I.J. Information Engineering and Electronic Business, **2025, 3, 49-61**
Published Online on June 8, 2025 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2025.03.04

Enhancing E-commerce Sentiment Analysis with Advanced BERT Techniques

Nusrat Jahan*
American International University-Bangladesh/Computer Science and Engineering, Dhaka, Bangladesh
E-mail: nusrat.jahan@aiub.edu
ORCID iD: https://orcid.org/0000-0002-4958-2676
*Corresponding Author

Jubayer Ahamed
American International University-Bangladesh/Computer Science and Engineering, Dhaka, Bangladesh
E-mail: jubayer@aiub.edu
ORCID iD: https://orcid.org/0000-0003-2076-9194

Dip Nandi
American International University-Bangladesh/Computer Science and Engineering, Dhaka, Bangladesh
E-mail: dip.nandi@aiub.edu
ORCID iD: https://orcid.org/0000-0002-9019-9740

Received: 04 June, 2024; Revised: 12 January, 2025; Accepted: 25 February, 2025; Published: 08 June, 2025

Abstract: This study introduces an improved BERT-based model for sentiment analysis in several languages, specifically focusing on analyzing e-commerce evaluations written in English and Bengali. Conventional sentiment analysis techniques frequently face difficulties in dealing with the subtle linguistic differences and cultural diversities present in datasets containing multiple languages. The model we propose integrates sophisticated methodologies and utilizes Local Interpretable Model-agnostic Explanations (LIME) to enhance the accuracy, interpretability, and dependability of sentiment assessments in various language situations. To tackle the challenges of sentiment categorization in a multilingual setting, we enhance the pre-trained BERT architecture by incorporating extra neural network layers. Compared to traditional machine learning and current deep learning methods, the model underwent a thorough evaluation, showcasing its superior capabilities with accuracy, precision, recall, and F1-score of 0.92. Including LIME improves the model's transparency, allowing for a better understanding of the decision-making process and increasing user confidence. This research highlights the potential of utilizing advanced deep learning models to address the difficulties of sentiment analysis in global e-commerce environments, providing major implications for both academic research and practical applications in industry.

Index Terms: E-commerce, Sentiment analysis, Machine Learning, BERT, Explainable AI, LIME

1. Introduction

In the burgeoning field of e-commerce, customer reviews significantly influence purchasing decisions and brand reputation. These reviews, rich in sentiment and opinion, require effective analytical strategies to harness their full potential for market analysis and customer service improvement [1]. Sentiment analysis extends beyond classification to shape e-commerce strategies by refining marketing, brand perception, and customer retention. Platforms like Amazon and Alibaba leverage it for recommendation systems, automated customer service, and fraud detection. By analyzing large-scale consumer opinions, businesses can optimize pricing, product offerings, and customer experiences, making sentiment analysis indispensable in the competitive digital marketplace. The primary challenge is accurately classifying sentiments as positive or negative, especially when dealing with reviews across multiple languages [2]. This challenge is amplified by the linguistic diversity and complexity of user-generated content, which often includes colloquial language, idioms, and cultural nuances.

Existing literature on sentiment analysis has predominantly focused on single-language datasets, particularly in English, using conventional natural language processing (NLP) techniques such as bag-of-words models and simple

lexical approaches [3]. However, these methods fall short in handling the subtleties of multilingual data and fail to capture the contextual dependencies within texts [4]. Furthermore, while providing a baseline, traditional machine learning models do not adequately address the deep semantic structures necessary for understanding complex sentiments.

This research introduces an advanced computational framework designed to enhance the sentiment classification of e-commerce reviews in both English and Bengali. By integrating the Bidirectional Encoder Representations from Transformers (BERT), a cutting-edge transformer-based model developed by Google, this study leverages state-of-the-art deep. Volume XX (XX), Issue 6 learning methodologies to capture contextual nuances across different languages. BERT's mechanism, which pre-trains on a vast corpus of unlabeled text before fine-tuning on specific tasks, provides a robust foundation for understanding and analyzing intricate language patterns [5].

This research aims to adapt the BERT model for bilingual sentiment analysis meticulously and innovate upon its architecture with additional custom neural network layers tailored for this task. These enhancements aim to improve sentiment analysis's accuracy and efficiency across diverse linguistic datasets. Additionally, the research employs Local Interpretable Model-agnostic Explanations (LIME) for explainability analysis [6], thereby increasing the transparency and interpretability of the machine learning models used, which is crucial for applications in dynamic and sensitive environments like e-commerce. Overall, the contributions of this research are:

1. **Advanced BERT Integration:** By enhancing a BERT-based model with additional neural network layers, this research tackles the intricate challenge of sentiment analysis across multiple languages. This integration significantly boosts the model's ability to understand and analyze complex linguistic structures, thereby improving sentiment classification accuracy.
2. **Multilingual Analysis Capability:** Addressing the challenge of multilingual sentiment analysis, this study provides a robust solution that can accurately process and analyze sentiments in both English and Bengali. This capability is crucial for global e-commerce platforms that operate across diverse linguistic landscapes.
3. **Empirical Validation:** Through extensive testing and validation, this research not only demonstrates the model's superior performance over traditional and existing deep learning approaches but also provides a detailed analysis of its effectiveness through various performance metrics such as accuracy, precision, recall, and F1-score.
4. **Practical Implications for E-commerce:** By improving the accuracy and reliability of sentiment analysis, this model can significantly enhance how businesses interpret and respond to customer feedback on a global scale. This has direct implications for marketing strategies, customer service, and product development within the e-commerce industry.
5. **Application of LIME for Explainability:** Incorporating Local Interpretable Model-agnostic Explanations (LIME) marks a substantial advancement in making deep learning models more transparent and understandable. This addition allows users to see which aspects of the input text most influence the sentiment prediction, thereby increasing trust in the model's outputs.

The paper is divided into five sections, which are arranged in the following sequence: The initial section comprises the introduction, while the subsequent section encompasses the related work. The dataset is delineated in Section 3, providing comprehensive information on our methodology. Section 4 presents a comparative analysis of the results and discusses the findings. The text is concluded in Section 5, which addresses the subsequent steps.

2. Related Work

The field of sentiment analysis has evolved significantly over the past decade, driven by advances in machine learning and natural language processing technologies. Traditional approaches have predominantly utilized statistical techniques and basic NLP tools to discern sentiment from text, largely focusing on English-language data. Notable studies have employed methods ranging from simple polarity-based algorithms to more complex machine learning techniques such as Support Vector Machines (SVM) and Naive Bayes classifiers. These foundational methods laid the groundwork for understanding text-based sentiment but often lacked the nuance necessary to accurately interpret context and idiomatic expressions. As the digital landscape becomes increasingly multilingual, the demand for effective sentiment analysis tools that can operate across diverse linguistic datasets has grown. Recent research has begun to explore the capabilities of deep learning models, particularly those based on the transformer architecture, to address these challenges. This review critically examines the trajectory of sentiment analysis methodologies from their inception to the current state-of-the-art, highlighting the technological advancements and identifying persistent gaps in multilingual sentiment analysis that this research aims to address.

2.1. Machine Learning

Machine learning (ML) sentiment analysis methods have been extensively investigated to improve the accuracy and efficiency of textual sentiment reading. Historical emotion classification and prediction require supervised learning.

