Ensemble and Transformer Models for Emotion Recognition in Bengali Consumer-Goods E-Commerce Comments

Md. Abdul Amin[1], Habiba Begum[1], Kazi Md. Jahid Hasan[1], and Md. Jalal Uddin Chowdhury[1,*]

Department of Computer Science and Engineering, Leading University, Sylhet 3100, Bangladesh
mdabdulamin11@gmail.com, habibabegum992@gmail.com, jahidce@lus.ac.bd, jalal_cse@lus.ac.bd

Abstract. The rapid growth of the internet and e-commerce in Bangladesh has generated a vast repository of customer reviews in Bengali across domains such as foods, fashions and electronics. It is important to understand what users say in these multilingual user opinions, which leads to a better understanding of customer satisfaction and improved service quality. But most of the sentiment analysis language models are for English only, which misses out on a major portion of fine-level emotions, indicating the need to be developed in the case of Bengali text. This paper introduces a complete Bengali emotion recognition framework for online business that leverages contextual embeddings from BanglaBERT and integrates a soft-voting ensemble of classical machine learning classifiers (Logistic Regression, Random Forest and Multilayer Perceptron). The dataset contains three emotion classes (Negative, Neutral, Positive) and a balanced 3972-comment corpus constructed through preprocessing, human annotation, and inter-annotator agreement validation. Four experimental pipelines were evaluated: (1) traditional machine-learning classifiers (Logistic Regression, Random Forest and Multilayer Perceptron); (2) BanglaBERT embeddings combined with these classifiers; (3) IndicBERT embeddings combined with the same set of classifiers; (4) the proposed soft-voting ensemble integrating Logistic Regression, Random Forest and Multilayer Perceptron. BanglaBERT achieved the highest performance of 93.33% when combined with Random Forest, and the proposed Voting Ensemble model obtained the highest macro-robustness F1-score of 0.90, offering the most consistent performance across all classes. Macro F1 is selected as the primary metric due to its balanced evaluation in multiclass settings. The results validate that the stacking diverse classifiers (Logistic Regression, Random Forest, MLP) on top of contextual embeddings improves robustness and interpretability for multi-domain sentiment analysis in low-resource languages. The proposed framework provides a scalable platform for developing automated sentiment and emotion analysis in Bengali e-commerce that may have implications for recommender systems, consumer analytics, and cross-language applications.

Keywords: Bengali Emotion Classification, E-commerce Feedback, Bengali Sentiment Analysis, Ensemble Methods, Transformer Models, BanglaBERT.

M. S. Arefin et al. (eds.), *Proceedings of the International Conference on Intelligent Data Analysis and Applications (IDAA 2025)*, Advances in Intelligent Systems Research 206,
https://doi.org/10.2991/978-94-6239-664-7_39

1 Introduction

The online business sector in Bangladesh has rapidly expanded to cover a wide range of consumer goods, including food, clothing, electronics, cosmetics, and digital products. Economic growth and entrepreneurial opportunities have been significantly increased, supported by rising internet penetration, mobile device usage, and enhanced consumer confidence in online shopping. Robust annual growth – estimated to exceed 30%-has been observed, reflecting shifts in consumption patterns and the emergence of new digital platforms [1].
The rapid expansion of online business has introduced several challenges for both buyers and sellers [2]. The accurate classification of consumer emotions has been identified as crucial for understanding sentiment, satisfaction, trust, and complaints across various product categories—not only food but also clothing, electronics, and daily necessities. Existing sentiment analysis approaches have been found to focus primarily on English or limited domains, leaving a gap in nuanced emotion detection for Bengali, particularly in multi-category e-commerce feedback [3].

Although some basic sentiment analysis tools have been developed, most have been found to be either language-independent or limited to simple positive/negative polarity. Consequently, the detection of more subtle emotions such as trust, disappointment, or excitement relevant to product and service experiences in Bengali has rarely been achieved [4]. Transformer-based deep learning models, including BanglaBERT, and ensemble-based techniques have recently demonstrated potential for domain-specific applications, particularly in the online business sector [5]. However, comprehensive solutions addressing heterogeneous online business categories remain underdeveloped in both academic research and industrial practice.

