Sentiment Analysis of Flipkart Product Reviews using Machine Learning, Deep Learning and Large Language Models

1st Sumedha Arya
aryasumedha06@gmail.com

2nd Nirmal Gaud
ThinkAI - A Machine Learning Community
Bhopal, Madhya Pradesh, India
nirmal.gaud.ai@gmail.com

***Abstract*—This study focuses on sentiment analysis of Flipkart product reviews using machine learning (ML), deep learning (DL), and transformer-based large language models (LLMs). The objective is to classify reviews into positive, negative, or neutral categories. In the literature, we highlighted key challenges in the field, including limited dataset diversity, class imbalance, and insufficient feature extraction techniques. To address these, various models based on ML, DL, and LLMs were implemented and evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Among the models, RoBERTa demonstrated the highest performance (accuracy: 0.9452, F1-score: 0.95), followed closely by CNN-LSTM (accuracy: 0.9454, F1-score: 0.9466) and LSTM (accuracy: 0.9482, F1-score: 0.9482). Among traditional ML models, Random Forest achieved the best results (accuracy and F1-score: 0.9173). The findings suggest that RoBERTa is well-suited for applications requiring high precision, while CNN-LSTM offers a robust alternative in resource-constrained environments. Future research will focus on hyperparameter tuning, and multimodal data analysis.**

Index Terms—Sentiment Analysis, Flipkart Reviews, Machine Learning, Deep Learning, Transformer Models, RoBERTa, CNN-LSTM.

I. Introduction

In today's data-driven world, sentiment analysis has become an essential tool for organizations. It helps in understanding and responding to customer feedback across various digital platforms. In addition, it plays a critical role in marketing, customer support, and product development. This helps companies extract actionable insights from unstructured data such as product reviews and comments on social media [1]. As digital content continues to grow exponentially in volume and complexity, traditional methodologies become impractical, necessitating the use of automated and scalable sentiment analysis systems [2]. Product-based companies, such as Flipkart, benefit from sentiment analysis to refine strategies, improve offerings, and maintain a positive brand image [3], [4]. Despite advances, challenges such as limited dataset diversity, class imbalance, and language complexity reduce performance [5]. Recent research emphasizes the potential of ML, DL and LLM to address these issues by interpreting context-rich reviews [6], [7]. Deep learning techniques, such as CNNs and LSTMs, provide improvements by automatically extracting hierarchical features and modeling sequential short-term dependencies in the data. However, they still struggle to capture complex patterns with long-range dependencies in lengthy or complex reviews. Large Language Models (LLMs), built on transformer architectures, overcome many of these challenges by using attention mechanisms that capture global context, semantic richness, and linguistic complexities [6], [7]. However, they do have high computational requirements for training and execution. In this research, we compared ML, DL models, with LLMs. The results show that LLMs such as RoBERTa, demonstrate superior performance in handling contextual sentiment variations, multilingual inputs, and domain-specific text. The paper is structured as follows: the Literature Review discusses existing sentiment analysis approaches and their limitations; the Problem Statement defines the key challenges specific to Flipkart product reviews; the Research Methodology outlines the dataset, preprocessing techniques, and implementation of ML, DL, and LLM-based model; the Results Analysis presents and interprets the comparative performance of these models; and finally, the Conclusion and Future Work summarizes the findings and suggests potential directions for further research.