Due to their simplicity and efficacy, Naive Bayes, SVM, and Decision Trees are popular feature-based classification methods [7]. These methods use preprocessed datasets and phrase frequency and emotion lexicons to capture textual sentiments.

This research enhances data preparation and feature selection for sentiment analysis machine learning models-Shafin et al. proposed classifying user reviews using modified logistic regression sentiment analysis and language processing [8]. Traditional methods struggle to accurately detect user opinion polarity (positive, negative, and neutral) because of the review of classification uncertainty. Support count estimation and simultaneous processing of synonymous phrases improve classification precision in this novel method. The improved method outperforms existing methods on a movie review dataset with 90% classification accuracy, 78.6% precision, 75.6% recall, and 76.5% F-measure. This research enhances social media sentiment analysis by improving interpretation. Another research demonstrated Bangla NLP and ML [9]. They wanted to apply NLP to assess online customer feedback and establish a positive/negative Bangla comment ratio. Over 1000 product reviews were collected for investigation. KNN, Decision Tree, SVM, Random Forest, and Logistic Regression classification algorithms were employed for sentiment analysis. SVM outperformed all algorithms with 88.81% accuracy. This research uses Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest to identify Malayalam tweets as good or negative [10]. Features like Bag of Words (BOW), TF-IDF, and Unigram with Sentiwordnet—with and without negation words—are used to build feature vectors. The study reveals that the Random Forest classifier is most accurate when utilizing Unigram with Sentiwordnet, which includes negation terms. This research shows that specialized feature selection improves social media sentiment analysis for certain languages.

Another machine learning-based emotion categorization study uses restaurant recommendations from SWOT's Guide to Karachi's Restaurants, a prominent Facebook community [11]. Based on meal flavor, ambiance, service, and value for money, Naive Bayes, Logistic Regression, SVM, and Random Forest classify client comments as positive or bad. Random Forest is the most accurate algorithm, with 95% accuracy on 4,000 hand-annotated records. According to one study, social media comments improve corporate insights and customer service. Star ratings are used to classify IMDB movie reviews by sentiment. This article filters and suggests movies using Cosine Similarity and user preferences [12]. The system improves movie suggestions by analyzing sentiment using Na¨ıve Bayes (NB) and Support Vector Machine (SVM) supervised machine learning. Sentiment analysis helps propose movies by assessing their appeal. SVM's 98.63% accuracy exceeds NB's 97.33%. SVM is great for movie recommendation user experience due to its sentiment analysis efficiency and accuracy.

This study creates a classifier model utilizing feature extraction and ranking to classify movie reviews as favorable or negative [13]. Using NLP on the IMDB movie review database, the model addresses movie comments' informal and linguistically heterogeneous nature. Comparative comparison with existing learning models shows that the suggested approach improves movie rating systems by 97.68%. This research extends sentiment analysis and refines entertainment media consumer opinion classification. This study explores different machine learning classifiers to detect Twitter cyberbullying, a widespread but difficult-to-treat problem because of victim interaction [14]. Analyzing 37,373 tweets with Logistic Regression (LR), LGBM, SGD, Random Forest (RF), AdaBoost (ADB), Naive Bayes (NB), and SVM. Performance was measured by accuracy, precision, recall, and F1. Logistic Regression has the highest median accuracy (90.57%, F1 score 0.928). Precision and recall are best with SGD and SVM. These findings imply that machine learning can detect cyberbullying autonomously, improving intervention. Tweepy and TextBlob in Python are used to evaluate Twitter sentiment on the global COVID-19 pandemic [15]. Seven days of tweets on COVID-19 and coronavirus were collected from April 9 to 15, 2020. Graphically representing public perspectives via sentiment research showed the spectrum of emotions and opinions expressed online. The findings reveal early pandemic mood and demonstrate the value of social media data in measuring real-time global responses.

Rustam et al. analyzed using supervised machine learning [16]. After testing several classifiers, the Extra Trees model performed best with 0.93 accuracy utilizing bag-of-words and TF-IDF. Although accurate, the study largely employs Twitter data, which may limit its application across social media platforms. This paper provides advanced neural network models, including n-grams, word embeddings, and emotion indicators to detect bogus reviews [17]. Deception can be detected by evaluating semantic content and emotional emotions in reviews with these models. These attributes integrate to classify across domains and product categories. Superior performance compared to leading methodologies shows the models' potential to improve e-commerce reliability. Virtual transactions obscure fraudulent e-commerce reviews [18]. The paper discusses it. Intelligent n-grams and sentiment scores from review texts identify fake reviews. Tripadvisor hotel reviews are TF-IDF preprocessed, and feature extracted. This detection model uses four supervised machine learning methods: Na¨ıve Bayes, Support Vector Machine, Adaptive Boosting, and Random Forest, attaining 88%, 93%, 94%, and 95% accuracy and F1-score the method works because these results outperform earlier studies on the same dataset. Domain-wide validation is needed because hotel reviews may not apply to other e-commerce businesses.

2.2. Deep Learning

Deep learning automates feature engineering by learning high-level features from data using complex architectures, improving sentiment analysis. CNNs, RNNs, LSTM networks, and GRUs offer context-aware textual sentiment analysis. These models excel in sequential data processing, which is critical for capturing text nuances, where

word sequence and contextual relevance strongly affect sentiment. Deep learning can identify complex patterns in large datasets, making sentiment analysis more precise and detailed across various data sources-

The 21st-century data revolution presents problems and opportunities for businesses to exploit consumer reviews and social media comments [19]. Deep learning technique Convolutional Neural Networks (CNN) are used to analyze text sentiment. CNNs outperform SVM and Na¨ıve Bayes in recognizing feelings in large text corpora. CNN is tested on Amazon purchase reviews and IMDB movie reviews for emotion and opinion recognition. Deep learning could improve Natural Language Processing sentiment analysis for customer sentiment insights. LSTM, a form of Recurrent Neural Network (RNN), is used to analyze IMDb movie reviews for sentiment analysis in this paper [20]. Governments and corporations need sentiment analysis to understand public opinion. The research optimizes classification by preparing data. LSTM performs well in text-based sentiment analysis with 89.9% classification accuracy. This achievement shows that LSTM can analyze complicated emotional information in movie reviews to reveal consumer feelings. This research evaluates sentiment analysis in MOOC reviews using 66,000 evaluations, comparing machine learning, ensemble learning, and deep learning methods. The study finds that LSTM networks with GloVe embeddings outperform others, achieving 95.80% accuracy. Although deep learning shows robust performance, the focus remains on textual data. Future research could enhance analysis by integrating user demographic data to uncover broader sentiment trends.

