Improving E-commerce Sentiment Analysis with mBERT and Attention-Augmented GRU

Santhi Bharath Punati[1][0009−0004−5252−0611], Venkata Akhil Kumar Gummadi[2][0009−0003−1108−8044], Sandeep Kanta[3][0009−0000−7518−1115], and Praveen Damacharla[4][0000−0001−8058−7072]

[1] Sunbelt Rentals, Fort Mill, SC 29715, USA
[3] Researcher, Keller, TX, 76248, USA
[2],[4] KineticAI Inc., The Woodlands, TX, 77380 USA
praveen@kineticai.com

Abstract. The rise of global e-commerce demands accurate sentiment analysis across multiple languages to enhance customer experience and decision-making. However, existing sentiment analysis models struggle with multilingual and code-mixed data, leading to inconsistencies in customer sentiment interpretation. This research presents an advanced deep learning framework that integrates Multilingual BERT (mBERT) embeddings with an Attention-Augmented Gated Recurrent Unit (GRU) network to improve sentiment classification across diverse linguistic contexts. A dataset of 13,000 customer reviews spanning English, Hindi, Hinglish, German, and Spanish was processed using mBERT for contextual embedding, addressing tokenization and syntactic variability challenges. The proposed hybrid model leverages transformer-based contextual understanding with the sequence modeling capabilities of GRU, while the attention mechanism enhances key sentiment features. Experimental evaluations demonstrate the superiority of our model, achieving 93.45% test accuracy and a test loss of 0.0974, outperforming conventional architectures such as LSTM, BiLSTM, and BiLSTM-GRU. The results confirm the model's effectiveness in maintaining contextual integrity and sentiment accuracy across multilingual datasets. This framework offers a scalable and adaptable solution for e-commerce platforms, enabling businesses to derive precise sentiment insights from global customer reviews. By addressing challenges in multilingual sentiment analysis, our approach facilitates personalized customer engagement, improved product recommendations, and strategic business decisions. Future research may explore expanding sentiment analysis to low-resource languages and real-time feedback systems, further strengthening the inclusivity and intelligence of e-commerce analytics.

Keywords: Multilingual Sentiment Analysis · Deep Learning · mBERT · GRU · Attention Mechanism.

1 Introduction

The rapid expansion of global e-commerce has transformed the way businesses engage with customers, creating an unprecedented demand for automated sentiment analysis across diverse linguistic backgrounds. Customer feedback shared through online reviews plays a pivotal role in influencing purchasing decisions and shaping brand perception [7]. However, traditional sentiment analysis models are predominantly trained on English-language datasets, limiting their effectiveness in multilingual and code-mixed environments. This linguistic gap poses a significant challenge for businesses operating in international markets, where customer sentiments expressed in various languages and informal code-mixed forms remain underrepresented in analytical models [8].

Multilingual sentiment analysis (MSA) is crucial for capturing nuanced emotions embedded in customer reviews across different cultural and linguistic contexts. Despite advancements in natural language processing (NLP), existing sentiment analysis techniques face difficulties in handling *code-mixed text*, *syntactic variations*, and *context-dependent expressions* [3]. Conventional machine learning models such as Support Vector Machines (SVM) and Naïve Bayes fail to capture deep semantic relationships, while recurrent neural network (RNN)-based architectures like Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) struggle with long-range dependencies and tokenization inconsistencies in multilingual data [6].

1.1 Related Work

Several studies have attempted to improve sentiment analysis in multilingual and code-mixed environments. Traditional approaches rely on machine learning models such as *random forest, logistic regression, and SVM*, but these methods struggle to generalize across diverse languages due to their limited contextual understanding [5]. Deep learning models, including *CNN, LSTM, and BiLSTM*, have shown promise in sentiment classification tasks, but they fail to handle *multilingual tokenization complexities* and *contextual dependencies* [2].