The main objective of our study is to develop a reliable framework for fine-grained emotion analysis in an imbalanced monolingual Bengali e-commerce review dataset by incorporating a soft voting ensemble and transformer-based emotional learning methods. Given the linguistic diversity and low-resource nature of Bengali, our proposed framework acts as a bridge between handcrafted feature engineering and sophisticated contextual representations for emotion classification across several product domains. The modeling purpose was established on a balanced and systematically curated dataset of 3,972 comments in the Bengali language that was prepared with thorough preprocessing followed by manual annotation and inter-annotator agreement to validate the reliability of labeling. This study makes the following key contributions:

1. We created a balanced and domain-rich Bengali emotional dataset for the e-commerce review domain to fill an important resource gap in low-resource language research.
2. A comparative study was performed between the classical feature extraction methods and transformer-based contextual embeddings (BanglaBERT, IndicBERT) to evaluate their performance in Bengali emotion recognition.
3. A soft-voting ensemble model that jointly combines heterogeneous classifiers and achieves a higher macro-robustness F1 score, class-level balance, and robustness than the individual learners.
4. Establishing a scalable foundation for future cross-lingual emotion recognition research and low-resource language processing in South Asian contexts.

2 Literature Review

Some recent studies have presented different techniques for sentiment and emotion analysis of Bengali text, including comments and opinions of users on several online platforms and application domains.

Hasan et al.(2024)[6], a Bangla aspect –based sentiment analysis dataset was built for four domains: car, Mobile Phone, Movie and restaurant. Comments were collected, manually annotated into positive and negative categories by three annotators, and decided using majority vote. Comments were domain-wise organised and annotated, covering performance, design, story, food, and related aspects (1149, 975, 800 and 1149 comments, respectively). The BANGLA_ABSA dataset was constructed using these labelled comments. Naive Bays, Logistic Regression, and SVM were applied, with SVM achieving the best performance. The dataset was domain-depended, and comment length (up to 35 words) limited the capture of long-term dependencies.

Hassan et al.(2025)[7], social media comments in Bangla related to crime were studied, and sentiments were classified as positive, negative, and neutral. A dataset of 28,528 Bangla comments from social media was collected and prepared for sentiment analysis. The XLM-RoBERTa Base transformer was applied and achieved 97% accuracy, with LIME used to explain keyword contributions. The dataset was mainly urban, errors happened in mixed sentiments, and the model was limited to text, excluding audio or video.

Al Amin et al.(2024)[8], sentiment polarity of Bangla food reviews was analyzed using ML and DL methods to support user decisions and provide business insights. A dataset of 1,484 Bangla reviews was collected from FoodPanda and HungryNaki and labeled as positive or negative, features were extracted using TF-IDF (uni-/bi-gram), count vectorizer, and Bangla pre-trained GloVe vectors. ML models (Logistic Regression, Decision Tree, Random Forest, Multinomial NB, KNN, Linear & RBF SVM, SGD) and DL models(LSTM, GRU) were applied. Logistic Regression with TF-IDF unigram features achieved the highest accuracy (90.91%). Dataset size was small for deep learning, only binary sentiment was considered, manual data collection may include bias.

Poyoi et al.(2023)[9], Tourists' local food experience before and during COVID-19 were analyzed using NLP to study changes in sentiment and experience sharing in online reviews. 35,001 Google Local Guide reviews were collected via web scarping and divided into pre-pandemic (22,561) and pandemic (12,485) periods, text was cleaned and preprocessed using Python (Pandas, NumPy, NLTK, SpaCy). Sentiment was scored using TextBlob. Words were generated for frequent terms, bigram features were extracted with CountVectorizer, exploratory analysis focused on describing patterns, not prediction. Only Goggle reviews were used, dataset was smaller during pandamic, bigram analysis sensitive to word order, findings limited to one area.

Carrasco & Dias.(2024)[10], customer sentiment in restaurants was analyzed using ABSA and NLP techniques to support management decisions. 880,000 TripAdvisor reviews from 1,581 Algarve restaurants(2010-2023) were collected, 1,000 reviews were manually labeled by three human groups for benchmarking, text was cleaned and prepared for analysis. BART, and ChatGPT were applied for automated sentiment and attribute extraction, F1 scores of 0.86-0.92 were achieved, ChatGPT gave the highest accuracy. The RestMON Algarve web app was used to track sentiment trends over time. Only 1,000 reviews used for testing, study limited to Algarve, more reviews from other places needed to check results.