Another paper develops a system to distinguish between spam and non-spam (" ham") tweets and analyze their sentiments using various machine learning and deep learning models, including logistic regression, SVM, random forest, and advanced neural networks like LSTM and BiLSTM [21]. The approach effectively classifies tweets based on preprocessed features, demonstrating strong spam detection and sentiment assessment capabilities. However, the focus on Twitter data might limit the applicability of finding platforms with different interaction dynamics and content structures, suggesting the need for cross-platform validation. Ray et al. presented a hotel recommendation system using BERT, Random Forest, sentiment analysis, and aspect-based classification [22]. TripAdvisor.com reviews are categorized by hotel qualities like cleanliness and service using fuzzy logic and cosine similarity. It outperforms prior systems with an 84% Macro F1-score and 92.36% accuracy. One platform's evaluations may limit input data diversity, therefore future study should use other sources for a more robust and generalized method. Deep learning is used to evaluate 112,412 Yelp.com restaurant reviews from the first half of 2020 to discover customer priorities during the COVID-19 epidemic [23]. The research extracts service, food, environment, and experience aspects and their sentiment scores to determine what factors customers love most in eating encounters. It tests deep learning models like Bidirectional LSTM and Simple Embedding + Average Pooling for sentiment classification and review rating prediction. These deep learning models outperform typical machine learning methods in such tasks.

Deep learning methods SRN, LSTM, and CNN are used to research how word count and review readability affect sentiment analysis performance on a movie reviews dataset [24]. Shorter, more readable reviews do best in classification. Statistical study shows that review duration affects accuracy more than readability. These findings show text evaluators or website plugins could improve crowd-sourced review quality. By improving textual quality, it can greatly improve sentiment analysis accuracy, helping consumers and organizations make better decisions. This study proposes the SLCABG model, which uses a sentiment lexicon, CNN, and attention-based Bidirectional Gated Recurrent Unit [25]. Attention classifies review sentiment features, CNN and BiGRU extract and contextualizes them, and the sentiment lexicon improves them. Text sentiment analysis on 100,000 Dangdang.com book reviews improve with SLCABG.

Multilingual sentiment analysis faces challenges like grammatical variations, cultural expressions, and code-mixed text, which traditional NLP models struggle with. Rule-based and machine learning methods rely on lexicons or curated datasets, limiting cross-language adaptability. While deep learning models (LSTMs, CNNs) enhance performance, they require extensive multilingual annotations.

2.3. Overall Findings

While sentiment analysis has been widely studied using machine learning and deep learning, challenges persist in multilingual settings. Traditional models (Naïve Bayes, SVM, Decision Trees) excel in single-language analysis but struggle with linguistic variations, while deep learning methods (LSTMs, CNNs) improve contextual understanding yet lack cross-language generalization. Our approach integrates an advanced BERT-based framework with additional neural layers, fine-tuned for English and Bengali. Unlike prior studies, we enhance model transparency using LIME, ensuring interpretability. This work bridges existing gaps, presenting a more robust, explainable, and multilingual-compatible sentiment analysis model. The emergence of deep learning, particularly through the implementation of CNNs, LSTMs, and Transformers, has significantly improved the ability to analyze intricate language patterns and sentiments without extensive manual feature engineering. This evolution has shown superior performance in detailed sentiment analysis across various data sources, including social media and specialized consumer reviews. Despite this research, the review identifies a notable gap in cross-linguistic applications and multimodal sentiment analysis. This gap underscores a crucial area for future research, aiming to enhance the versatility and accuracy of sentiment analysis tools in our increasingly digital and multilingual environment. This research contributes to the field by addressing these gaps and proposing novel approaches to multilingual and multimodal sentiment analysis, thereby advancing the state-of-the-art in sentiment analysis research.

3. Methodology

The methodology section outlines the systematic approach to conducting sentiment analysis on a multilingual dataset. The process involves translating Bangla text to English, preprocessing the data, training machine learning models, evaluating their performance, enhancing interpretability through explainability analysis, and proposing a novel model architecture based on BERT.

3.1. Dataset

For this study, we are utilizing benchmark datasets. We selected Kaggle datasets for their relevance in prior research and suitability for benchmarking sentiment analysis models. The English dataset covers diverse e-commerce reviews, while the Bengali dataset captures sentiment expressions unique to non-Latin script users. Both provide a balanced mix of positive and negative reviews, ensuring robust model training. In order to train and evaluate our suggested approach, we utilized various languages, necessitating the collection of two distinct datasets. The English and Bengali datasets were obtained from Kaggle and have been utilized in many research studies [26, 27]. We were required to translate the Bengali data and thereafter combine the two datasets. Due to the significantly increased number of instances, we were compelled to randomize them. Subsequently, we assigned 8924 instances to the bad class and 8922 instances to the good class data. Among them, we allocated 70% for training, 15% for validation, and 15% for testing.

3.2. Data Translation Process

The initial step involved translating Bengali text data into English. This process has been executed within the Google Colab environment using the Googletrans library. This translation process is crucial for enabling multilingual sentiment analysis. Though sentiment analysis in non-Latin scripts like Bengali poses challenges such as morphological complexity, transliteration inconsistencies, and script ambiguities. To address these, we employed structured preprocessing: Bangla-to-English translation via Googletrans with quality checks, text normalization (stop-word removal, tokenization, stemming), and fine-tuning on both Bengali and Banglish datasets. Additionally, BERT-based embeddings were enhanced with Bengali corpora to improve semantic understanding, strengthening the model's multilingual sentiment analysis capabilities which we are describing through next few sub sections.

1. Setting Up Environment: The first step involved setting up the Google Colab environment. This included integrating Google Drive into Colab and importing essential libraries such as pandas, matplotlib, seaborn, and scikit-learn.
2. Loading and Preprocessing Dataset: Bangla text data stored in CSV files were loaded into dataframes. These dataframes were then merged to create a unified dataset. Null values in the "cleaned sentence" column were removed to ensure data integrity.
3. Translation Process: The Googletrans library facilitates batch translation of Bangla text segments into English. This step is crucial for enabling multilingual sentiment analysis.
4. Error Handling and Data Integrity: Robust error handling mechanisms is implemented to address translation errors. Despite occasional challenges, the integrity of the translated dataset was maintained.
5. Dataset Enrichment: Each iteration of the translation process enriched the dataset with English equivalents of the original Bengali text. This iterative approach resulted in a comprehensive dataset ready for analysis.

Though translating Bengali to English poses challenges like meaning loss, idiomatic expressions, and context variations, leading to potential sentiment misclassification. To mitigate this, we applied quality control through manual verification, cross-validation with native speakers, and context-aware translation tools. Despite these measures, some nuances may be lost, highlighting the need for native Bengali sentiment analysis in future research.

3.3. Data Preprocessing and Exploration

Following translation, the preprocessed dataset underwent exploratory analysis and preparation for sentiment analysis. To ensure high-quality input data for sentiment analysis, we applied multiple preprocessing techniques to clean and standardize the text. Specifically, contraction expansion was used to replace common contractions (e.g., 'don't' → 'do not', 'can't' → 'cannot'), reducing ambiguity and improving tokenization. Lemmatization was applied to convert words to their root forms, ensuring consistency in word representation (e.g., 'running' → 'run', 'better' → 'good'). Additionally, HTML tags, special characters, and redundant spaces were removed to enhance text clarity. These preprocessing steps collectively improved the model's ability to capture sentiment-relevant features while reducing noise in the dataset.

1. Mounting Google Drive and Importing Libraries: The Google Drive was integrated into the Colab environment, and necessary data preprocessing and exploration libraries were imported.
2. Loading and Concatenating Datasets: Individual datasets were loaded into dataframes. These dataframes were then concatenated into a single dataframe to ensure comprehensive coverage and analysis.