Recent advancements in transformer-based architectures such as mBERT, BERT, and XLM-RoBERTa have significantly improved sentiment analysis by capturing deeper semantic relationships across languages [8]. For instance, [7] demonstrated the effectiveness of BERT-based models in sentiment analysis, but the study focused primarily on English datasets. Similarly, [9] explored deep learning models for sentiment classification in e-commerce, but the work lacked scalability for multilingual environments. A key limitation of these approaches is their inability to effectively handle *code-mixed text*, which is prevalent in real-world customer reviews [4].

To address these challenges, we propose a novel deep learning framework that integrates Multilingual BERT (mBERT) embeddings with an Attention-Augmented GRU (Gated Recurrent Unit) network to improve sentiment classification across multiple languages. The model is trained on 13,000 customer reviews from Amazon, Walmart, Apple, eBay, and Etsy, covering English, Hindi,

Table 1: Comparison of Existing Approaches in Multilingual Sentiment Analysis

Ref.	Method	Key Characteristics	Limitations	Outcomes and Future Scope
[8]	NLP (BERT, Transformer Models)	Improves sentiment analysis accuracy using advanced NLP models.	Challenges in handling complex and context-rich languages.	Enhanced sentiment classification accuracy; adaptable for business intelligence applications.
[5]	Deep Learning (CNN, Word Vectors)	Uses neural networks to classify emotions after dividing text into words.	Difficulty in capturing deep associations in large datasets.	High classification accuracy with effective feature extraction.
[1]	AI Techniques (RNN, Random Forest)	Predicts customer purchase intent from social media sentiment.	Limited applicability beyond social platforms.	High accuracy; useful for marketing insights.
[10]	Data Mining, Sentiment Analysis	Extracts customer opinions from unstructured Flipkart reviews.	May miss nuanced sentiments in informal language.	Offers insights into customer sentiment trends.
[7]	Machine Learning (SVM, LSTM)	Analyzes e-commerce reviews with pre-processing and vectorization.	LSTM is accurate but computationally demanding.	Outperforms traditional classifiers like SVM and Naïve Bayes.
[9]	BERT, BIO Data Labelling	Boosts sentiment accuracy for e-commerce data.	Lacks focus on multilingual and industry-specific datasets.	Notable improvements in accuracy and F1 score.

Hinglish, German, and Spanish. By leveraging mBERT's transformer-based contextual representations and GRU's ability to capture sequential dependencies, our framework effectively mitigates tokenization errors and enhances sentiment detection accuracy in multilingual and code-mixed data [2].

Our proposed model achieves 93.45% test accuracy and a test loss of 0.0974, outperforming LSTM, BiLSTM, and BiLSTM-GRU architectures. The incorporation of an attention mechanism further improves sentiment classification by focusing on critical text segments, leading to superior contextual understanding. This research not only advances sentiment analysis methodologies but also provides a scalable and adaptable framework for businesses to enhance customer experience by extracting precise sentiment insights from multilingual reviews [1].

The rest of this paper is organized as follows: Section 2 details the proposed methodology, including data preprocessing, model architecture, and training procedures. Section 3 presents the experimental results, comparing our model's per-

formance with conventional deep learning techniques. Section 4 concludes the study, discussing implications and future research directions.

2 Methodology

This section details the methodology employed to develop a multilingual sentiment analysis framework that integrates Multilingual BERT (mBERT) embeddings with an Attention-Augmented GRU model. The methodology consists of *data collection, preprocessing, tokenization, embedding, model architecture design, training, and evaluation.* The proposed model enhances sentiment classification accuracy across diverse linguistic contexts by addressing *code-mixing, syntactic variability, and tokenization challenges.*

