emotion recognition. As a result, the use of multimodal data for emotion recognition has gained significant attention, driven by the exponential growth of such data on social media platforms. Implicit affective cues are embedded in the digital media we commonly engage with, such as text, images, audio, and videos. These subtle cues provide an indirect way to gauge emotions, enriching our understanding of affective states. The fusion of different modalities in multimodal emotion classification offers several advantages over unimodal approaches, enabling a comprehensive interpretation of emotions [1] [2].

Despite Bangla's prominent presence on the internet, most studies on emotion recognition are rooted in languages like English, overlooking the unique linguistic landscape of Bangla. Notably, numerous studies on Bangla emotion recognition have predominantly focused on text data, though emerging research suggests that the inclusion of images alongside text yields more precise emotional insights. However, research on this specific subject, particularly within the context of the Bangla language, remains scarce. This lacuna served as the impetus for our endeavor to develop an automated multimodal emotion classification system adept at discerning emotions in posts. An inherent challenge in multimodal emotion recognition lies in the paucity of annotated training data. Existing multimodal datasets primarily revolve around facial and human images, sidelining generic images featuring non-human entities. Furthermore, available multimodal datasets featuring generic images tend to offer emotion labels classifying sentiments into broad categories, such as positive, negative, or other, rather than furnishing a comprehensive spectrum of multi-class emotion labels. In response, this work culminated in the creation of a Bangla multimodal dataset and an associated emotion classification system. Leveraging advanced deep learning techniques, including BiLSTM, CNN, and pre-trained models such as VGG16, Resnet50, and InceptionV3, we explored various multimodal learning strategies, encompassing feature fusion and decision fusion, to seamlessly integrate visual and textual representations. The significant contributions of this research are outlined below:

- We developed a Bangla multimodal dataset comprising 4660 Bangla social media posts.
- We explored multimodal learning techniques to fuse visual and textual features for comprehensive emotion recognition.

II. Related Studies

Researchers have made notable progress in Bengali text-based emotion recognition, detecting six basic emotions. Parvin et al. [3] developed a Bengali emotion corpus to recognize 6 basic emotions from the text. They used 8 popular machine learning algorithms to detect emotions, and tf-idf yielded the highest weighted f1-score of 62.39%. Islam et al. [4] proposed BERTBSA for sentiment analysis. For multilingual emotion recognition, Bianchi et al. [5] introduced

XLM-EMO across 19 languages. Transformer models were effective for Bengali emotion classification [6], with XLM-R achieving a weighted f1-score of 69.73%. 'BEmoC' [7] provided a robust emotion dataset, achieving a high Cohen's kappa score of 0.969. 'BSaD' [8] enabled sentiment classification with an ensemble approach reaching 82% accuracy. 'BEmoD' [9] enhanced emotion analysis with a Cohen's kappa score of 0.920. An ensemble-based technique [8] achieved a weighted f1-score of 62.46% for textual emotion classification. A study [10] introduced the Multimodal EmotionLines Dataset (MELD), which encompasses dialogues from the Friends TV series with both audio and textual modalities, labeled with emotion and sentiment descriptors. Yang et al. utilized MVSA-Single, MVSA-Multiple, and TumEmo datasets to classify emotions into seven classes using features extracted from images and text, achieving an f1-score of 63.39% [1]. Another study [2] effectively detected seven emotions by combining text, emoji, and image features with an accuracy of 71.98%. Eftekhar et al. introduced the Bengali dataset MemoSen, containing memes with annotated sentiment labels, leading to improved multimodal models compared to unimodal ones [11]. Additionally, a Multimodal Emotion Recognition (MER) dataset was constructed from diverse sources, with a framework encompassing representation learning, feature fusion, and classifier optimization [12]. Gundapu and Mamidi proposed a novel multimodal neural network approach for classifying Internet memes into sentiment categories, humor types, and expression extents, achieving a micro f1-score of 58.96% [13].
Our work addresses notable contributions through an expanded emotion spectrum, advanced multimodal fusion, specialized Bengali corpus, and benchmark performance metrics in comparison to prior studies. We collected 4660 Bangla social media posts covering seven emotional classes and proposed a multimodal early fusion approach, leveraging InceptionV3 and BiLSTM models.