Zahoor et al.(2020)[11], customer reviews of Karachi restaurants were studied using ML and NLP to find sentiment and classify feedback by food, service, ambiance, and value for money. About 4,000 English reviews from a Facebook food group were collected, text was cleaned with tokenization, stop-word removal, lemmatization, stemming, and TF-IDF using NLTK. Naive Bays, Logistic Regression, SVM, and Random Forest were applied. Random Forest gave the best accuracy (95%), food taste was easiest to classify, ambiance/value harder. Only English reviews from one Facebook group , deep learning and more languages suggested for future work.

Asani et al.(2021)[12], a system was made to suggest restaurants based on user reviews and food preference using sentiment analysis. Reviews from 100 TripAdvisor users were collected. Food words were picked and grouped using Wu-Palmer similarity. Sentiment scores were calculated with SentiWordNet. Low quality restaurants were removed. A hybrid method was combining content-based and collaborative filtering. Cosine similarity matched user tastes with menus, Location, time, and restaurant status were also considered. Only small dataset was used. Accuracy drops with more nearby restaurants or longer distances. Future work can use bigger datasets, group recommendations, and multiple languages.

Bhuiyan et al.(2020)[13], sentiment of food reviews was analyzed using deep learning. CNN with attention was used to classify reviews as positive or negative. 1600 Food Panda reviews were collected and labeled. Spelling was corrected, words were simplified, special characters removed, and text was tokenized. CNN, LSTM, and CNN with attention were tested. Attention highlighted important words. CNN with attention gave the best accuracy (98.45%) and F1-score(0.94). Small dataset and simple word embeddings were used. Bigger models and more data could improve results.

Hasan et al.(2020)[14], sentiment classification for Bangla text was studied using classical, deep learning, and transformer models. Multiple datasets (tweets, reviews, social media posts) were collected, cleaned, and split into train, validation, and test sets. SVM, Random Forest, CNN, FastText, and transformers (BERT, XLM-RoBERTa) were tested. Transformers performed best, with BERT and XLM-RoBERTa achieving highest F1 scores. Small datasets and high computation needs were noted. Future work can use larger datasets and more efficient models.

Shamael et al.(2024)[15], a large Bangla-English and code-mixed review dataset, BanglishRev, was created with 1.74M reviews from Daraz Bangladesh. Reviews and metadata (ratings, dates, likes, dislikes, seller replies) were collected. BanglishBERT was trained for binary sentiment classification, achieving 94% accuracy and F1 0.94. Data was imbalanced, from one platform only.

Finally, the prior works have made commendable progress on Bengali sentiment analysis which includes approaches based on traditional machine learning as well as transformer-based models. However, most of the existing works are domain dependent, applicable to only binary sentiment polarity and cannot perform fine-grained emotion classification across various e-commerce domains.

3 Data Analysis

This section provides a detailed explanation of the methods applied for data collection, preprocessing, and annotation. A systematic workflow was employed careful curation and organization following a consistent protocol to guarantee accurate and dependable experimental results.

3.1 Data Collection

In this investigation, a supervised dataset containing Bengali online business emotion was adopted. the dataset aimed to provide reliable input for both model training and performance assessment. The dataset comprised Bengali-language-online-business-related comments and phrase gathered from multiple social media sources such as Facebook and YouTube. Every comment reflected a specific emotional tone based linked to online business choices or consumption experiences. The original dataset contained 2,044 labeled samples, distributed across three target sentiment categories: Positive(412), Negative(308), and Neutral(1324). Once balanced using Random Oversampler, each class reached 1324 samples , producing a 3,972- entry dataset ideal for fair evaluation. Table I presents the comparative distribution of samples and after balancing.

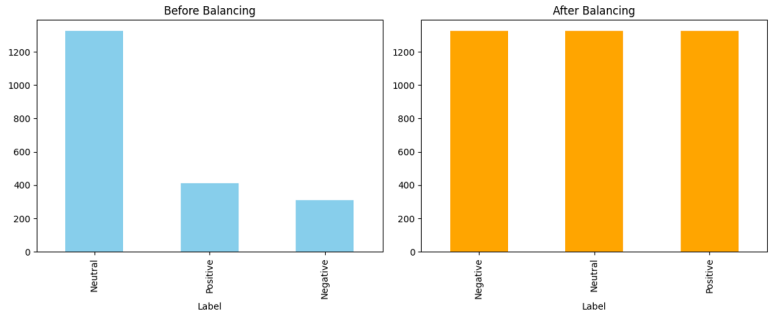

Fig. 1. Class Distribution Diagram

3.2 Data Preprocessing

To improve data quality and consistency, the text underwent preprocessing prior to annotation and modeling. Each sentence was standardized by lowercasing and filtering out irrelevant symbols and non-Bengali or Banglish scripts. Empty or duplicate entries were excluded to maintain dataset integrity. In addition, label encoding was performed to convert the categorical emotion tags(Positive,Negative,Neutral) into numerical form. facilitating compatibility worth standard machine learning tools. This preprocessing stage ensured consistency, reproducibility, and analysis-ready data.