II. Literature Review

This section provides a comprehensive review of recent research on text-based sentiment analysis. Highlights key algorithms, methodologies, and techniques used to analyze sentiments from textual data across various domains such as finance, e-commerce, and social media. The authors [8] in their research performed sentiment analysis for the prediction of the stock market from 2018 to early 2023 using ML techniques. Their work explores how investor sentiment, derived from sources such as social media and news outlets, affects the dynamics of the stock market. They emphasized the role of ML models such as SVM, Naive Bayes, LSTM, CNN, and transformer-based architectures such as BERT and GPT to perform sentiment analysis. However, there are certain limitations of the research as it focused only on non-US stock markets. In addition, the best predictive model accuracy was only between 85–89%. The multimodal sentiment analysis (MSA) is done by the authors [9] by a novel approach in their work. They incorporated multi-layer feature fusion and multi-task learning, addressing challenges in extracting and fusing emotional information across different modalities. The proposed Unimodal Feature Extraction Network (UFEN) enables intra-modal feature extraction using convolutional layers and self-attention mechanisms, while the Multi-Task Fusion Network (MTFN) enhances inter-modal complementarity with cross-modal attention and transformer-based architectures. Experimental results on benchmark datasets show superior performance compared to state-of-the-art architectures, despite challenges in handling modal imbalances and implicit expressions. The study highlights limitations in dataset diversity and proposes future integration of physiological signals through non-contact sensors. The authors [10] explored sentiment analysis and emotion detection in both handwritten and electronic text (e-text) documents. They introduced sentiment insights in handwritten and e-text through the advanced ML model. This model utilizes SVM, Random Forest, Na¨ıve Bayes, Logistic Regression, Stochastic Gradient Descent, and CNN, LSTM techniques to classify sentiments and detect emotions across different datasets. Achieving over 90% accuracy, with CNNs, the study outperforms traditional methods by addressing challenges like handwriting variability, noise in scanned documents, and limited datasets through advanced preprocessing and feature extraction. The authors [11] examined the sentiment analysis of product reviews using traditional ML models and the latest LLMs. They focussed on their efficiency in classifying sentiments as positive, negative, or neutral. The study was conducted on a Flipkart dataset. It shows that SVM performed well in processing short, concise texts, while GPT-4 outperforms in analyzing longer reviews. The research shows the superiority of LLMs in explaining and its ability to handle complex language, making them valuable for customer-centric businesses. It also identifies challenges with unstructured and voluminous data. Aspect-Based Sentiment Analysis (ABSA), a key natural language processing task that identifies sentiments toward specific aspects within a given text is explored by the authors [12]. The study reviews 60 post 2019 articles, and analyzes traditional and deep learning approaches, including BERT, RNNs, and novel models like LadaBERT and Feature-Enhanced Attention CNN-BiLSTM. The proposed models optimize aspect extraction, sentiment polarity classification, and real-time processing while achieving quality performance on benchmark datasets. The review identifies gaps in current ABSA models, such as handling complex sentiment patterns, and emphasizes their applications in social media monitoring, brand reputation management, and customer feedback analysis. A comprehensive study of sentiment analysis, covering methods, applications, and challenges is done by the authors [13] in their work. They categorized techniques into lexicon-based, machine learning, and transformer-based approaches, including LLMs like ChatGPT, ERNIE Bot, and PaLM. The performance across document-level, sentence-level, and aspect-level is evaluated for all models. This study identified challenges such as the handling of multilingual texts, semantic nuances, and data scalability, proposing solutions such as advanced preprocessing and multimodal integration. The authors [14] examined sentiment analysis on e-commerce product reviews data, analyzing 20 studies from 2018 to 2024 using the PRISMA methodology. The research identifies key trends, cross-validation as the primary model testing method. The SVM and LSTM were found to be the most effective techniques, with LSTM and BERT-BiGRU achieving the highest accuracies, up to 98.43%. The review also addresses challenges such as fake reviews and limited language diversity. The authors [15] examined the sentiment analysis and classification of Flipkart product reviews using NLP and ML techniques. Their study used steps such as preprocessing, n-gram analysis, and TF-IDF vectorization for feature extraction, addressing class imbalance with the Synthetic Minority Oversampling Technique (SMOTE). ML models such as Logistic Regression, Decision Tree, K-Nearest Neighbor (KNN), and Na¨ıve Bayes algorithms were applied. Among them, Logistic Regression achieved the highest accuracy, which was validated through 10-fold cross-validation and evaluation metrics. The research focuses on the importance of automated sentiment analysis to understand customer behavior in e-commerce, providing insights from phrases such as "good product" and "worst experience ever." An investigation of sentiment analysis in e-commerce product descriptions, is done by the authors [16] in their work. They focussed on reducing asymmetry in customer-to-customer (C2C) platforms, particularly for high-priced items. This research utilizes a novel method called SATORE. It utilizes Latent Dirichlet Allocation (LDA) with Renyi Entropy to extract topics. Then sentiment scores are calculated, and integrated into a logistic regression model to examine their impact on customer's purchase intentions. The study analyzed over 4.4 million product descriptions from Mercari, a Japanese C2C platform. Their results show that context-sensitive emotional content in descriptions highly impacting buyer decisions. The authors [17] explored sentiment analysis for predicting product reviews in e-commerce recommendation systems, using deep learning and transformer-based modeling. The research utilized CNN, RNN, and Bi-LSTM models in comparison to traditional ML models such as Na¨ıve Bayes and SVM. They used word embeddings from BERT, RoBERTa, FastText, and Word2Vec on a dataset of clothing reviews. Results demonstrate that use of RoBERTa embeddings outperformed others models. In addition, the ensemble models achieved higher accuracy than the independent ones. The problem of information overload in e-commerce platforms was addressed by the authors in their research [18]. They suggested a hybrid framework which combines sentiment analysis and aspect-based sentiment analysis (ABSA). The research utilized TextBlob for sentiment polarity, NLTK for aspect extraction, and ML algorithms for classification. The customer reviews dataset is taken to classify sentiments and identify opinions on specific product features such as quality, price, and delivery. The framework improves classification accuracy and provides actionable insights for businesses. In similar research, the authors [19] also investigated ABSA to extract detailed insights from customer reviews on e-commerce platforms. They analyzed around 2,700 manually annotated reviews, identifying 14 key aspects of the purchase process that influences customer satisfaction. In research methodology, advanced ML models, like RoBERTa, were used, which achieved over 90% accuracy. The research findings show high potential of ABSA in providing actionable business insights, thus improving customer satisfaction and supporting strategic decisions. The authors [20] investigated sentiment analysis in online product reviews. They proposed a robust framework for sentiment classification with advanced NLP techniques. Their study combines feature engineering with word embedding techniques. In feature engineering extracting lexical, syntactic, and semantic features like n-grams, sentiment lexicons, part-of-speech tags were used. While in word embeddings, deep learning models, specifically CNNs and RNNs were used. The model was evaluated on a large dataset of product reviews, demonstrating superior performance in accuracy, recall, and F1-score across various sentiment classes and product domains. In summary, the literature highlights significant advancements in sentiment analysis through the use of traditional ML and DL algorithms along with LLMs. However, key challenges such as limited datasets, lack of feature extraction techniques, and low classification accuracy remain. These research challenges helped in formulation of the current research problem.