III. Dataset Development

To address the lack of an existing multimodal emotion dataset in Bangla, we took the initiative to create our own corpus. We followed a thorough process inspired by the methodology presented in [7].

A. Data Accumulation

We curated a Bangla-language social media corpus of 4786 image-caption pairs, primarily sourced from Facebook and Instagram. To ensure annotation consistency, we implemented several cleaning steps, removing non-Bangla texts, duplicate texts, links, and blurry images. This resulted in a refined corpus of 4726 image-caption pairs suitable for manual annotation.

B. Data Annotation and Quality

The dataset underwent rigorous manual annotation, classifying posts into seven distinct emotional categories: happy, sad, disgust, fear, surprise, angry, and other. This classification adhered to Ekman's guidelines [14], ensuring consistency and minimizing annotation bias. Three annotators independently labeled each instance, employing a majority voting approach. To ensure the quality of the data annotation, we used Cohen's kappa coefficient to calculate inter-annotator agreement. The mean kappa score of 0.77 indicates a moderate level of agreement between the annotators.

C. Data Statistics

The dataset contains 4660 image-caption pairs after cleaning and annotation, occupying 3.24 GB of storage. It is divided into training (3833 samples), validation (413 samples), and test (414 samples) sets. The training set captions were analyzed in detail. A summary of the caption statistics in the training set is presented in Table I. In the dataset, the "happy" category

TABLE I: Data statistics for the captions

Class	Train	Validation	Test	Total	NW	NUW	AW
Happy	861	91	93	1045	9440	3933	11.25
Angry	477	66	44	587	5686	2528	12.44
Disgust	348	39	40	427	4610	1979	13.89
Fear	400	44	44	488	4341	1484	11.19
Sad	621	70	68	759	7809	3099	13.04
Surprise	410	43	47	500	4817	2460	11.86
other	716	60	78	854	7361	3487	10.41

displayed the highest number of words(NW) at 9440, followed by "disgust" with 5686 words, while "fear" had the lowest count at 4610 words. Examining the number of unique words uniqueness(NUW), "happy" led with 3933 distinct words, followed by "other" with 3487, and "disgust" with the fewest at 1979. Notably, "disgust" also had the highest average word in texts(AW) count per text at 13.89, followed by "sad" at 13.04. Conversely, "other" exhibited the lowest average at 10.41 words. These insights are invaluable for tailoring text processing techniques in emotion classification tasks. Jaccard similarity index measures overlap between two sets to determine shared and distinct members. The resulting Jaccard similarity scores are presented in Table II.

TABLE II: Jaccard similarity of 400 most frequent words between each pair of classes

	Happy	Angry	Disgust	Fear	Sad	Surprise	Other
Happy	1	-	-	-	-	-	-
Angry	0.29	1	-	-	-	-	-
Disgust	0.24	0.27	1	-	-	-	-
Fear	0.24	0.38	0.24	1	-	-	-
Sad	0.36	0.38	0.28	0.32	1	-	-
Surprise	0.35	0.29	0.25	0.26	0.34	1	-
Other	0.31	0.30	0.25	0.24	0.33	0.28	1

IV. Methodology

This study investigates multimodal emotion classification in social media posts, exploiting visual (VGG16, InceptionV3, ResNet50) and textual (BiLSTM, CNN) features. ModelCheckpoint saves the best model based on validation accuracy to prevent overfitting, while class weights address class imbalance.

A. Data Preprocessing

In our study, we utilized deep learning techniques to process visual and textual data. Images were standardized to 150 x 150 x 3 dimensions and normalized before feeding into neural networks, streamlining computational complexity during training. For text data, we removed emojis, special characters, and non-bangla characters, as they were deemed potential sources of noise during classification. We also used a tokenizer function to convert input texts into uniform vectors of distinctive integers with equal lengths ensured through padding. We determined the maximum text length based on frequency distribution analysis, setting it at 180 for the dataset. For example, the text, "একদিকে ১৫ কোটি টাকা খরচ করে কনসার্ট করে, অন্য দিকে মানুষ খাদ্যের জন্য টিসিবির গাড়ির পিছনে লাইন ধরে। সাধ্বীনতার ৫০ বছরে সাধ্ব নিচ্ছে মানুষ।" . After performing automatic preprocessing using a Python script, the given example comes into "একদিকে ১৫ কোটি টাকা খরচ করে কনসার্ট করে অন্য দিকে মানুষ খাদ্যের জন্য টিসিবির গাড়ির পিছনে লাইন ধরে সাধ্বীনতার ৫০ বছরে সাধ্ব নিচ্ছে মানুষ".