3. Handling Class Imbalance: Techniques were employed to address class imbalance within the dataset. This ensured a balanced representation for sentiment analysis.
4. Text Preprocessing: Various text preprocessing techniques such as HTML tag removal, contraction expansion, and lemmatization were applied. This helped to standardize and cleanse the textual data.
5. Exploratory Data Analysis (EDA): Visualization techniques, including count plots and statistical summaries, were utilized to explore the dataset's characteristics. EDA provided insights into data distribution, feature importance, and potential patterns.

3.4. BERT Model

The Bidirectional Encoder Representations from Transformers (BERT) model, created by Google, is a pre-trained language model based on transformers that has significantly transformed activities related to natural language processing. BERT undergoes unsupervised learning on a vast collection of textual material, enabling it to grasp the contextual connections among words and phrases.

The architecture of BERT is composed of several transformer layers, each containing self-attention mechanisms and feedback forward neural networks. The model acquires contextual embedding for every token in a phrase, providing comprehensive representations encompassing both syntactic and semantic information. The formula for self-attention is given by:

$$\text{Attention}(Q,\ K,\ V) = \text{softmax}\left(\frac{QK^T}{\sqrt{\text{d}_\text{k}}}\right)V \tag{1}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors.

3.5. Proposed Model Architecture

The proposed model architecture integrates the power of BERT embeddings with traditional neural network layers for sentiment analysis on the multilingual dataset. The architecture comprises the following components:

In summary, explainability analysis using LIME enables us to gain a deeper understanding of the model's inner workings, making it easier to interpret and trust its predictions.

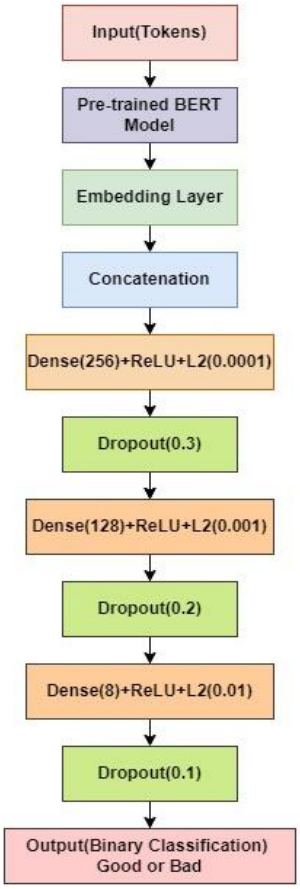

Fig. 1. Enhanced Proposed model

Table 1. Overview of Model Architecture and Explainability Analysis

BERT Embedding Layer	The input sequences are tokenized and passed through the pre-trained BERT model to obtain contextual embeddings for each token.
Dense Layers	Multiple dense layers with ReLU activation functions are stacked on top of the pooled embeddings to learn complex patterns and relationships in the data. The equation for a dense layer is given by: $Y = \mathrm{ReLU}(X \cdot W+b)$ (2) where X is the input tensor, W is the weight matrix, b is the bias vector, and Y is the output tensor.
Dropout Regularization	Dropout layers are inserted between dense layers to prevent overfitting and improve model generalization. The equation for dropout regularization is given by: $\mathrm{Dropout}(X) = X \odot M$ (3) where X is the input tensor and M is a binary mask matrix generated randomly with a specified dropout rate.
Output Layer	The final dense layer with a sigmoid activation function generates the output predictions, indicating the sentiment polarity of the input text. Explainability Analysis (XAI) An investigation of explainability is essential for improving the interpretability of machine learning models. This section outlines the systematic methodology employed to do explainability analysis using LIME (Local Interpretable Model-agnostic Explanations), which offers insights into the model's decision-making process at a local level. Installation and Setup: To initiate the explainability analysis process, the initial step entails installing the LIME library and configuring the essential dependencies. This guarantees that the necessary tools are accessible for generating explanations. Explanation Generation: After the setup is finished, LIME produces explanations for the model's predictions on example input sentences. LIME operates by perturbing the input data and analyzing the resulting model output alterations. This enables it to approximate the model's behavior in a localized manner. Visualization: The explanations produced by LIME are thereafter presented visually to offer clear and understandable insights into the prediction process of the model. Visualization methods, such as bar charts, heat maps, and saliency maps, can be employed to emphasize the most significant features or tokens that contribute to the decision made by the model.

4. Results

We conducted an experiment using a transferable knowledge-based model to evaluate the sentiment of e-commerce reviews in many languages. To further our understanding of the benefits of our methodology, we conducted a thorough investigation, comparing it directly with well-established methodologies. Our model's assessment entailed utilizing various measures, including accuracy, precision, recall, and AUC value. SVM is widely recognized for its exceptional performance in machine learning, but BERT has emerged as a prominent competitor in deep learning. We have enhanced the BERT model to target this problem's importance precisely. After conducting thorough testing and making necessary revisions, our proposed model has demonstrated outstanding performance compared to the current prevailing model. The model has achieved test accuracy, precision, recall, and f-1 score of 0.92 each. Previously, we analyzed various machine-learning models using various vectorizers. The TF-IDF vectorizer demonstrated exceptional performance when combined with the Support Vector Machine, outperforming other methods. The assessment of all the machine learning and deep learning models that were assessed is documented in Tables 2. The given example shows that the BERT model outperforms conventional machine learning techniques in terms of performance. The technique we endorse demonstrated exceptional performance compared to all benchmarks in this classification task, closely approximating the precision of the most accurate deep-learning model. A confusion matrix was generated to examine each model thoroughly. Figures 2 and 3 depict the complete set of confusion matrices.

After examining the confusion matrix, it is clear that some basic machine learning models, such as Logistic regression, Gradient boosting, Decision Tree, Random Forest, and Support Vector Machine, showed below-average performance in a multilingual setting. In contrast, the BERT model has demonstrated remarkable achievements in the domain of "Deep" learning. The suggested approach, employing Bert, improved the overall performance. The shade of blue along the diagonal in the confusion matrix diagram represents the instances when the model's forecast closely matched the actual value. The outcomes derived from our model are displayed in Fig.4, Fig.5, and Fig.6. The figures illustrate quantitative accuracy, loss, precision, recall, and AUC measurements.