III. Problem Statement

Based on the literature review, the following key challenges were identified in text sentiment classification: (i) the limited availability of large, annotated, and diverse datasets, for model generalization; (ii) inadequate feature extraction and preprocessing techniques in existing models, leading to poor contextual understanding; (iii) the presence of class imbalance, which restricts the model's ability to learn equally from all sentiment categories; and (iv) the need for more accurate and efficient algorithms capable of capturing complex language patterns while maintaining high precision across multiple sentiment classes. Based on these problem statements, we aimed to apply various ML and DL algorithms along with LLMs on a benchmark dataset to demonstrate their effectiveness on text sentiment classification data.

IV. Research Methodology

The research methodology comprises data collection, preprocessing, exploratory data analysis (EDA), feature engineering, model development, and performance evaluation.

A. Data Collection

Dataset Source: The dataset is sourced from Kaggle. It contains product reviews with metadata including product name, price, rating, review summary, and sentiment labels (positive, negative, neutral). The analysis focuses on Summary for Textual content of the reviews and Sentiment for Categorical labels.

B. Data Preprocessing

The dataset is pre-processed to ensure quality and suitability for model training:

1) Data Cleaning:

- Removed 111,129 duplicate reviews.
- Dropped 2 rows with missing Summary values.
- Reset the DataFrame index and drop the index column.

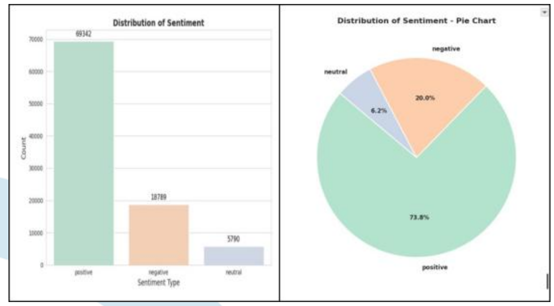

Fig. 1. Sentiment Distribution - Positive, Negative and Neutral.

2) Text Preprocessing: Applied two text cleaning approaches:

- **Initial Cleaning**: Converted text to lowercase, removed URLs, special characters, digits, and extra whitespace. Applied tokenization, stop word removal (using NLTK's stopwords), and lemmatization (using WordNetLemmatizer) to standardize text.
- **Transformer-Specific Cleaning**: For TinyBERT, ALBERT, and RoBERTa models, further simplified text by retaining only alphanumeric characters and spaces.