3.3 Data Annotation

The annotation process was conducted collaboratively by two human annotators, both of whom were trained to distinguish emotional tones specific online-business-related expressions. Through discussion and samples-based calibration , both annotators ensured consistent interpretation of

Table 1. Class Definitions, Sample Comments, and Sample Counts in Bengali Online Business Emotion

Class (Label)	Definition and Sample Comment	Count
Negative (0)	Negative emotions such as disappointment, dislike, or criticism toward food quality, taste, or service. *Example:* , (This biryani tastes awful, only oil and spice smell are there.)	1324
Neutral (1)	Emotionally balanced or factual statements expressing neither strong satisfaction nor dissatisfaction. *Example:* , (Went to a new restaurant today, the price of rice and fish was moderate.)	1324
Positive (2)	Positive emotions such as satisfaction, delight, or appreciation for food taste, service, or ambiance. *Example:* , (The food at this restaurant is excellent, the chicken is very soft and delicious.)	1324

emotion categories. Following manual labeling , a one-shot validation was performed using ChatGpt to verify consistency and annotation reliability as a hot topic in world wide and used in annotation. To disambiguate between the *Negative* and *Positive* sentiment categories, the annotation guidelines adopted a simple stance-based cue: **Negative** = content expressing dislike, criticism, or unfavorable emotion toward online business platform such as food and cloth: **Positive** = content showing satisfaction, appreciation, or favorable emotion towards the online business products.See Table 1 for the label mapping, definitions, and representative examples.To assess labeling consistency, both human annotators and the large language model (*ChatGPT*) were evaluated using Cohen's Kappa coefficient [**?**]. The resulting scores—**0.704** for human annotation and **0.658** for ChatGPT validation—reflect a substantial level of agreement and confirm the reliability of the final annotated dataset.

Each class has different comment lengths, with class 0 (Negative) containing comments with the maximum words and class 2 (Positive) having the shortest. This comparison is shown in Table 2.

Table 2. online business Comments Word Count by Class

Class	Maximum Length	Minimum Length	Average Length
Negative (0)	42	3	11
Neutral (1)	38	2	9
Positive (2)	35	2	8

3.4 Emotion Classes

The final dataset includes three emotion classes: 1. Negative(0) : dissatisfaction, criticism, complaint 2. Neutral(1): factual, mixed, or emotion-less statements 3. Positive(2): satisfaction, praise, recommendation

Although the term "fined-grained" indicates nuanced emotional expressions in text, e-commerce comments are generally characterized by evaluative sentiment rather than specific affective subclass such as fear, anger, or disgust. Therefore, only these three categories are considered in this study.

4 Methodology

In this study, we conducted a systematic analysis and comparison of supervised machine learning models for sentiment analysis in Bangla-language reviews about restaurants. The overall goal is to compare traditional feature engineering approaches against transformer based contextual embeddings for sentiment classification of low resource languages. To accomplish this objective, the methodology is composed of various interconnected phases such as holistic non-linearity passed label encoding followed by dataset balancing permitting feature extraction, model training and a rigorous performance evaluation across four experimental setups.

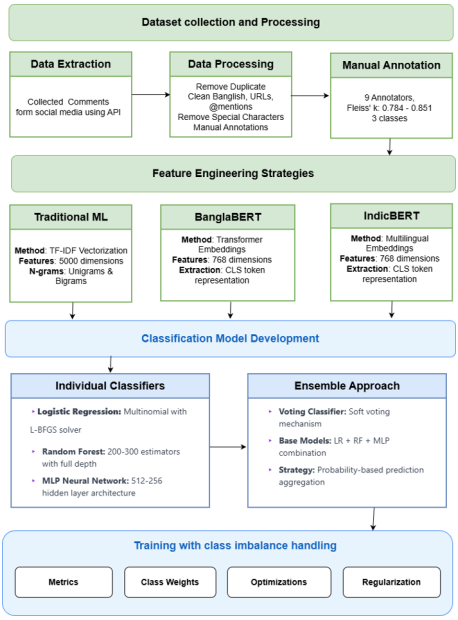

Fig. 2. Overall Proposed Workflow Diagram

4.1 Experimental Design and Feature Engineering

The purpose of this experimental framework was to investigate several feature extraction and classification approaches in a controlled and reproducible context, comparatively. The data was split in an 80/20 train–test stratified method where the class distribution was preserved for each of the two sets. The experiments were performed with 4 different configurations in order to examine the feature representation and classifier ensemble strategy combination variation.