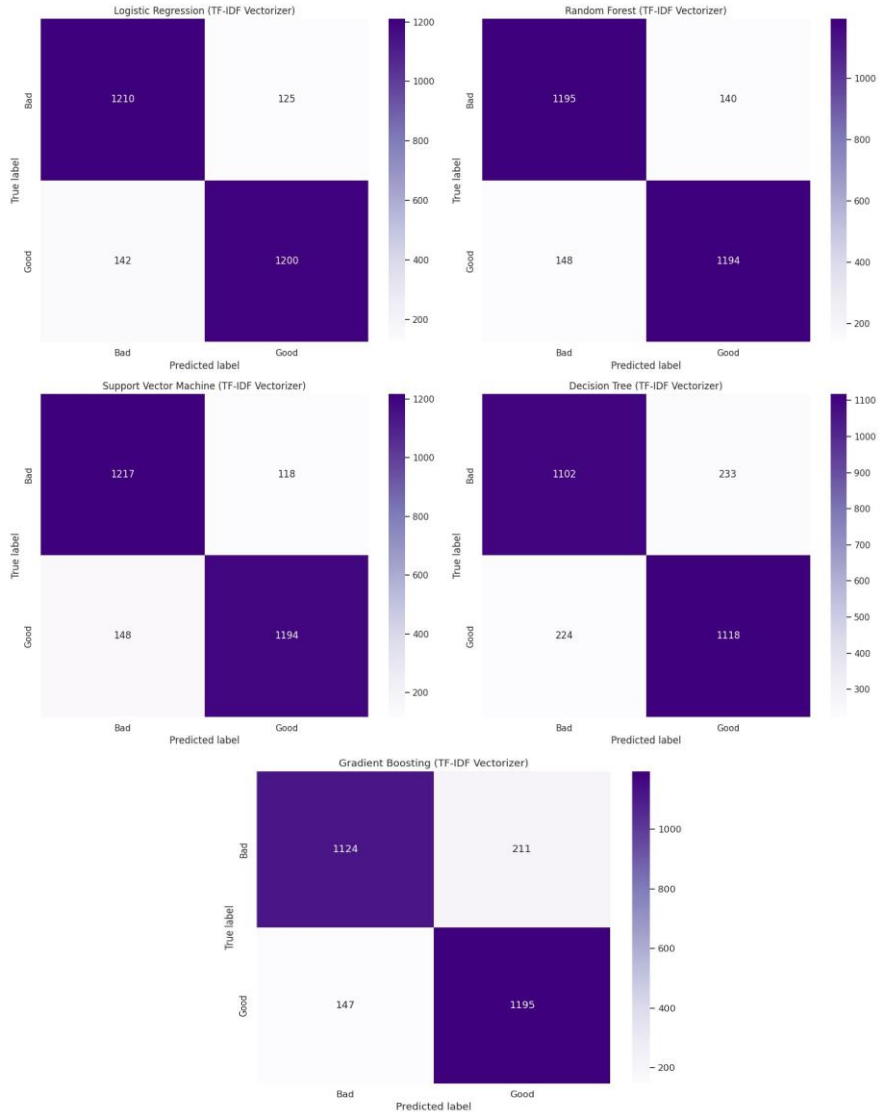

Fig. 2. This comprehensive visualization showcases the performance of various machine learning models through the use of confusion matrices.

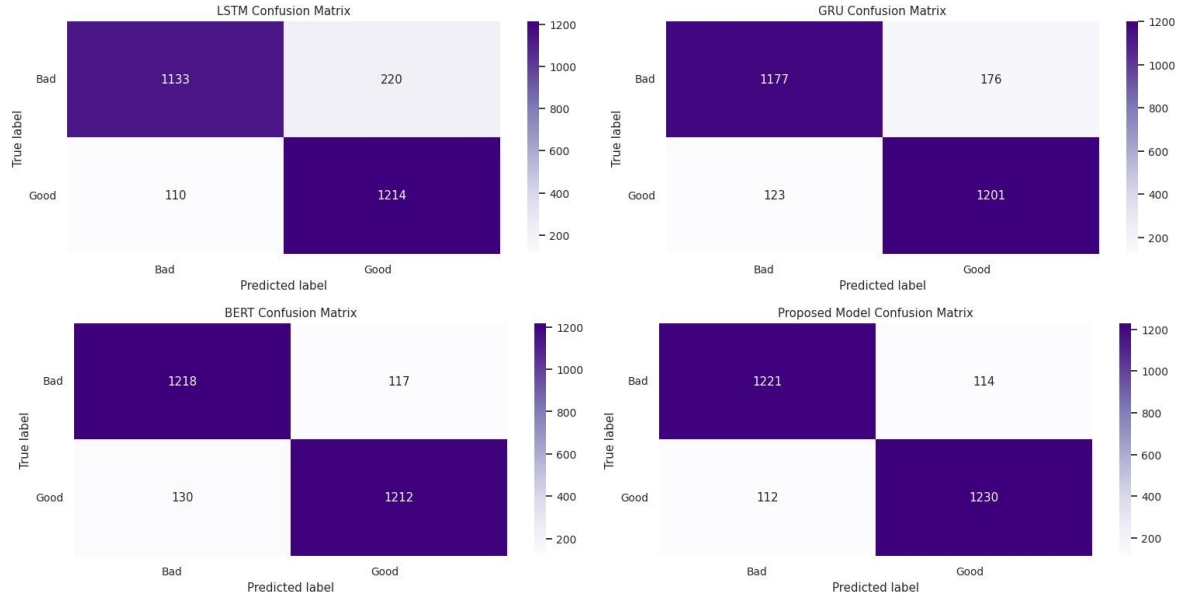

Fig. 3. This comprehensive visualization showcases the performance of various deep learning models using confusion matrices.

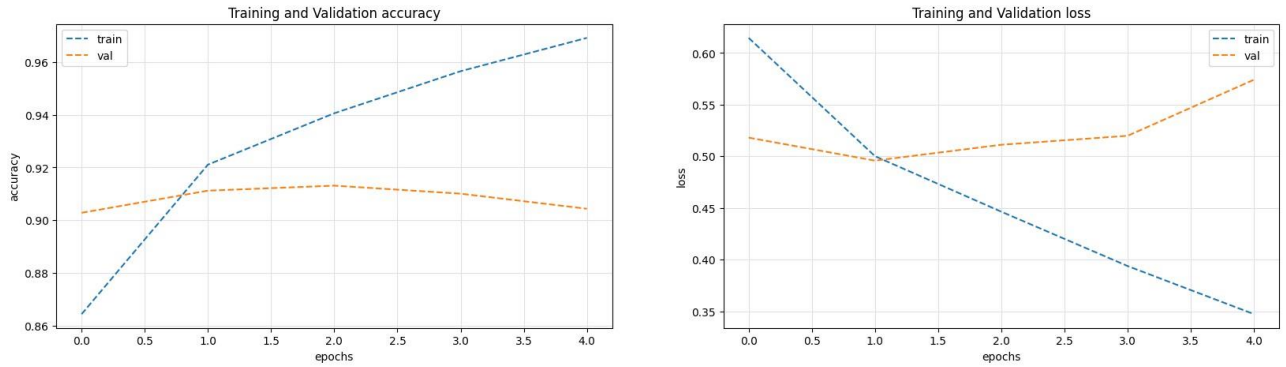

Fig. 4. On the far left side of the picture, we can see a graph that displays how accurate the training and validation were. On the far right, we can see a graph depicting the loss during the validation and training processes.

The efficacy of our model was exceptional when we utilized visualizations to represent the data through various graphical formats. The model experienced a slight problem of overfitting. The model demonstrated effective learning from the training data, as evidenced by the consistent decrease in the ratio between the accuracy of the training and validation curves. We refined our model to address overfitting concerns and achieve maximum accuracy. The loss graph illustrates the impressive efficacy of our approach, as the curves display a significant level of closeness to each other. The graph labelled as Fig.5 illustrates the ratio of training and testing for recall and precision. It visually represents these measures. In order to improve our model analysis, we have obtained the area under the curve (AUC) plot, as shown in Fig.6. The outcomes of our model are contrasted with those of other models in Table 2.

The AUC score of our model, as indicated by the area under the curve (AUC) graph, represents a level of performance that is very close to perfection, reaching a value of 1. A larger area under the curve (AUC) indicates enhanced consistency in identifying data across multiple genres. The algorithm we suggested demonstrated exceptional performance in the task of classifying e-commerce reviews.