3) Class Balancing: Addressed class imbalance by upsampling minority classes (negative and neutral) to match the majority class (positive) using resampling technique. The balanced dataset contains 208,026 reviews (69,342 per sentiment class).

4) Feature Extraction:

- **Machine Learning Models**: Converted text to TF-IDF vectors using TfidfVectorizer with a maximum of 10,000 features.
- **Deep Learning Models**: Tokenized text using Tokenizer and padded sequences to a fixed length (max_len=128) using pad_sequences.
- **Transformer Models**: Tokenized text using BertTokenizer (for TinyBERT), AlbertTokenizer (for ALBERT), and RobertaTokenizer (for RoBERTa) with a maximum length of 50 tokens, suitable for short summaries. Generated input_ids and attention_mask for model input.

C. Exploratory Data Analysis (EDA)

EDA was conducted to understand dataset characteristics and sentiment distribution:

- **Sentiment Distribution**: Visualized using count plots and pie charts to confirm class balance post-sampling.
- **Word Cloud Analysis**: Generated word clouds for all sentiments using WordCloud, highlighting frequent terms.
- **Top Words Analysis**: Identified the top 10 words per sentiment class, revealing sentiment-specific vocabulary.

D. Model Development

The study implemented a comprehensive set of models for text sentiment classification, ranging from traditional ML algorithms to advanced DL architectures and transformer-based

Fig. 2. Word Cloud of All Sentiments.

language models. Each category of models was carefully designed and optimized to address the challenges of text sentiment classification.

1) Machine Learning Models: Machine learning models were implemented using scikit-learn with TF-IDF features. Logistic Regression was applied with default parameters for simplicity and robustness. A Decision Tree Classifier was configured with max_depth=200 and random_state=42 to balance learning complexity and reproducibility. Multinomial Na¨ıve Bayes was employed due to its effectiveness in handling text classification tasks. Random Forest was also implemented with n_estimators=10, max_depth=200, and random_state=42, using ensemble learning for improving accuracy.

2) Deep Learning Models: Deep learning models were implemented using TensorFlow and Keras, with text sequences padded to ensure uniform input lengths. A simple neural network was constructed with an embedding layer (128 dimensions), a flatten layer, and dense layers (10 units with ReLU activation and 3 units with softmax). Recurrent architectures included LSTM and BiLSTM, each with 64 units in the recurrent layer, preceded by an embedding layer and followed by dense layers for classification. To improve contextual modeling, LSTM and BiLSTM networks were also integrated with a custom attention mechanism, enabling the models to focus on relevant parts of the sequence. Dropout regularization (0.4) was used to prevent overfitting. Furthermore, CNN-LSTM hybrids were developed in two variants—LSTM and BiLSTM—featuring Conv1D layers (128 filters, kernel size=5), MaxPooling1D, recurrent layers, dense layers, and Dropout (0.5) for improved generalization.

3) Transformer-Based Models: Transformer-based models were implemented using the Transformers library and TensorFlow. Three pre-trained architectures—TinyBERT, ALBERT, and RoBERTa—were fine-tuned for sentiment classification. TinyBERT (huawei-noah/TinyBERT_General_4L_312D) was employed with a tokenizer (max_len=50), three output labels, and trained using AdamWeightDecay optimizer (learning rate=2e-5, weight decay=0.01) for 15 epochs with batch size 32. ALBERT (albert-base-v2) followed the same training configuration and hyperparameters. RoBERTa (roberta-base) was also fine-tuned with identical optimizer settings, training epochs, and batch size. All transformer models were trained with GPU acceleration to efficiently process large tokenized datasets.

E. Model Training and Validation

The dataset was split into 60% training, 20% validation, and 20% testing using traintestsplit, with randomstate=42 and stratification applied to maintain balanced class distributions. For transformer models, the inputids and attentionmask were split accordingly, while labels were one-hot encoded for training and preserved as integers for evaluation. Machine learning models were trained on TFIDF features, while deep learning models were trained on padded sequences. Transformer models were trained on tokenized inputs with GPU support (Tesla T4, 13942 MB memory). During training, both accuracy and loss were monitored across epochs for all models. Hyperparameters were selected to balance performance and computational efficiency. For machine learning models, default parameters were used for Logistic Regression and Na¨ıve Bayes, while max_depth and n_estimators were specified for Decision Tree and Random Forest. Deep learning models were trained for 15 epochs with batch size 64, embedding dimensions set to 128 (100 for CNN-LSTM), and maximum sequence length of 128. Transformer models were trained for 15 epochs with batch size 32, sequence length 50, and learning rate 2e-5.