In the first experiment, a classic model with TF–IDF was used as baseline. TfidfVectorizer from Scikit-learn was used with maximum vocabulary size set to 5,000 terms and an n-gram range between (1,2) to cover both unigram and bigram level information. This configuration forced the model to be able to learn both single word-level semantics and short-range contextual dependencies, whilst still being tractable in computation.

The second experiment was with BanglaBERT, a monolingual transformer model pre-trained only on Bangla corpora. Using 'csebuetnlp/banglabert' model of Hugging Face Transformers library, each review was tokenized and converted to a 768-dimensional contextual embedding which represents the final hidden state for [CLS] token. These embeddings represent the semantics of a reviewer's opinion as it lives at the level of a sentence. For numerical stability, the embeddings were standardized by zero mean and unit variance to facilitate downstream learning using Scikit-learn's StandardScaler.

The third experimental set-up used IndicBERT, a multilingual transformer model pre-trained on twelve major Indian languages including Bangla. We used the ai4bharat/indic-bert model for generation of sentence level embeddings in a similar manner as we presented for BanglaBERT setup. The embeddings obtained were also normalized in order to remain consistent across conditions. This experiment helped us explore cross-lingual generalization of transformer-based embeddings in multilingual settings.

In the fourth and last experimental setting, we investigated an ensemble learning technique where we created a Voting Classifier on top of embeddings extracted from BanglaBERT. In this approach, three base classifiers namely Logistic Regression, Random Forest and Multilayer Perceptron (MLP) were used in a soft voting ensemble, which aggregating the probabilistic predictions of each classifier to make sentiment label classification. -Soft voting instead of Hard voting was implemented to make the ensemble take into account how confident were perturbation effects generated by base classifiers and therefore increase robustness and stability of prediction.

4.2 Classifier Training and Optimization

All experiments, except the ensemble setting, used a shared base of three supervised machine learning algorithms: Logistic Regression (LR), Random Forest (RF), and Multilayer Perceptron (MLP). Each model was tuned in order to maximize the performance on its feature set.

The Logistic Regression classifier was configured as a multinomial model with the lbfgs optimization solver, and one maximum iteration limit of 1000 to guarantee convergence. This algorithm was selected for its efficient and interpretable feature learning on the high dimensional feature spaces like TF-IDF. For TF–IDF features, 200 decision trees and for transformer-based embeddings 300 decision trees were used in the Random Forest classifier to extract complex nonlinear decision boundaries. The loss criterion was the entropy that also favors well diverse decision paths in ensemble.

The Multilayer Perceptron network serving for the deep learning part of the experimental setting was composed by two hidden layers fully connected with 512 and 256 neurons, respectively. The ReLU was used as an activation function to provide non-linearity and the Adam optimizer enabled adaptive learning with quick convergence of gradients. The network was trained up to 100 epochs with early stopping based on validation loss for avoiding overfitting.

To ensure reproducibility, random seed=42 was fixed for Numpy, PyTorch, and Scikit-learn. All embeddings were generated using PyTorch 2.0 and HuggingFace Transformers 4.37 on a NVIDIA GPU. We used GPU-based machinery for training and evaluation to speed up the embedding generation as well as model optimization.

4.3 Proposed Ensemble Model

Results To test the potential utility of fusion, an ensemble learning approach was implemented based on the Voting Classifier meta-estimator from Scikit-learn. The ensemble combined three base learners – Logistic Regression, Random Forest and Multilayer Perceptron, separately trained on the centered BanglaBERT embeddings computed from second Experiment. We used 'soft' on voting the ensemble calculated class output prediction by averaging the predicted probabilities of all base classifiers. This probabilistic averaging helped the ensemble to take into account the confidence of each individual model and thus mitigating the impact of outlier predictions while increasing overall decision stability.

Soft voting is chosen instead of hard voting because it usually achieves higher performance in multiclass problems as taking advantage of the complementary strengths of heterogeneous models. Logistic Regression brought linear decision boundaries and interpretability; Random Forest strength against noise and feature variance; and MLP the ability to model complex nonlinear relationships over embedding space. Having these various types of learners allowed the ensemble to generalize more effectively over the feature terrain.