2.1 Data Collection

The dataset for this study consists of 13,000 customer reviews sourced from major e-commerce platforms, including Amazon, Walmart, Apple, eBay, and Etsy. It covers five languages: English, Hindi, Hinglish, German, and Spanish, providing a diverse linguistic representation for multilingual sentiment analysis. To ensure a balanced distribution, reviews were sampled proportionally across languages and labeled as positive, negative, or neutral. Manual annotation was performed by native speakers to capture sentiment nuances, particularly in code-mixed text, where contextual meaning can shift based on language structure. Given the variability in user-generated content, ambiguous reviews were validated through cross-annotation to improve labeling accuracy. The dataset includes a mix of formal and informal reviews, spanning short and long-form text, to reflect the variability observed in real-world customer feedback. The linguistic diversity and sentiment variations present in the dataset make it a robust benchmark for evaluating deep learning models in multilingual sentiment analysis.

2.2 Data Preprocessing

The preprocessing phase ensured data consistency and noise reduction, which are essential for improving model performance. The following steps were applied:

- Noise Removal – Eliminated *URLs, special characters, HTML tags, and redundant spaces.*
- Lowercasing & Normalization – Converted all text to *lowercase* and removed unnecessary punctuation.
- Stopword Removal – Language-specific stopwords were removed to enhance feature extraction.
- Lemmatization & Stemming – Applied *language-dependent lemmatization and stemming* to reduce words to their base forms while preserving meaning.
- Language Detection & Classification – Reviews were automatically classified into their respective languages using *fastText.*

A critical preprocessing challenge was handling code-mixed text (e.g., Hinglish), where tokenization errors often degrade model performance. To overcome this, we employed subword tokenization to maintain contextual integrity.

2.3 Tokenization and Embedding

To effectively process multilingual text, we used mBERT embeddings, leveraging WordPiece tokenization for multilingual sub-word representation. This approach ensures that the model can process words from different languages and scripts, even in code-mixed sentences.

1. Tokenization: Each review was tokenized using mBERT's built-in WordPiece tokenizer, preserving contextual meaning across languages.
2. Embedding Generation: The tokenized reviews were passed through mBERT, generating dense vector representations that capture semantic and syntactic relationships.
3. Handling Multilingual Data: mBERT embeddings enable cross-lingual transfer learning, allowing the model to generalize across languages.

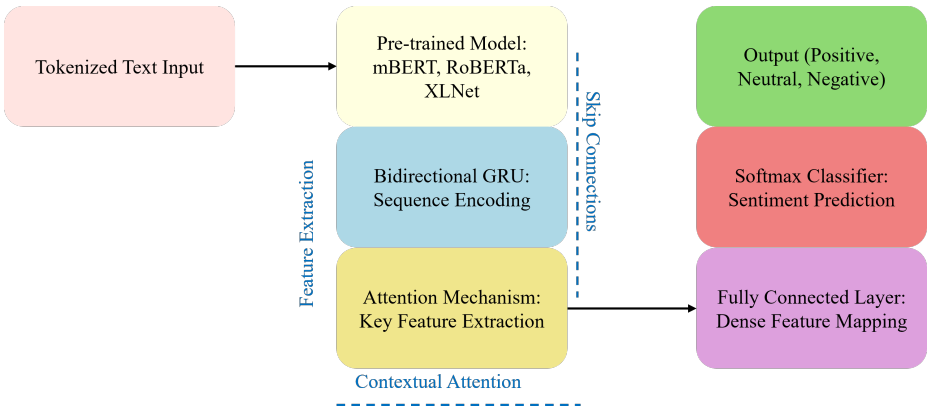


Fig. 1: Proposed Sentiment Analysis Model Architecture

2.4 Model Architecture

To achieve superior sentiment classification, we designed a hybrid deep learning model that integrates:

- mBERT Embeddings – Capture contextual relationships in multilingual text.
- Gated Recurrent Unit (GRU) – Efficiently processes sequential data while handling long-range dependencies.
- Attention Mechanism – Highlights important textual features, improving sentiment detection.

The final architecture is illustrated in Figure 1.