F. Model Evaluation

The performance of all models was evaluated using accuracy, precision, recall, F1-score, confusion matrices, and training/validation curves. Accuracy measured the proportion of correctly classified reviews, while precision, recall, and F1-scores were computed per class and averaged using both weighted and macro methods. Confusion matrices were visualized with seaborn heatmaps using the Pastel2 colormap, providing insight into misclassification patterns. Additionally, training and validation curves were plotted for accuracy and loss across epochs using matplotlib with the Accent colormap, offering a clear view of model convergence and generalization behavior.

V. Results Analysis

Among the machine learning models, the Random Forest Classifier achieved the highest performance with an accuracy of 0.9173 and an F1-score of 0.9173. The Decision Tree Classifier followed with an accuracy of 0.8947 and an F1-score of 0.8947. Logistic Regression yielded a moderate accuracy of 0.8073 and an F1-score of 0.8075, limited by convergence issues that suggest potential improvements with hyperparameter tuning or data scaling. The Multinomial Naive Bayes model performed the least effectively, with an accuracy of 0.7664 and an F1-score of 0.7660, likely due to its simplistic assumptions about feature independence. Deep learning models outperformed machine learning models overall. The LSTM model achieved the highest accuracy of 0.9482 and an F1-score of 0.9482, showing its ability to capture sequential dependencies in text. The CNN-LSTM model closely followed with an accuracy of 0.9454 and an F1-score of 0.9466, benefiting from the combination of convolutional feature extraction and recurrent processing. The Bidirectional LSTM (BiLSTM) model recorded an accuracy of 0.9461 and an F1-score of 0.9460, slightly underperforming the

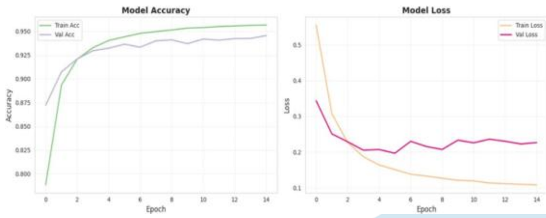

Fig. 3. CNN-LSTM Model Accuracy vs Loss.

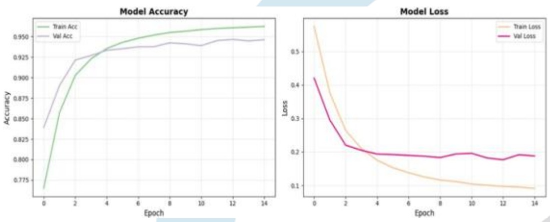

Fig. 4. RoBERTa Model Accuracy vs Loss.

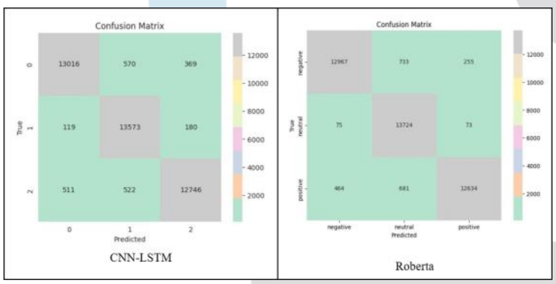

Fig. 5. Confusion Matrix of CNN-LSTM and RoBERTa Model.