B. Visual Approach

We use transfer learning to extract visual features from images. We leverage pre-trained ResNet50, VGG16, and InceptionV3 models trained on ImageNet. To obtain visual features, we remove the top two layers of the pre-trained models.

VGG16: To accomplish our task, we modified the pre-trained VGG16 model by freezing its top layers and adding a global average pooling layer to reduce spatial information. Finally, we included a dense output layer for classification purposes.

ResNet50: We adopted the transfer learning approach to adapt the pre-trained ResNet50 model to our specific task. The last layers, including a global average pooling layer and a dense layer, were fine-tuned with new weights for our task.

InceptionV3: To complete our objective, we incorporated a global average pooling layer, followed by a fully connected layer consisting of 7 neurons, and a softmax layer for class prediction.

C. Textual Approach

We explored various widely used deep learning models, such as BiLSTM and CNN, for the task of textual emotion classification. These models relied on word embedding features for their training.

1) Word Embedding: Textual feature extraction is essential for training classifier models such as BiLSTM and CNN, which cannot be learned from raw text data. Word embedding captures the context, semantic and syntactic similarity, and relationships of words in a document, facilitating the learning process, especially in deep neural networks. Keras word embedding is used to train these models, transforming each word into a 100-dimensional vector representation using a vocabulary size of 50,000, an embedding dimensionality of 100, and a maximum sequence length of 180.

2) Deep Learning-Based Methods: This study investigated several deep learning-based models, such as BiLSTM and CNN, for the task of textual sentiment classification.

BiLSTM: We constructed a Bidirectional Long Short-Term Memory (BiLSTM) network, consisting of two layers with 128 units in the first layer. The output of the last BiLSTM layer is fed into a dense layer with a softmax activation function for emotion classification.

CNN: We used a CNN for text classification. The input text is embedded into 100-dimensional vectors, and then passed through two convolutional layers with 64 and 32 filters, respectively. Max pooling is used to reduce dimensionality. The resulting feature map is flattened and fed into an output layer.

D. Multimodal Approach

To incorporate multiple modalities, we adopted two different techniques, namely feature fusion and decision fusion [15]. Decision fusion combines the softmax outputs of visual and textual models, whereas feature fusion aggregates a flexible hidden layer from diverse modalities. Specifically, we used the three visual models (i.e., VGG16, ResNet50, and InceptionV3) and two textual models (i.e., BiLSTM and CNN) to construct the six fusion models (i.e., VGG16 $\oplus$ BiLSTM, VGG16 $\oplus$ CNN, ResNet50 $\oplus$ BiLSTM, ResNet50 $\oplus$ CNN, InceptionV3 $\oplus$ BiLSTM, and InceptionV3 $\oplus$ CNN) in our experiment. We combined features from visual and textual models. The resulting outputs were concatenated, and a dense layer consisting of 64 neurons followed by dropout with a probability of 0.2 was applied to the combined feature. Finally, a softmax layer was employed to classify the data.

E. Proposed Method

The proposed framework for emotion classification from social media posts is depicted in Figure 1, which utilizes a feature fusion technique of image and text. Inception V3 handles visual feature extraction, fine-tuned via transfer learning with frozen top layers and augmented by a global average pooling layer. Keras embedding is applied to the BiLSTM's embedding layer for text processing, with specific dimensions tailored to the dataset's vocabulary and sequence length. The BiLSTM comprises two layers with 128 units in the first layer. Hyperparameters are detailed in Table **??**. Concatenating output from the dense layers of InceptionV3 and BiLSTM forms the multimodal model's output. Subsequently, a dense layer with 64 neurons, followed by a 0.2 dropout, is integrated. The final step involves an output dense layer. The model employs categorical crossentropy loss and the Adam optimizer during training. Concatenation is chosen for fusion as it effectively combines visual and textual features, preserving critical information from both modalities for improved performance.