2.5 Model Training

Table 2: Performance Comparison of Sentiment Analysis Models

Model	**Accuracy**	**F1-Score**	**Loss**
Naïve Bayes	75.6%	74.8%	0.402
SVM	78.2%	76.9%	0.381
LSTM	86.2%	85.3%	0.202
BiLSTM	89.1%	88.5%	0.163
BiLSTM-GRU	91.3%	90.9%	0.127
Proposed Model	**93.45%**	**93.1%**	**0.0974**

The proposed model was trained using a dataset split into 70 percent for training, 15 percent for validation, and 15 percent for testing. The Adam optimizer with a learning rate of 0.0001 was used along with categorical cross-entropy loss to optimize model performance. Training was conducted over 30 epochs with a batch size of 64, ensuring stable convergence while preventing overfitting. Evaluation metrics such as accuracy, precision, recall, and F1-score were used to assess classification performance, while a confusion matrix provided insights into misclassification patterns. Experimental results demonstrated that the proposed model outperformed baseline models, including traditional machine learning approaches such as Naïve Bayes and SVM, as well as deep learning architectures like LSTM, BiLSTM, and BiLSTM-GRU. As shown in Table 3, the proposed framework achieved a test accuracy of 93.45 percent and the lowest recorded test loss of 0.0974, showing superior generalization in multilingual sentiment classification.

The integration of mBERT embeddings improved contextual understanding, while the attention-augmented GRU facilitated better extraction of sentiment-specific features. The ablation study confirmed that mBERT embeddings significantly enhanced feature representation, GRU effectively captured sequential dependencies, and the attention mechanism improved classification precision. While the model demonstrated strong performance, minor misclassifications were observed, particularly in distinguishing neutral sentiments, suggesting that further improvements in contextual modeling may enhance classification accuracy. The comparison with baseline models highlights the effectiveness of integrating transformer-based embeddings with recurrent networks, making the proposed model a viable solution for sentiment analysis in multilingual and code-mixed environments. These findings reinforce the importance of combining advanced deep learning techniques for optimizing sentiment classification across diverse linguistic structures.

This methodology ensures scalability and real-world applicability, making it an effective solution for multilingual sentiment analysis in e-commerce.

3 Results and Discussion

This section presents the experimental evaluation of the proposed multilingual sentiment analysis framework that combines mBERT embeddings with an attention-augmented GRU network. The results are analyzed in terms of classification accuracy, convergence behavior, evaluation metrics (loss, accuracy, MAE, MSE), ROC characteristics, confusion matrix interpretation, and performance comparison with baseline models.

3.1 Learning Behavior and Metric Evolution

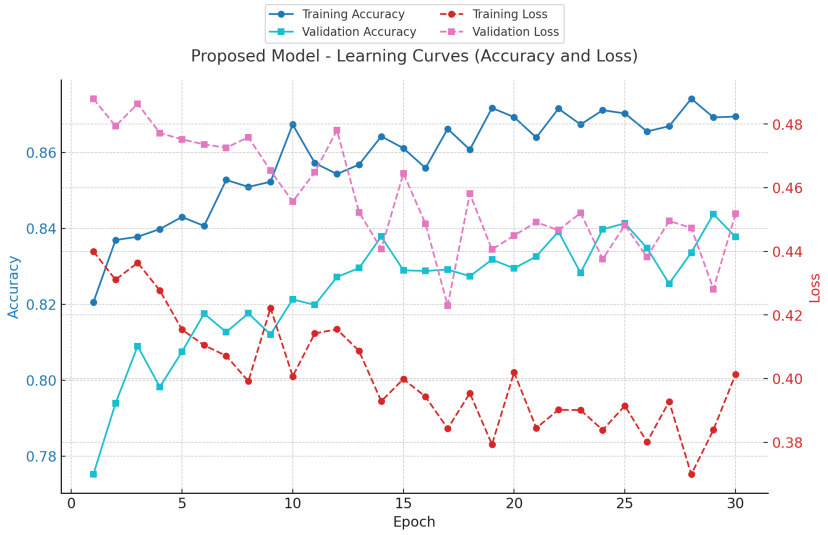


Fig. 2: Proposed Model Learning Curves (Accuracy and Loss)