The ensemble model was tested with the same protocol as for the individual classifiers, including the same training and testing partitions. Its performance was then compared with that of the individual models to determine if collaborative decision-making can have a tangible impact in terms of classification accuracy, precision, recall and F1-score. This combined ensemble approach symbolized the end of the proposed experimental design, providing insight for a synergistic relationship between traditional machine learning and recent deep learning methodologies in Bangla sentiment analysis.

Each classifier receives the 768-dimensional BanglaBERT embedding and outputs class probabilities. Soft voting these probabilities to produce the final prediction.

5 Result

5.1 Performance Measure for Traditional Machine Learning

In traditional machine learning approach, three models were selected- Logistic Regression, Random Forest, and MLP and TF-IDF along with Label Encoder were applied. An accuracy of approximately 81.13% was achieved by Logistic Regression, and accuracy of about 85.91% was recorded for Random Forest, while MLP classifier was measured with an accuracy of around 83.90%.

Random Forest achieved the highest accuracy. At the highest class level, Negative sentiment was detected with excellent performance of f1 score 0.92. However, Neutral sentiment f1 score 0.82 and Positive sentiment f1 score 0.84 remain challenging, indicating some confusion occurred when classifying mixed or unclear content. Further details are presented in Table 3.

Table 3. Performance Measure for Traditional Machine Learning Models

Models	Accuracy	Precision	Recall	F1-score (Macro Avg)
Logistic Regression	0.8113	0.82	0.81	0.81
Random Forest	0.8591	0.87	0.86	0.86
MLP	0.8390	0.84	0.84	0.84

5.2 Model Performance Using BanglaBERT and IndicBERT

Transformers like BanglaBERT and IndicBERT were evaluated for Bangla sentiment classification. These transformer models were used for converting text to numeric representations. Both models outperformed traditional machine learning approaches in capturing contextual semantics of Bangla text.

IndicBERT was evaluated, and Random Forest obtained the highest accuracy of 91.95%, which is slightly lower than BanglaBERT. Logistic Regression achieved 78.87% and MLP reached 86.54%. At the highest class level, Negative sentiment was detected efficiently with an f1 score of 0.97, while Neutral and Positive Sentiments showed f1 scores between 0.89 and 0.90, indicating some inconsistency. A detailed breakdown of the results is provided in Table 4.

Table 4. Performance Using IndicBERT for Sentiment Classification

Models	Accuracy	Precision	Recall	F1-score (Macro Avg)
Logistic Regression	0.7887	0.79	0.79	0.78
Random Forest	0.9192	0.92	0.92	0.92
MLP	0.8654	0.87	0.87	0.86

BanglaBERT achieved an accuracy of 93.33% using Random Forest model, while Logistic Regression reached 84.40% and MLP around 88.55%. At the highest class level, Negative sentiment was detected very accurately with an f1 score of 0.97, whereas Neutral and Positive sentiments achieved f1 scores of 0.90 and 0.92, showing slight variability. Complete results are shown in Table 5.

Table 5. Performance Using BanglaBERT for Sentiment Classification

Models	Accuracy	Precision	Recall	F1-score (Macro Avg)
Logistic Regression	0.8440	0.84	0.84	0.84
Random Forest	0.9333	0.93	0.93	0.93
MLP	0.8855	0.89	0.89	0.89