V. Results

We conducted all experiments on Google Colaboratory using Python 3, pandas, numpy (v1.18.5), scikit-learn (v0.22.2), Keras (v2.4.0), and TensorFlow (v2.3.0) for deep learning. We

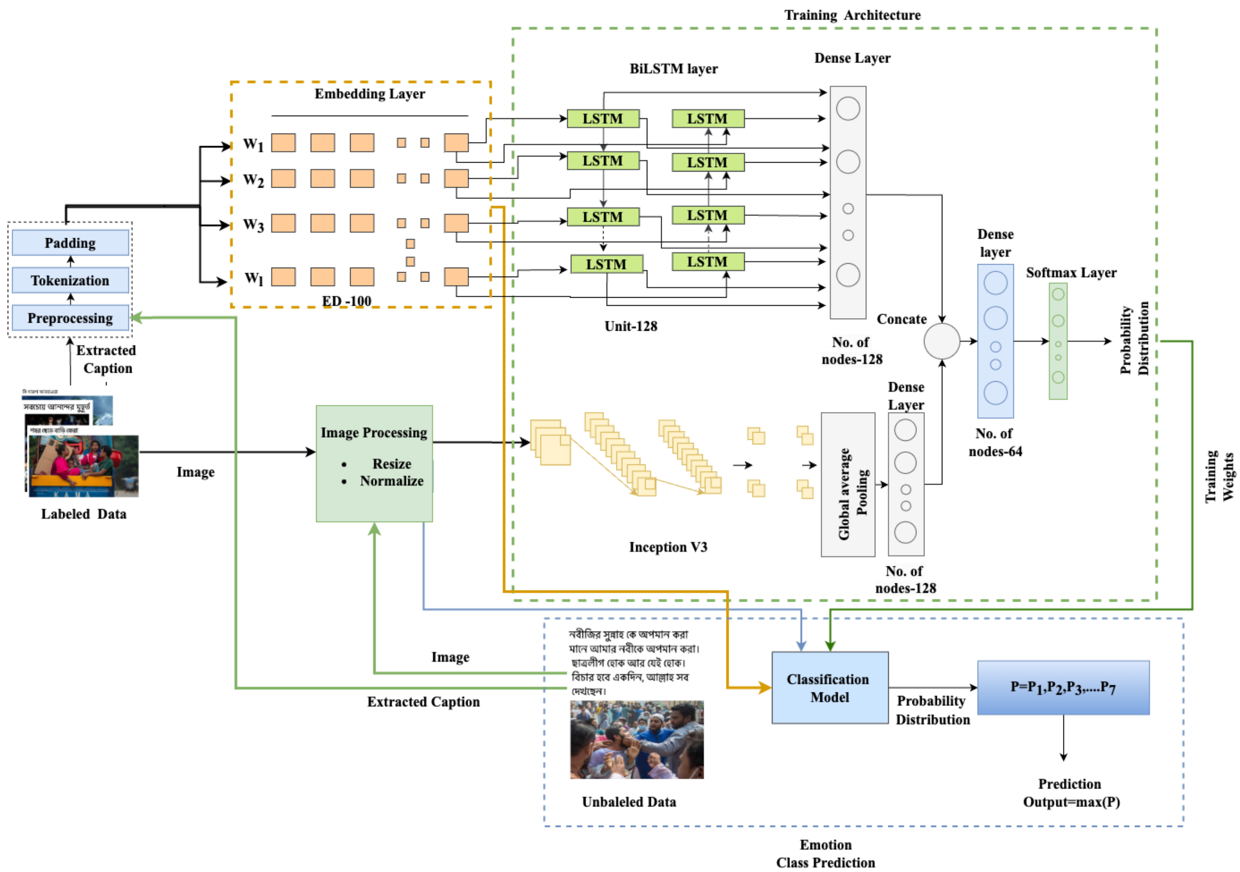


Fig. 1. Overall architecture of the proposed framework for emotion classification

TABLE III: List of hyperparameters values for the proposed model.