LSTM due to increased complexity. The LSTM with Attention and BiLSTM with Attention models achieved accuracies of 0.9435 and 0.9400, with F1-scores of 0.9435 and 0.9399, respectively, indicating that attention mechanisms did not significantly enhance performance in this context. The Simple Neural Network, with an accuracy of 0.9311 and an F1-score of 0.9311, provided a baseline deep learning performance, surpassing most machine learning models but falling short of recurrent architectures. Transformer-based models demonstrated superior performance, reflecting their advanced capability to model complex text relationships. The RoBERTa model achieved the highest test accuracy of 0.9452 and a weighted F1-score of 0.95. Its class-specific metrics include negative (precision=0.96, recall=0.93, F1-score=0.94), neutral (precision=0.91, recall=0.99, F1-score=0.95), and positive (precision=0.97, recall=0.92, F1-score=0.94), indicating robust performance across all sentiments, particularly in identifying neutral reviews. The ALBERT model followed closely with a test accuracy of 0.9421 and a weighted F1-score of 0.94. Its class-specific metrics (negative: precision=0.95, recall=0.93, F1-score=0.94; neutral: precision=0.91, recall=0.99, F1-score=0.95; positive: precision=0.97, recall=0.91, F1-score=0.94) are comparable to RoBERTa but slightly less consistent. TinyBERT recorded a test accuracy of 0.9344 and a weighted F1-score of 0.93. Its class-specific

TABLE I
A COMPARATIVE SUMMARY OF MODEL PERFORMANCE.

Model	**Test Accuracy**	**F1-Score (Weighted)**
Machine Learning Models		
Logistic Regression	0.8073	0.8075
Decision Tree	0.8947	0.8947
Naive Bayes	0.7664	0.7660
Random Forest	0.9173	0.9173
Deep Learning Models		
Simple Neural Network	0.9311	0.9311
LSTM	0.9482	0.9482
BiLSTM	0.9461	0.9460
LSTM with Attention	0.9435	0.9435
BiLSTM with Attention	0.9400	0.9399
CNN-BiLSTM	0.9450	0.9449
CNN-LSTM	0.9454	0.9466
Transformer-Based Models		
TinyBERT	0.9344	0.93
ALBERT	0.9421	0.94
RoBERTa	0.9452	0.95

metrics (negative: precision=0.94, recall=0.92, F1-score=0.93; neutral: precision=0.91, recall=0.97, F1-score=0.94; positive: precision=0.96, recall=0.90, F1-score=0.93) indicate slightly lower performance, particularly for positive reviews. The comparative analysis shows that transformer-based models generally outperform machine learning and deep learning models, with RoBERTa achieving the highest test accuracy (0.9452) and F1-score (0.95), closely followed by the CNN-LSTM (0.9454, 0.9466) and LSTM (0.9482, 0.9482) models. RoBERTa's superior performance is attributed to its robust pre-trained architecture, optimized for natural language understanding, and its ability to effectively capture contextual relationships in short review texts. The high validation accuracy (0.9462) and the low validation loss (0.1879) of RoBERTa indicate stable generalization, although its high training accuracy (0.9624) suggests a slight risk of overfitting, similar to ALBERT (training accuracy=0.9625). The CNN-LSTM and LSTM models offer competitive performance with lower computational complexity, making them viable alternatives for resource-constrained environments. Among machine learning models, Random Forest performs best but falls short of deep learning and transformer models, likely due to the limitations of TF-IDF features in capturing semantic nuances. The RoBERTa model is recommended as the best model for sentiment classification in this study due to its highest accuracy. The CNN-LSTM model is a compelling alternative, offering comparable performance with lower resource demands. The choice between RoBERTa and CNN-LSTM should consider the trade-off between accuracy and computational efficiency, with RoBERTa preferred for high-accuracy requirements and CNN-LSTM for resource-constrained settings.

VI. Conclusion and Future Work

This study presents a comprehensive methodology for sentiment classification of product reviews, integrating traditional machine learning, deep learning, and transformer-based models.

The RoBERTa model achieved the highest performance, outperforming other models, followed closely by the CNN-LSTM model and ALBERT. This research contributes to advances in natural language processing, offering robust solutions for automated sentiment analysis in e-commerce applications. In future research, we will address identified limitations to enhance model performance and generalizability. Advanced preprocessing techniques, such as n-grams or stemming, could improve the quality of the features. Experimentation with additional transformer models or embeddings may yield further improvements. Hyperparameter optimization via grid or random search, alongside regularization and data augmentation, could overcome overfitting observed in deep learning and transformer models. Testing models on diverse datasets would enhance generalizability beyond product reviews.

REFERENCES

[1] S. Chaturvedi, V. Mishra, and N. Mishra, "Sentiment analysis using machine learning for business intelligence," in *Proc. 2017 IEEE Int. Conf. Power, Control, Signals and Instrumentation Eng. (ICPCSI)*, Chennai, India, Sep. 21–22 2017, pp. 2162–2166.