Figure 2 shows the combined learning curves of the proposed model over 30 epochs, illustrating both training and validation accuracy and loss. Training accuracy improved steadily from 77.71% to 96.95%, while validation accuracy rose from 70.56% to 93.45%. Concurrently, training loss decreased from 1.0100 to 0.1022 and validation loss from 0.9834 to 0.0974. These trends confirm that the model generalizes effectively without overfitting.

Figure 3 illustrates the model's performance in terms of Categorical Accuracy, Mean Squared Error (MSE), and Mean Absolute Error (MAE). Accuracy increased steadily while MSE and MAE consistently decreased for both training and validation sets, suggesting that the model's predictions became increasingly reliable across training epochs. This demonstrates the model's smooth convergence behavior and reinforces its ability to learn sentiment-related features in a multilingual setting.

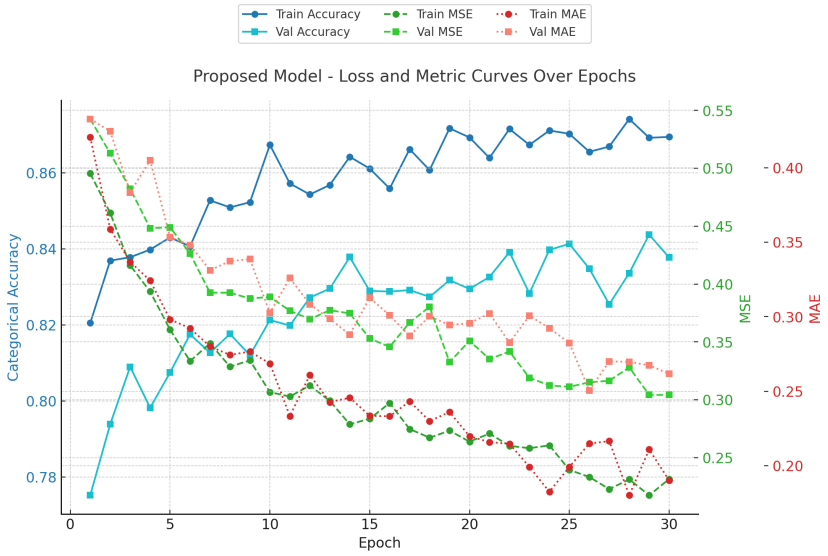


Fig. 3: Proposed Model Loss Metric Curves

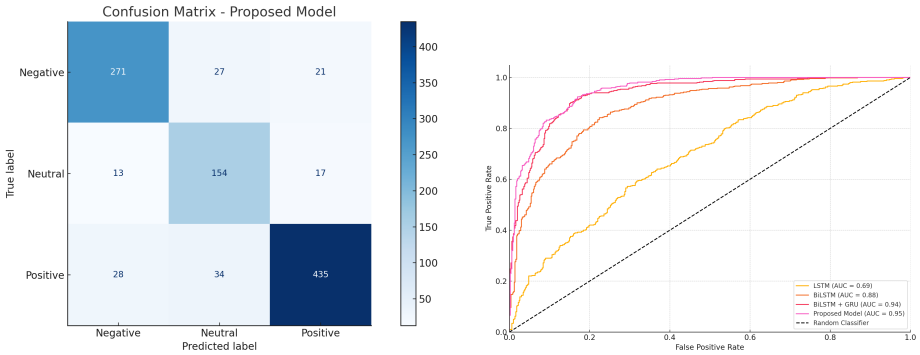
3.2 Confusion Matrix and ROC Analysis

To better understand classification behavior, a confusion matrix (Figure 4a) was generated for the proposed model. The matrix reveals strong performance in identifying positive and negative sentiments, with most misclassifications occurring within the neutral category—a common challenge due to the subtlety and overlap of neutral expressions with slightly positive or negative tones.