Hyperparameters	Optimum Value
Input text length	180
Embedding dimension	100
LSTM hidden units	128
Neurons(Last FC Layer)	64, 128
Pooling type	'average'
Loss function	Categorical crossentropy
Optimizer	Adam
Batch size	32
Epochs	50

evaluated four categorization approaches for Bangla-language social media posts: visual, textual, feature fusion, and decision fusion. We compared the performance of different models based on weighted f1-score, precision, recall, and MCC. The multimodal approach, combining visual and textual architectures with concatenation (⊕)), exhibits promising performance, as detailed in Table IV. The study employed three pre-trained models (VGG16, ResNet50, InceptionV3) in the visual approach. ResNet50 exhibited slightly superior results with an f1-score of 70.36%. For the textual approach, BiLSTM demonstrated a notable f1-score (75.72%). In the decision fusion approach, the combination of ResNet50 and BiLSTM stood out, yielding an f1-score of 75.87%. The exploration of feature fusion revealed promising results. The combination of the InceptionV3 ⊕ BiLSTM model displayed the highest f1-score (77.50%). Table V offers a comprehensive comparison of previous methodologies for multimodal emotion classification, integrating both image and text data in our dataset. RestNet50⊕CNN, as detailed in [11], achieved an f1-score of 74.86%. Meanwhile, the RestNet152⊕BiLSTM approach, presented in [2], demonstrated f1-score of 75.88%. Our proposed InceptionV3⊕BiLSTM method, however, surpassed both prior techniques, securing the highest f1-score which is 2.98% and 2.16% higher than the previous works, respectively.

Examining the classification reports and confusion matrix in Figure 2 allows us to gain a more in-depth understanding of how different models perform in the various classes. The precision and recall scores for all classes for the suggested multimodal model (InceptionV3⊕BiLSTM) were discovered to be noticeably higher. With f1-scores ranging from 0.79 to 0.85, the model specifically earned high scores for the happy, disgust, fear, and other classes. With f1-scores of 0.75 and 0.80, respectively the model likewise attained a moderate f1-score for the sad and surprise classes. The model only struggled with the angry class, achieving a low f1-score of

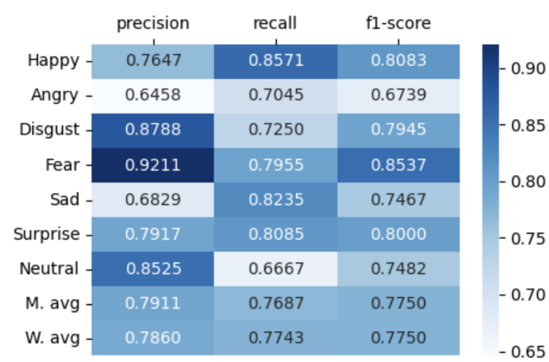
TABLE IV: Evaluation results of models on the test set

Approach	Classifier	Precision(%)	Recall(%)	F1-score(%)	MCC(%)
Visual	VGG16	71.75	68.44	70.06	63.15
	ResNet50	73.36	70.53	70.36	65.36
	InceptionV3	72.09	68.62	70.31	64.23
Textual	BiLSTM	77.75	75.36	75.72	71.07
	CNN	76.72	75.12	75.63	70.63
Decision Fusion	VGG16 ⊕ BiLSTM	78.18	74.88	75.73	70.53
	VGG16 ⊕ CNN	78.30	73.91	74.49	69.40
	ResNet50 ⊕ BiLSTM	77.08	75.06	75.87	71.17
	ResNet50 ⊕ CNN	77.20	74.15	74.38	69.56
	InceptionV3 ⊕ BiLSTM	79.87	74.40	75.39	70.50
	InceptionV3 ⊕ CNN	73.89	71.50	72.05	66.62
Feature Fusion	VGG16 ⊕ BiLSTM	77.42	75.49	75.51	71.15
	VGG16 ⊕ CNN	76.41	75.00	75.42	70.51
	ResNet50 ⊕ BiLSTM	78.11	75.73	75.88	71.57
	ResNet50 ⊕ CNN	77.23	74.51	74.80	70.06
	InceptionV3 ⊕ CNN	77.97	75.49	75.76	71.15
	InceptionV3 ⊕ BiLSTM	78.60	77.43	**77.50**	73.34