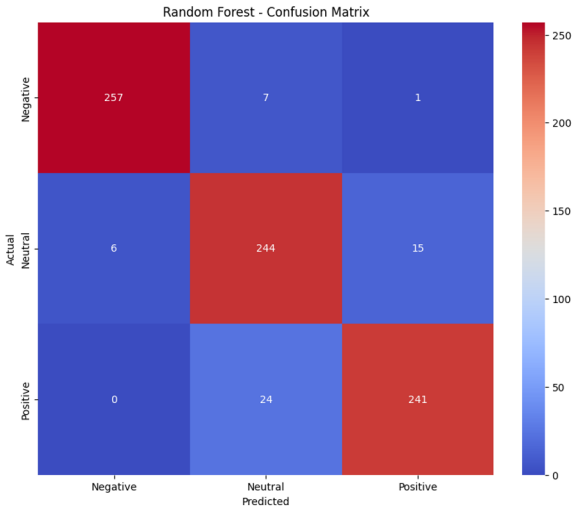

Fig. 3. Confusion Matrix for Random Forest

5.3 Proposed Model Using Ensemble Technique

The Voting Ensemble model was evaluated for Bangla Sentiment classification and achieved an overall accuracy of 89.94%, demonstrating strong performance across all classes. At the class level, Negative sentiment was detected with excellent precision, recall, and f1 score of 0.97, indicating the model's ability to identify clear negative samples consistently. Neutral sentiment achieved an f1 score of 0.85, with precision and recall of 0.85 and 0.86, showing improved stability compared to traditional ML models where Neutral was often the weakest. Positive sentiment reached an f1 score of 0.88, with precision and recall values of 0.89 and 0.87, reflecting reliable detection of positive samples. The macro average f1-score also reached 0.90, confirming that the model balanced performance across all sentiment classes.

Compared to traditional ML models and transformer-based models, the Voting Ensemble demonstrated the most consistent performance across all sentiment classes, especially improving the detection of Neutral and Positive sentiments. Complete class-wise metrics are summarized in Table 5 and Table 4.

Table 6. Performance of Voting Ensemble Model

Models	Accuracy	Precision	Recall	F1-score (Macro Avg)
Proposed Ensemble Model	0.8994	0.90	0.90	0.90

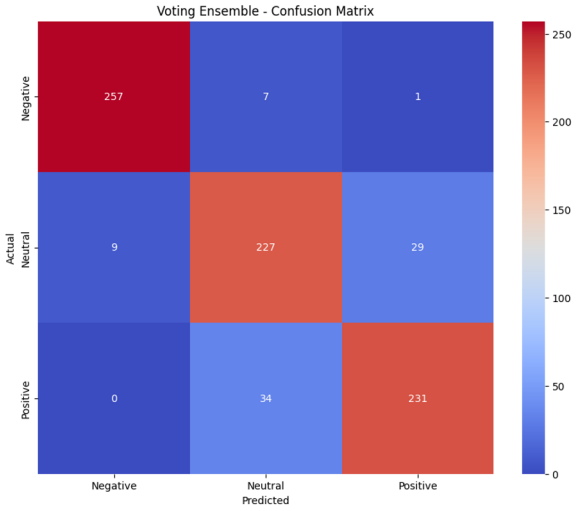

Fig. 4. Confusion Matrix for Voting Ensemble

6 Discussion

The comparative results indicate a clear gap in performance between the traditional feature-based classifiers and transformer-based representation for Bengali emotion classification within e-commerce contexts. Classical TF–IDF models had moderate performance but they were not able to capture implicit sentiment drifts and domain-specific language nuances, especially in Neutral class. This echoes previous findings regarding the limitations of mostly sparsely encoded features in low-resource languages to which contextual and morphological variation is crucial.

Transformer embeddings, particularly BanglaBERT, showed a significant improvement establishing the advantage of contextualized language representations as for morphologically rich languages such as Bengali. The BanglaBERT–Random Forest ensemble had the best accuracy 93.33%, showing that using deep contextual representation with non-linear ensemblers is one of the good choices if you have moderate size of data. It not only achieved competitive performance but also exhibited absolute below BanglaBERT, demonstrating the effectiveness of monolingual pre-training on domain-aligned tasks. The soft-voting ensemble had the best trade-off in macro-performance where F1 is 0.90. Even though it didn't go above and beyond BanglaBERT–RF in raw accuracy, its reliability over all the classes, Neutral and Positive notably, shows higher

robustness. Complementary strengths of LR, RF and MLP versioned overfitting and class-level sensitivity dropping which are essential in practical e-commerce applications with the diverse set of sentiment.

The results as a whole suggest that the hybrid architectures based on transformer embeddings with heterogeneous ensemble learners discussed in this paper are capable of significantly improving emotion classification performance on low-resource cases. This suggests promising applications in scalable sentiment-aware business analytics and multilingual downstream systems. In terms of future work, a host of extensions are possible. The extensions to include richer linguistic features, build fusion-based transformer architectures as well as consider the adaptive class-weights strategies for imbalance handling signify promising directions to improve performances and design a more accurate domain-independent emotion recognition system.