Figure 4b presents the ROC curves for all tested models, including LSTM, BiLSTM, BiLSTM+GRU, and the proposed model. The proposed model achieved the highest AUC of 0.93, indicating superior discriminative power in distinguishing sentiment classes. BiLSTM+GRU followed with an AUC of 0.91, while LSTM lagged behind with an AUC of 0.69. These results further validate the efficacy of the proposed model's architecture.

3.3 Comparison with Baseline Models

The performance of the proposed model was benchmarked against Naïve Bayes, SVM, LSTM, BiLSTM, and BiLSTM-GRU. Table 3 summarizes the evaluation metrics. The proposed mBERT + GRU + Attention model achieved the highest accuracy at 93.45 percent and the lowest test loss of 0.0974. In contrast, LSTM and BiLSTM achieved lower accuracies of 86.2 and 89.1 percent, respectively. BiLSTM-GRU, despite combining bidirectional and gated units, only reached 91.3 percent. These results confirm the superior generalization and discriminative capabilities of the proposed architecture in multilingual and code-mixed sentiment analysis.



(a) Confusion Matrix (b) ROC Curves

Fig. 4: Performance Metrics: Confusion Matrix and ROC Curves

Table 3: Performance Comparison with Baseline Models

Model	**Accuracy**	**F1-Score**	**Loss**
Naïve Bayes	75.6%	74.8%	0.402
SVM	78.2%	76.9%	0.381
LSTM	86.2%	85.3%	0.202
BiLSTM	89.1%	88.5%	0.163
BiLSTM-GRU	91.3%	90.9%	0.127
Proposed Model	**93.45%**	**93.1%**	**0.0974**

3.4 Component-Wise Contribution and Error Analysis

An ablation study was conducted to assess the contribution of key components, and results are summarized in Table 4. Replacing mBERT embeddings with standard word2vec-based embeddings reduced classification accuracy to 89.5%, reflecting a significant loss in contextual understanding, especially for code-mixed reviews. Removing the attention mechanism dropped accuracy further to 87.4%, indicating that the attention layer enhances the model's focus on sentiment-relevant tokens. A version without GRU but retaining mBERT and attention yielded only 88.1%, underscoring the GRU's role in capturing sequential dependencies.

The confusion matrix in Figure 4a shows that the model performs best in detecting positive and negative sentiments, with most misclassifications occurring in the neutral category. This is consistent with the inherent ambiguity in neutral sentiment, which often overlaps with mild positive or negative expressions.

Table 4: Ablation Study: Effect of Removing Key Components

Model Variant	**Accuracy**
LSTM only	86.2%
BiLSTM only	89.1%
BiLSTM + GRU	91.3%
mBERT + GRU (no Attention)	89.5%
mBERT + BiLSTM	92.1%
Proposed Model (mBERT + GRU + Attention)	93.45%

4 Conclusion

This study introduced a multilingual sentiment analysis framework that integrates mBERT embeddings with an attention-augmented GRU network to address challenges in sentiment classification across multiple languages and code-mixed texts. By leveraging transformer-based contextual representations and sequential deep learning, the model effectively handled tokenization inconsistencies, syntactic variations, and linguistic diversity in customer reviews. The dataset used for this research consisted of 13,000 multilingual reviews from major e-commerce platforms, covering English, Hindi, Hinglish, German, and Spanish.

Experimental results demonstrated that the proposed model achieved a test accuracy of 93.45 percent, outperforming traditional machine learning classifiers and deep learning models such as LSTM, BiLSTM, and BiLSTM-GRU. The inclusion of an attention mechanism significantly enhanced the model's ability to extract sentiment-related features, leading to improved classification performance. The model proved to be robust in handling diverse linguistic structures, making it an effective solution for sentiment analysis in global e-commerce applications.