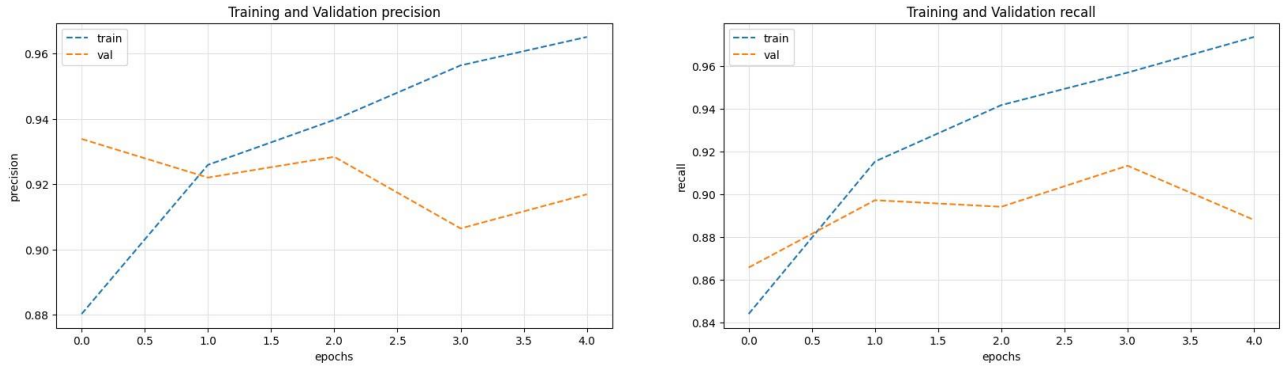

Fig. 5. The picture shows the recall graphs from the identical training and validation procedures in the right-hand corner, and the precision graphs in the left-hand corner.

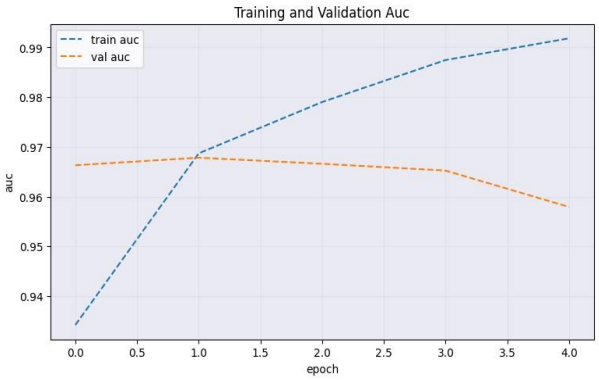

Fig. 6. Training and validation AUC values are displayed in this figure.

Table 2. Performance Metrics for Machine Learning Models

ML Method Name	Accuracy (%)	Loss	Precision	Recall	F1-score	ROC-AUC
Support Vector Machine (TF-IDF Vectorizer)	90.06	0.099	91.01	88.97	89.98	90.07
Logistic Regression (TF-IDF Vectorizer)	90.03	0.100	90.57	89.42	89.99	90.03
Random Forest (TF-IDF Vectorizer)	89.24	0.108	89.51	88.97	89.24	89.24
Gradient Boosting (TF-IDF Vectorizer)	86.63	0.134	84.99	89.05	86.97	86.62
Decision Tree (TF-IDF Vectorizer)	82.93	0.171	82.75	83.31	83.03	82.93
LSTM	88.00	0.320	85.00	92.00	88.00	88.00
GRU	89.00	0.300	87.00	91.00	89.00	89.00
BERT	91.00	0.240	91.00	90.00	91.00	97.00
Proposed Model	92.00	0.490	92.00	92.00	92.00	97.00

The subplot of Fig.8 displays the images and their corresponding descriptions. For each photograph, we displayed both the original image and LIME. In addition, we have included both the current label and the predicted label for the image. Our research findings indicate that the BERT-based model consistently and accurately classified each image into its respective category. The provided explanations enabled us to gain a deep understanding of the model's prediction and the specific image elements that significantly influenced it.

The superiority of our model above all previous models is clearly demonstrated by the performance assessment table. The model achieved an accuracy of 0.92, precision of 0.92, recall of 0.91, f-1 score of 0.92, and area under the curve of 0.97. Fig.7 presents a visual representation of each supported language's different degrees of precision. The model demonstrated a significantly higher level of performance in Bengali and English languages. An important hindrance is the limited availability of resources to prepare Bangla. Considering that this is the initial attempt to analyze sentiment in Banglish, the accuracy level is really high. While performance metrics such as accuracy, precision, recall, and F1-score indicate the effectiveness of our proposed model, their practical implications are significant for real-world e-commerce applications. A high accuracy of 92% ensures that sentiment predictions align closely with actual user opinions, making the model reliable for automated review classification. The precision and recall scores highlight the model's ability to correctly identify positive and negative sentiments while minimizing misclassifications. This level of performance is particularly beneficial for businesses that rely on sentiment analysis for customer feedback monitoring, product recommendations, and automated response systems. By accurately detecting sentiment patterns, companies can enhance their marketing strategies, improve customer engagement, and proactively address negative feedback, ultimately leading to better customer satisfaction and retention.

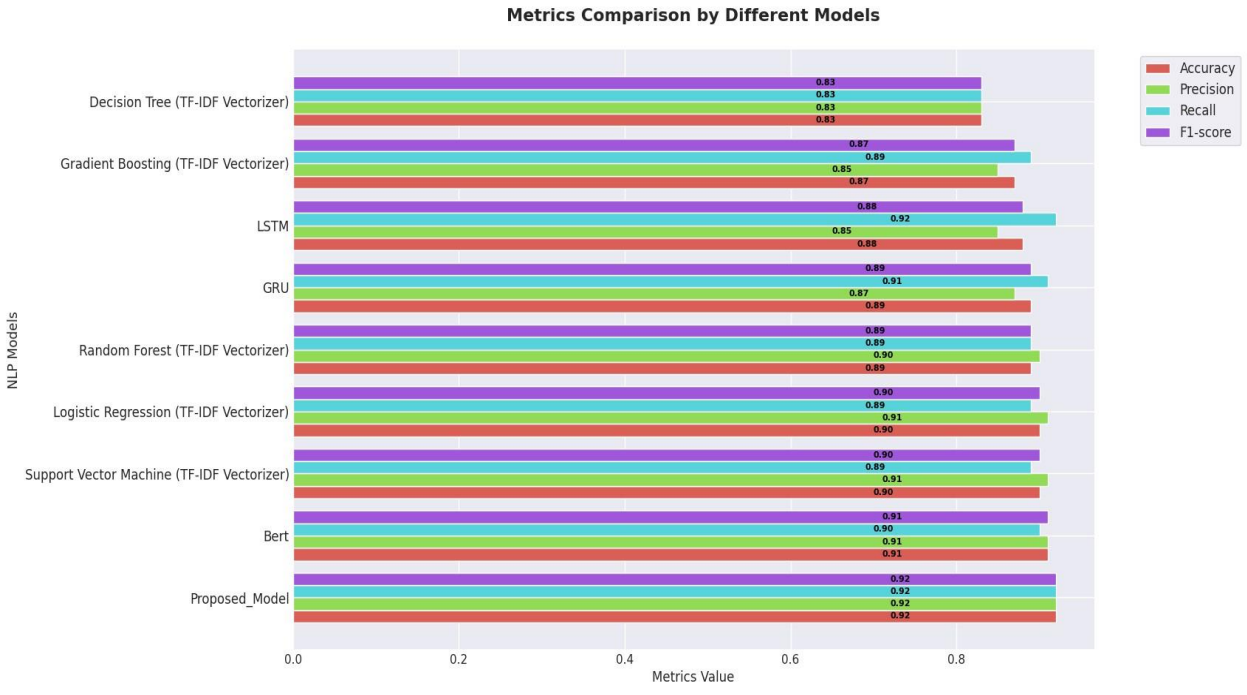

Fig. 7. Accuracy scores for both languages using different approaches

We employ the Local Interpretable Model-Agnostic Explanation (LIME) technique to demonstrate how our model discerns whether a phrase elicits a positive or negative response by analyzing particular phrases. LIME demonstrates the categorization of models according to a specific area. The model output displayed in Figure 8 offers valuable information regarding the transparency of the model.