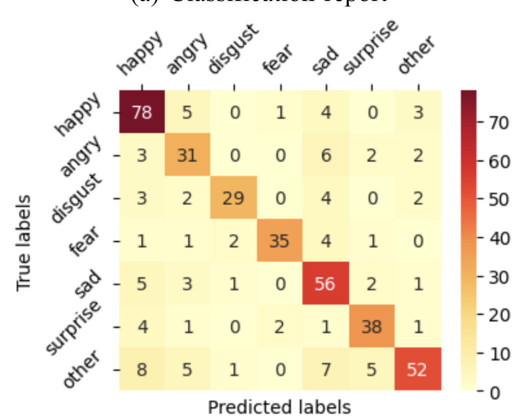
TABLE V: Performance comparison with previous approaches

Method	F1-score
RestNet50⊕CNN [11]	75.19%
RestNet152⊕BiLSTM [2]	75.86%
InceptionV3⊕BiLSTM(Proposed Method)	**77.50%**

0.67, suggesting that it might have had trouble distinguishing this class from others. The confusion matrix for our proposed method (InceptionV3⊕LSTM) shows that the model was the most accurate in detecting the happy emotion, with 78 true positives, followed by sad and other emotions with 56 and 52 true positives, respectively. The model performed relatively well for angry and surprise emotions, with 31 and 38 true positives, respectively. It struggled more with disgust and fear emotions, with 29 and 35 true positives, respectively. Table VI provides descriptive insights into the model's predictions for different image-caption pairs. In the first two data, the model accurately predicted the emotion as "Angry" and "Sad" respectively. However, the error analysis also uncovers instances where the model encountered difficulties in making accurate predictions. In the third example, the model incorrectly predicted the emotion as "Sad". The error analysis indicates that the limited amount of data available for the "Disgust" class played a role in hindering the model's ability to make a precise prediction.



(a) Classification report



(b) Confusion matrix

Fig. 2. Quantitative performance analysis of the proposed multimodal (InceptionV3⊕BiLSTM).

VI. Conclusion

To conclude, this work focuses on the important task of classifying emotions in Bangla social media posts. The

TABLE VI: Output Analysis. The symbol (✓) indicates the correct predictions.

Image	Caption	Actual Label	Predicted Label	Reason
	নবীজির সুন্নাহ কে অপমান করা মানে আমার নবীকে অপমান করা। ছাত্রলীগ হাকে আর যেই হাকে। বিচার হবে একদিন, আল্লাহ সব দেখছেন (Insulting the Prophet's Sunnah means insulting my Prophet. Be it Chatra League or whoever. Judgment will come one day, God is watching everything)	Angry	Angry	✓
	ভুল সময়ে জন্ম নেয়া সেরা গালেকিপারদের একজন এডারসন। এলিসন কে জায়গা দিতে ওয়ার্ল্ড কাপের মত মঞ্চে বেঞ্চে বসে থাকা লাগে। আমরা আসলেই ভাগ্যবান যে সময়ের দুই সেরা গালেকিপার একসাথে পেয়েছি।(Ederson is one of the best goalkeepers born at the wrong time. Allison needs to sit on the bench on a stage like the World Cup to make room. We are indeed lucky to have two of the best goalkeepers of all time together)	Sad	Sad	✓
	একদিকে ১৫ কাটি টাকা খরচ করে কনসার্ট করে, অন্য দিকে মানুষ খাদ্যের জন্য টিসিবির গাড়ির পিছনে লাইন ধরে। সাধ্বীনতার ৫০ বছরে সাধ্ব নিচ্ছে মানুষ।(15 crores spent on concerts on one side, on the other side people lined up behind TCB vehicles for food. People are enjoying 50 years of independence.)	Disgust	Sad	Due to insufficient data in disgust class, it's hard to relate this pair to disgust class