[2] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, pp. 5731–5780, 2022.

[3] G. Bonifazi, F. Cauteruccio, E. Corradini, M. Marchetti, G. Terracina, D. Ursino, and L. Virgili, "A framework for investigating the dynamics of user and community sentiments in a social platform," *Data Knowl. Eng.*, vol. 146, p. 102183, 2023.

[4] F. Cauteruccio and Y. Kou, "Investigating the emotional experiences in esports spectatorship: The case of league of legends," *Inf. Process. Manag.*, vol. 60, p. 103516, 2023.

[5] R. K. Schmidt, "Automatic document classification in technical logbooks: A comparative study of supervised, weakly supervised and unsupervised machine learning approaches," Master's thesis, Universidade NOVA de Lisboa, Lisboa, Portugal, 2024.

[6] M. F. Abdussalam, D. Richasdy, and M. A. Bijaksana, "Bert implementation on news sentiment analysis and analysis benefits on branding," *J. Media Inform. Budidarma*, vol. 6, pp. 2064–2073, 2022.

[7] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intell. Syst.*, vol. 28, pp. 15–21, 2013.

[8] J. Chandra and A. C. Mondal, "Studies of sentiment analysis for stock market prediction using machine learning: A survey towards new research direction," *Scholars Journal of Engineering and Technology*, vol. 13, no. 1, p. 007, 2025.

[9] Y. Cai, X. Li, Y. Zhang, J. Li, F. Zhu, and L. Rao, "Multimodal sentiment analysis based on multi-layer feature fusion and multi-task learning," *Scientific Reports*, vol. 15, no. 2126, 2025.

[10] R. Ahamad and K. N. Mishra, "Exploring sentiment analysis in handwritten and e-text documents using advanced machine learning techniques: a novel approach," *Journal of Big Data*, 2025.

[11] P. S. Ghatora, S. E. Hosseini, S. Pervez, M. J. Iqbal, and N. Shaukat, "Sentiment analysis of product reviews using machine learning and pre-trained llm," *Big Data and Cognitive Computing*, vol. 8, no. 12, p. 199, 2024.

[12] A. Goud and B. Garg, "Advancements in aspect-based sentiment analysis: Leveraging deep learning and machine learning techniques," *Journal of Information Systems Engineering and Management*, vol. 10, no. 21s, 2025.

[13] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *Journal of King Saud University - Computer and Information Sciences*, vol. 36, p. 102048, 2024.

[14] A. Daza, N. D. Gonza´lez Rueda, M. S. Aguilar Sa´nchez, F. Robles Esp´ıritu, and M. E. Chauca Quin˜ones, "Sentiment analysis on e-commerce product reviews using machine learning and deep learning algorithms: A bibliometric analysis, systematic literature review, challenges and future works," *International Journal of Information Management Data Insights*, vol. 4, p. 100267, 2024.

[15] A. Godia and L. K. Tiwari, "Sentiment analysis and classification of product reviews: A comprehensive study using nlp and machine learning techniques," in *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, Mar 2024.

[16] Y. Sun, K. Sekiguchi, and Y. Ohsawa, "Optimizing sentiment analysis in product descriptions: effects on customer purchase intentions," *Information Technology and Management*, Feb 2025.

[17] O. Bellar, A. Baina, and M. Ballafkih, "Sentiment analysis: Predicting product reviews for e-commerce recommendations using deep learning and transformers," *Mathematics*, vol. 12, no. 15, p. 2403, 2024.

[18] K. Kumar, S. Singh, K. Tyagi, S. Chaudhary, N. Prakash, and P. Dhanraj, "Sentiment analysis and opinion mining of amazon reviews," Meerut Institute of Engineering and Technology, Meerut, India, Tech. Rep., 2024.

[19] L. Davoodi, J. Mezei, and M. Heikkila¨, "Aspect-based sentiment classification of user reviews to understand customer satisfaction of e-commerce platforms," *Electronic Commerce Research*, Jan 2025.

[20] L. Bharadwaj, "Sentiment analysis in online product reviews: Mining customer opinions for sentiment classification," *International Journal for Multidisciplinary Research (IJFMR)*, vol. 5, no. 5, p. 1, Sep–Oct 2023.