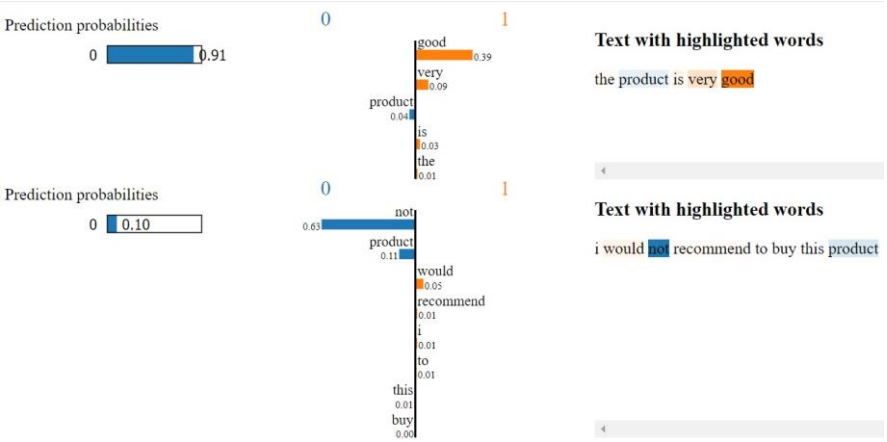

Fig. 8. Providing explainable AI solutions for all classes

5. Discussion

This study contributes to the field of sentiment analysis by showing the effectiveness of a new BERT-based model designed specifically for analyzing multilingual e-commerce evaluations. The model achieves a high level of consistency in accuracy, precision, recall, and F1-score, with a value of 0.92. The superiority of our model stems from its architectural improvements, such as the incorporation of extra neural network layers and the enhancement of its ability to recognize intricate semantic links in both English and Bengali languages. Despite its strengths, the model encounters limitations with colloquial language and the nuances lost in translation from Bengali to English, suggesting areas for future development. Our approach addresses critical gaps in handling linguistic diversity within sentiment analysis compared to existing literature. The practical implications of our findings are significant for global e-commerce markets, enabling businesses to harness sophisticated sentiment analysis for better customer engagement. Future research should explore direct sentiment analysis in Bengali to eliminate translation biases and investigate unsupervised learning techniques to capture informal language usage better. Theoretically, our study enriches the understanding of applying deep learning to NLP while also emphasizing the need for ethically leveraging AI to interpret diverse consumer feedback in an expanding digital economy. Though the confusion matrix provides valuable insights into the model's misclassifications, revealing key areas for further refinement. Most misclassifications occurred in sentimentally ambiguous reviews or those containing mixed sentiment (e.g., 'The product quality is great, but the delivery was extremely late'). This indicates that while our model effectively captures dominant sentiment trends, it sometimes struggles with contextual subtleties and sentiment shifts within a single review. Additionally, errors were more frequent in translated Bengali reviews, suggesting potential information loss or contextual mismatches during translation. To address these issues, future work could incorporate direct sentiment analysis on native Bengali text without translation, employ context-aware attention mechanisms, and explore multimodal approaches integrating textual and visual sentiment cues. By refining the handling of ambiguous and mixed-sentiment reviews, the model's overall accuracy and reliability can be further enhanced. While the proposed model demonstrates high accuracy in multilingual sentiment classification, certain limitations remain. One key challenge is the handling of informal language, including slang, abbreviations, and code-mixed text (e.g., Banglish). Since sentiment is often conveyed through informal expressions, the model may misinterpret reviews that deviate from standard grammar and syntax. Additionally, the translation process from Bengali to English introduces the risk of meaning distortion, particularly for idiomatic phrases and culturally specific expressions. Despite quality control measures, some nuances may be lost, affecting sentiment classification accuracy. To mitigate these issues, future iterations of the model could incorporate direct analysis of Bengali text using transformer-based architectures pre-trained on native Bengali datasets, eliminating the need for translation. Furthermore, integrating a context-aware sentiment classification mechanism could help refine the model's ability to interpret mixed and ambiguous sentiments more effectively.

6. Conclusion

The success of our improved BERT-based model marks a significant advancement in sentiment analysis for e-commerce reviews across various languages. An impressive score of 0.92 across key metrics like accuracy, precision, recall, and F1 score underscores the strength of our approach. The inclusion of extra neural network layers has enabled us to effectively address the challenges posed by linguistic diversity, highlighting the potential of advanced architectures

in managing such complexities. Despite the persistent obstacle of informal language, our study lays the groundwork for future exploration. Integrating unsupervised learning methods could enhance our model's ability to interpret informal expressions, while direct analysis without translation holds promise in preserving subtle nuances. These avenues for further research deepen our grasp of natural language processing and have far-reaching implications for e-commerce platforms worldwide. Our findings open doors to enhanced customer experiences by improving comprehension and engagement strategies. We advocate for ongoing advancements in deep learning techniques for sentiment analysis, aiming to enrich both academic research and practical applications across diverse linguistic landscapes. Building upon this work, future research could explore the integration of multimodal approaches, incorporating visual and auditory cues alongside textual data for sentiment analysis. These work adaptation techniques could also refine the model's performance in specific e-commerce contexts, catering to different product categories and customer demographics. Furthermore, analyzing sentiment evolution over time could offer valuable insights into changing consumer perceptions, informing proactive decision-making for businesses. Collaborative efforts with industry partners could drive the development of tailored solutions to address real-world challenges encountered by e-commerce platforms. Ultimately, we aim to instigate a shift in sentiment analysis methodologies, fostering deeper insights and more effective strategies to enhance global customer engagement and satisfaction. Future research should refine the model by integrating domain-specific sentiment lexicons and fine-tuning on larger datasets. Using models like BanglaBERT would improve sentiment classification for non-Latin scripts. Multimodal sentiment analysis, combining text, visual, and audio cues, could enhance accuracy. Testing in live e-commerce settings and deploying as an API would enable continuous refinement based on real-world feedback.

References

[1] Aytug Onan. Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks.˘ *Concurrency and computation: Practice and experience*, 33(23):e5909, 2021.

[2] Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226:107134, 2021.

[3] El Mahdi Mercha and Houda Benbrahim. Machine learning and deep learning for sentiment analysis across languages: A survey. *Neurocomputing*, 531:195–216, 2023.