7 Conclusion

In this paper, we proposed an end-to-end framework for fine-grained Bengali emotion recognition in e-commerce reviews based on mixture of transformer based and ensemble learning. By systematic experimentation with TF–IDF, BanglaBERT, IndicBERT and soft voting ensemble the research illustrated how contextual embeddings outperform traditional feature engineering in managing subtle emotional expression in Bengali text. The experimental results showed that BanglaBERT with Random Forest attained the highest accuracy of 93.33% and the ensemble model obtained a macro-average F1-score of 0.90, which evidences the strong positive generalization on all negative, neutral and positive classes. These results, illustrate the utility of combining heterogeneous classifiers -Logistic Regression, Random Forest and MLP- in preventing overfitting as well as improving prediction stability within multilingual settings. The results have clear implications for the design of e-commerce websites, by enabling better customer experience analytics, product recommendation systems and feedback-based business strategies. In future investigations, the framework could be expanded by domain adaptation, extend large amount of dataset, incorporating multimodal features like audio or images, or adding large scale generative transformer models for improving emotion understanding in under-resourced languages.

References

1. Łobejko, S., & Bartczak, K. (2021). The role of digital technology platforms in the context of changes in consumption and production patterns. Sustainability, 13(15), 8294.
2. Sikder, A. S., & Rolfe, S. (2023). The Power of E-Commerce in the Global Trade Industry: A Realistic Approach to Expedite Virtual Market Place and Online Shopping from anywhere in the World.: E-Commerce in the Global Trade Industry. International Journal of Imminent Science & Technology., 1(1), 79-100.
3. Sakib, S., Chowdhury, M. J. U., Hira, M. M. H., Choudhury, T. E., & Das, S. (2025, April). Deep Learning-Based Sentiment Analysis of Social Media and E-Commerce Reviews. In 2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI) (pp. 1-5). IEEE.
4. Khare, B. K., & Khan, I. (2024, March). Transforming Emotions: A Comprehensive Review of Text Emotion Detection with Transformer Models. In International Conference on Emerging Trends and Technologies on Intelligent Systems (pp. 515-534). Singapore: Springer Nature Singapore.

5. Mereddy, D., & Beedareddy, J. S. R. (2024, December). Enabling Next-Generation Smart Homes Through Bert Personalized Food Recommendations-RecipeBERT. In 2024 IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (pp. 796-803). IEEE.
6. Hasan, M., Ghani, M. R., & Hasan, K. A. (2024). Aspect based sentiment analysis datasets for Bangla text. *Data in Brief, 57*, 111107.
7. Hassan, F. B., Jubair, M. A., Hasan, M. M., Hossain, T., Shuvo, S. M., & Arefin, M. S. (2025). Understanding Public Perception of Crime in Bangladesh: A Transformer-Based Approach with Explainability. *arXiv preprint arXiv:2507.21234*.
8. Amin, A., Sarkar, A., Islam, M. M., Miazee, A. A., Islam, M. R., & Hoque, M. M. (2024). Sentiment polarity analysis of Bangla food reviews using machine and deep learning algorithms. *arXiv preprint arXiv:2405.06667*.
9. Poyoi, P., Gassiot-Melian, A., & Coromina, L. (2023). Local food experiences before and after COVID-19: a sentiment analysis of EWOM. *Tourism and Hospitality Management, 29*(4), 477–493.
10. Carrasco, P., & Dias, S. (2024). Enhancing restaurant management through aspect-based sentiment analysis and NLP techniques. *Procedia Computer Science, 237*, 129–137.
11. Zahoor, K., Bawany, N. Z., & Hamid, S. (2020, November). Sentiment analysis and classification of restaurant reviews using machine learning. In *2020 21st International Arab Conference on Information Technology (ACIT)* (pp. 1–6). IEEE.
12. Asani, E., Vahdat-Nejad, H., & Sadri, J. (2021). Restaurant recommender system based on sentiment analysis. *Machine Learning with Applications, 6*, 100114.
13. Bhuiyan, M. R., Mahedi, M. H., Hossain, N., Tumpa, Z. N., & Hossain, S. A. (2020, July). An attention based approach for sentiment analysis of food review dataset. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)* (pp. 1–6). IEEE.
14. Hasan, M. A., Tajrin, J., Chowdhury, S. A., & Alam, F. (2020, December). Sentiment classification in Bangla textual content: A comparative study. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)* (pp. 1–6). IEEE.
15. Shamael, M. N., Nawshin, S., Shatabda, S., & Islam, S. (2024). BanglishRev: A Large-Scale Bangla-English and Code-mixed Dataset of Product Reviews in E-Commerce. *arXiv preprint arXiv:2412.13161*.