research provides a multimodal dataset containing 4660 pairs of images and captions, representing seven emotion classes including happy, angry, disgust, fear, sad, and other. A fusion technique that combines InceptionV3 and BiLSTM as base classifiers is proposed and demonstrates exceptional performance in accurately identifying and categorizing emotions. The early fusion approach, which combines hidden layers from multiple modalities, outperforms existing machine learning and deep learning benchmarks. With a weighted f1-score of 77.50%, the model effectively recognizes and classifies various forms of emotions in Bangla posts. This methodology contributes to the advancement of emotion classification in Bangla and lays the groundwork for developing an online system that can efficiently filter posts on different internet platforms based on user requirements. However, further research is needed to address limitations such as including more emotion classes, expanding the dataset, addressing class imbalances, and exploring automated approaches for tuning hyperparameters.

References

[1] X. Yang, S. Feng, D. Wang, and Y. Zhang, "Image-text multimodal emotion classification via multi-view attentional network," *IEEE Transactions on Multimedia*, vol. 23, pp. 4014–4026, 2021.

[2] A. Illendula and A. Sheth, "Multimodal emotion classification," in *Companion Proceedings of The 2019 World Wide Web Conference*. ACM, may 2019. [Online]. Available: https://doi.org/10.1145%2F3308560.3316549

[3] T. Parvin and M. M. Hoque, "An ensemble technique to classify multi-class textual emotion," *Procedia Computer Science*, vol. 193, pp. 72–81, 2021, 10th International Young Scientists Conference in Computational Science, YSC2021, 28 June – 2 July, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050921020494

[4] K. I. Islam, M. S. Islam, and M. R. Amin, "Sentiment analysis in bengali via transfer learning using multi-lingual bert," *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pp. 1–5, 2020.

[5] F. Bianchi, D. Nozza, and D. Hovy, "XLM-EMO: Multilingual emotion prediction in social media text," in *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 195–203. [Online]. Available: https://aclanthology.org/2022.wassa-1.18

[6] A. Das, O. Sharif, M. M. Hoque, and I. H. Sarker, "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*. Online: Association for Computational Linguistics, Jun. 2021, pp. 150–158. [Online]. Available: https://aclanthology.org/2021.naacl-srw.19

[7] M. A. Iqbal, A. Das, O. Sharif, M. M. Hoque, and I. H. Sarker, "Bemoc: A corpus for identifying emotion in bengali texts," *SN Comput. Sci.*, vol. 3, no. 2, jan 2022. [Online]. Available: https://doi.org/10.1007/s42979-022-01028-w

[8] M. M. R. Mamun, O. Sharif, and M. M. Hoque, "Classification of textual sentiment using ensemble technique," *SN Computer Science*, vol. 3, 2021.

[9] A. Das, A. Iqbal, O. Sharif, and M. Hoque, *BEmoD: Development of Bengali Emotion Dataset for Classifying Expressions of Emotion in Texts*, 02 2021, pp. 1124–1136.

[10] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 527–536. [Online]. Available: https://aclanthology.org/P19-1050

[11] E. Hossain, O. Sharif, and M. M. Hoque, "MemoSen: A multimodal dataset for sentiment analysis of memes," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 1542–1554. [Online]. Available: https://aclanthology.org/2022.lrec-1.165

[12] S. Zhao, G. Jia, J. Yang, G. Ding, and K. Keutzer, "Emotion recognition from multiple modalities: Fundamentals and methodologies," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 59–73, 2021.

[13] S. Gundapu and R. Mamidi, "Gundapusunil at SemEval-2020 task 8: Multimodal memotion analysis," in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, Dec. 2020, pp. 1112–1119. [Online]. Available: https://aclanthology.org/2020.semeval-1.147

[14] P. Ekman, "Facial expression and emotion," *Am. Psychol.*, vol. 48, no. 4, pp. 384–392, Apr. 1993.

[15] Y. Zhou, Y. Yang, H. Liu, X. Liu, and N. Savage, "Deep learning based fusion approach for hate speech detection," *IEEE Access*, vol. 8, pp. 128 923–128 929, 2020.