[4] Dina Oralbekova, Orken Mamyrbayev, Mohamed Othman, Dinara Kassymova, and Kuralai Mukhsina. Contemporary approaches in evolving language models. *Applied Sciences*, 13(23):12901, 2023.

[5] Xianchang Luo, Yinxing Xue, Zhenchang Xing, and Jiamou Sun. Prcbert: Prompt learning for requirement classification using bert-based pretrained language models. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, pages 1–13, 2022.

[6] Shinji Kawakura, Masayuki Hirafuji, Seishi Ninomiya, and Ryosuke Shibasaki. Analyses of diverse agricultural worker data with explainable artificial intelligence: Xai based on shap, lime, and lightgbm. *European Journal of Agriculture and Food Sciences*, 4(6):11–19, 2022.

[7] Tofayet Sultan, Nusrat Jahan, Ritu Basak, Mohammed Shaheen Alam Jony, and Rashidul Hasan Nabil. Machine learning in cyberbullying detection from social-media image or screenshot with optical character recognition. *Int. J. Intell. Syst. Appl*, 15:1–13, 2023.

[8] Sanjay Dey, Sarhan Wasif, Dhiman Sikder Tonmoy, Subrina Sultana, Jayjeet Sarkar, and Monisha Dey. A comparative study of support vector machine and naive bayes classifier for sentiment analysis on amazon product reviews. In *2020 International Conference on Contemporary Computing and Applications (IC3A)*, pages 217–220. IEEE, 2020.

[9] Minhajul Abedin Shafin, Md Mehedi Hasan, Md Rejaul Alam, Mosaddek Ali Mithu, Arafat Ulllah Nur, and Md Omar Faruk. Product review sentiment analysis by using nlp and machine learning in bangla language. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5. IEEE, 2020.

[10] S Soumya and KV Pramod. Sentiment analysis of malayalam tweets using machine learning techniques. *ICT Express*, 6(4):300–305, 2020.

[11] Kanwal Zahoor, Narmeen Zakaria Bawany, and Soomaiya Hamid. Sentiment analysis and classification of restaurant reviews using machine learning. In *2020 21st International Arab Conference on Information Technology (ACIT)*, pages 1–6. IEEE, 2020.

[12] N Pavitha, Vithika Pungliya, Ankur Raut, Roshita Bhonsle, Atharva Purohit, Aayushi Patel, and R Shashidhar. Movie recommendation and sentiment analysis using machine learning. *Global Transitions Proceedings*, 3(1):279– 284, 2022.

[13] Sachin Chirgaiya, Deepak Sukheja, Niranjan Shrivastava, and Romil Rawat. Analysis of sentiment based movie reviews using machine learning techniques. *Journal of Intelligent & Fuzzy Systems*, 41(5):5449–5456, 2021.

[14] Amgad Muneer and Suliman Mohamed Fati. A comparative analysis of machine learning techniques for cyberbullying detection on twitter. *Future Internet*, 12(11):187, 2020.

[15] Kamaran H Manguri, Rebaz N Ramadhan, and Pshko R Mohammed Amin. Twitter sentiment analysis on worldwide covid-19 outbreaks. *Kurdistan Journal of Applied Research*, pages 54–65, 2020.

[16] Furqan Rustam, Madiha Khalid, Waqar Aslam, Vaibhav Rupapara, Arif Mehmood, and Gyu Sang Choi. A performance comparison of supervised machine learning models for covid-19 tweets sentiment analysis. *Plos one*, 16(2):e0245909, 2021.

[17] Petr Hajek, Aliaksandr Barushka, and Michal Munk. Fake consumer review detection using deep neural networks integrating word embeddings and emotion mining. *Neural Computing and Applications*, 32(23):17259–17274, 2020.

[18] S Nagi Alsubari, Sachin N Deshmukh, A Abdullah Alqarni, Nizar Alsharif, TH Aldhyani, Fawaz Waselallah Alsaade, and Osamah I Khalaf. Data analytics for the identification of fake reviews using supervised learning. *Computers, Materials & Continua*, 70(2):3189–3204, 2022.

[19] Srikanth Tammina and Sunil Annareddy. Sentiment analysis on customer reviews using convolutional neural network. In *2020 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–6. IEEE, 2020.
[20] Saeed Mian Qaisar. Sentiment analysis of imdb movie reviews using long short-term memory. In *2020 2nd International Conference on Computer and Information Sciences (ICCIS)*, pages 1–4. IEEE, 2020.
[21] Anisha P Rodrigues, Roshan Fernandes, Adarsh Shetty, Kuruva Lakshmanna, R Mahammad Shafi, et al. Real-time twitter spam detection and sentiment analysis using machine learning and deep learning techniques. *Computational Intelligence and Neuroscience*, 2022, 2022.
[22] Biswarup Ray, Avishek Garain, and Ram Sarkar. An ensemble-based hotel recommender system using sentiment analysis and aspect categorization of hotel reviews. *Applied Soft Computing*, 98:106935, 2021.
[23] Yi Luo and Xiaowei Xu. Comparative study of deep learning models for analyzing online restaurant reviews in the era of the covid-19 pandemic. *International Journal of Hospitality Management*, 94:102849, 2021.
[24] Lin Li, Tiong-Thye Goh, and Dawei Jin. How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis. *Neural Computing and Applications*, 32:4387–4415, 2020.
[25] Li Yang, Ying Li, Jin Wang, and R Simon Sherratt. Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning. *IEEE access*, 8:23522–23530, 2020.
[26] The Devastator. Amazon customer reviews with 2013-2019 sentiment. Kaggle Dataset, 2021.
[27] Kowshir Bitto. Bangla e-commerce sentiment. Kaggle Dataset, 2022.

Authors' Profiles

Nusrat Jahan completed her Master of Science in Computer Science with a specialization in Intelligent Systems in 2024 and received a Bachelor of Science in Computer Science and Engineering (CSE) with a major in Software Engineering from the American International University-Bangladesh (AIUB) in 2022. Her research interests are in Machine Learning, Deep Learning, and Natural Language Processing.

Jubayer Ahamed completed his Bachelor of Science in Computer Science from American International University-Bangladesh (AIUB), Dhaka, Bangladesh. He obtained his MSc from American International University-Bangladesh (AIUB), Dhaka, Bangladesh. Currently, Jubayer Ahamed is working as a Lecturer in the Department of Computer Science (CS) at American International University- Bangladesh (AIUB). His research interest includes Software Engineering and Machine Learning.

Dip Nandi is a Professor and the Associate Dean of the Faculty of Science and Technology at American International University-Bangladesh (AIUB). Previously, he served as the Director of the Computer Science Department at AIUB. He completed his PhD at RMIT University in Australia and earned an MSc from The University of Melbourne, Australia. His research interests span Software Engineering, E-Learning Technologies, Data Mining, and Information Systems. He is involved with several professional organizations, including IEEE and ACM, and has contributed to numerous peer-reviewed journals. Additionally, he is a member of the Academic Council at AIUB.

How to cite this paper: Nusrat Jahan, Jubayer Ahamed, Dip Nandi, "Enhancing E-commerce Sentiment Analysis with Advanced BERT Techniques", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.17, No.3, pp. 49-61, 2025. DOI:10.5815/ijieeb.2025.03.04