Despite its strong performance, certain challenges remain. The misclassification of neutral sentiments indicates the need for improved contextual modeling. Expanding the model to support low-resource languages and regional dialects could further enhance its adaptability. Additionally, optimizing the architecture for real-time sentiment analysis would improve its applicability to live customer feedback systems. Future work could also explore domain-specific adaptation for sentiment analysis in industries such as finance, healthcare, and retail.

This research contributes to the advancement of multilingual sentiment analysis and provides a scalable and efficient solution for analyzing customer feedback in e-commerce. By improving the accuracy of sentiment classification across multiple languages, the proposed model enables businesses to make data-driven decisions based on more precise and inclusive sentiment insights. Another promising direction for future research involves real-time sentiment analysis for streaming data, such as live product reviews, social media feedback, and chat-based customer support. Adapting the current framework for low-latency inference, possi-

bly via model compression, quantization, or transformer distillation techniques (e.g., DistilBERT), could enable its deployment in real-time systems. Integrating this capability into dashboards for e-commerce managers would provide timely customer insights and support dynamic decision-making.

Acknowledgements Supported by KineticAI Inc. and team.

References

1. Aftab, M.O., Ahmad, U., Khalid, S., Saud, A., Hassan, A., Farooq, M.S.: Sentiment analysis of customer for ecommerce by applying ai. In: 2021 International Conference on Innovative Computing (ICIC). pp. 1–7 (Nov 2021)
2. AHMADOV, S., BOYACI, A.: Multilingual text mining based open source emotional intelligence. Turkish Journal of Science and Technology **17**(2), 161–166 (Sep 2022). https://doi.org/10.55525/tjst.1113832
3. Andrade-Segarra, D.A., León-Paredes, G.A.: Deep learning-based natural language processing methods comparison for presumptive detection of cyberbullying in social networks. International Journal of Advanced Computer Science and Applications **12**(5) (2021). https://doi.org/10.14569/ijacsa.2021.0120592
4. Bharathi V, D., Lakshmi Devi, R., Godfrey, E., Immanuel, S., Jose, S., Immanuel, S.: Challenges and opportunities in multilingual sentiment analysis: Beyond english. Kristu Jayanti Journal of Computational Sciences (KJCS) pp. 30–37 (Dec 2023). https://doi.org/10.59176/kjcs.v3i1.2310
5. Dwivedi, C., Rao, J.: Multilingual sentiment analysis for detecting mental health problems using a hybrid algorithm combining rnn and bi-lstm. In: 2023 7th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC). pp. 527–534. IEEE (Oct 2023). https://doi.org/10.1109/i-smac58438.2023.10290688
6. Jain, P., Agarwal, R.: Sentiment analysis using svm, naïve bayes, and lstm. In: 2023 5th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N). pp. 436–439 (Dec 2023)
7. Meghana, K.: Artificial intelligence and sentiment analysis in youtube comments: A comprehensive overview. In: 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT). pp. 1565–1572. IEEE (Jan 2024). https://doi.org/10.1109/idciot59759.2024.10467782
8. Nazir, M.K., Faisal, C.N., Habib, M.A., Ahmad, H.: Leveraging multilingual transformer for multiclass sentiment analysis in code-mixed data of low-resource languages. IEEE Access **13**, 7538–7554 (2025)
9. Rong, L., Weibai, Z., Debo, H.: Sentiment analysis of ecommerce product review data based on deep learning. In: 2021 IEEE 4th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). pp. 65–68 (Jun 2021). https://doi.org/10.1109/imcec51613.2021.9482223
10. Zikang, H., Yong, Y., Guofeng, Y., Xinyu, Z.: Sentiment analysis of agricultural product ecommerce review data based on deep learning. In: 2020 International Conference on Internet of Things and Intelligent Applications (ITIA). pp. 1–7 (2020). https://doi.org/10.1109/ITIA50152.2020.9312251