

*Digital Object Identifier*

SENTIMENT TRENDS IN WALMART SERVICES: A DATA DRIVEN Approach Using HYBRID MODELS, XLNET, LSTM,LOGISTIC REGRESSION, and CNN

**Dr. MARY SHAMALA .L , SAIRAM GOPAL NIDUMOLU ,D L K TRINADH, RISHI PATRI.**

School of Computer Science and Engineering, Vellore Institute of Technology, Chennai 600127, India

Corresponding author: Mary Shamala L (e-mail: [maryshamala.l@vit.ac.in).](mailto:maryshamala.l@vit.ac.in)

 **ABSTRACT** In this project, we present a comprehensive sentiment analysis framework tailored to assess customer perceptions of Walmart's services using real-world review data. Leveraging reviews sourced from Trustpilot through web scraping, we perform extensive preprocessing, including cleaning and tokenization, to prepare the dataset for emotion and sentiment analysis. Utilizing the zero-shot classification capabilities of the pre-trained facebook/bart-large-mnli model, we categorize each review into ten fine-grained emotion labels: happy, joyous, sad, angry, frustrated, excited, disappointed, hopeful, relieved, and surprised. To deepen our analysis, we implement and compare a variety of machine learning and deep learning models— Logistic Regression, CNN, LSTM, RoBERTa, XLNet, and a hybrid model combining Bi-LSTM, BERT, and Attention mechanisms—for their effectiveness in emotion classification. Recognizing that customer feedback often spans multiple service dimensions, we incorporate aspect-based sentiment analysis to identify key focus areas such as Customer Service, Delivery & Orders, In-Store Experience, and Pricing & Billing. Using facebook/bart-large-mnli, we extract aspects and compute confidence scores, followed by sentiment classification through cardiffnlp/twitter-roberta-base-sentiment-latest. This dual-layered approach enables us to determine the sentiment polarity (positive, negative, or neutral) associated with each aspect, offering granular insights into Walmart's service performance from the customer perspective. The results provide valuable implications for customer experience management and service improvement strategies.

 **INDEX TERMS** Diabetic Retinopathy (DR), Interpretability, Trustworthy, Explainable AI (XAI), Mo- bileNet, Graph Neural Network (GNN), Recurrent Neural Network (RNN), Grad-CAM, Asia Pacific Tele- Ophthalmology Society (APTOS).

1. **INTRODUCTION**

Sentiment analysis is an effective NLP method utilized for assisting enterprises to understand customers' emotion, opinion, and sentiment. Reviews made by customers play a critical role in the calculation of service quality, responding to pain areas, and succeeding in improving retail organizations. Walmart is the largest chain of retailers receiving thousands of comments on multiple platforms, which helps in understanding diverse aspects of its services. Sentiment analysis of Walmart's services has also been conducted in this study through web scraping of customers' reviews on Trustpilot. The data gathered is tokenized and pre-processed to ensure consistency and uniformity before classification

In Hugging face, we have used a model by Facebook’s BART large MLNI Model, a zero-shot classifier, we classify reviews into ten emotion labels: happy, joyous, sad, angry, frustrated, excited, disappointed, hopeful, relieved, and surprised. Other than that, aspect-based sentiment analysis (ABSA) is performed using CardiffNLP’s Twitter- RoBERTa Model enabling us to find sentiment polarity (positive, negative, or neutral) for key service aspects such as customer service, delivery & orders, in-store experience, and pricing & billing.

To improve the accuracy in classification, several machine learning and deep learning models such as logistic regression, CNN, LSTM, RoBERTa, XLNet, and a hybrid model (Bi- LSTM + BERT + Attention Mechanism) are utilized and compared in terms of accuracy, precision, recall, and F1-score. Sarcasm, ambiguous sentences, and class imbalance, which may degrade the performance of classification, are among the most challenging problems for sentiment analysis. Besides, retail sentiment analysis must be domain-adapted to harvest the subtlety of customer interaction. Through aspect-based sentiment classification, this research endeavours to offer concrete recommendations under which Walmart can make optimal service delivery, handle issues with customers effectively, and improve overall customer satisfaction.

* 1. ***RESEARCH OBJECTIVE***

The rapid growth of e-business has led to a mushrooming of customer-generated content, including reviews, comments, ratings, and postings on social media. All these vast streams of feedback contain a treasure of information pertaining to likes, dislikes, satisfaction levels, and expectations of the customers. Yet, analyzing so vast an amount of unstructured data is labor-intensive and inefficient to do by hand. Such an issue provides a strong rationale for employing automated techniques to analyze and interpret customers' sentiment quickly and efficiently.Sentiment analysis fulfils this need by enabling businesses to examine customer opinions and emotions conveyed on various internet sites in a structured way.

Customer sentiment is adopted by online businesses in an effort to improve their services and offerings, increase the customer experience, and adjust their marketing strategies based on feedback in real-time. Moreover, the ability to monitor and react to public sentiments helps businesses manage their brand reputation more effectively in a competitive market.

By leveraging the sentiment analysis, the online shopping websites are able to move beyond the traditional feedback system mechanism and gain access to the vast online customer information bank. It not only aids in taking informed, data-driven decisions, but it also results in more cohesive customer relationships and continued business growth

The primary objective of this research is to analyze and predict sentiment trends in customer reviews related to Walmart services by leveraging a hybrid approach that combines deep learning and traditional machine learning models. With the rise of e-commerce and digital retail platforms, customer feedback has become a crucial component in understanding consumer satisfaction and business performance. This study aims to harness the power of advanced Natural Language Processing (NLP)

techniques to extract meaningful insights from unstructured text data. This study seeks to leverage advanced Natural Language Processing (NLP) models to extract and interpret sentiments from large volumes of unstructured textual data. The research involves the implementation of multiple models including XLNet, LSTM, CNN, and Logistic Regression, each offering unique advantages in capturing semantic, sequential, and structural features of the text. Furthermore, a key aim is to develop and evaluate hybrid models that combine the strengths of these individual approaches, with the goal of improving sentiment classification accuracy and robustness. By visualizing sentiment trends over time and across various service dimensions such as delivery, customer support, and pricing, the study also aims to generate actionable insights for Walmart’s strategic decision-making. Additionally, model interpretability techniques will be employed to identify influential keywords and patterns that drive customer sentiment, thereby contributing to a more nuanced understanding of consumer behavior. Overall, the research aspires to provide a comprehensive and scalable sentiment analysis framework tailored to large-scale retail services. The core objective of this research is to investigate and understand sentiment trends in customer feedback related to Walmart services by utilizing a comprehensive, data-driven approach rooted in both traditional and modern machine learning paradigms. As customer opinions shared through online platforms have become a significant factor in shaping a company’s public image and strategic direction, it is essential to systematically analyze these sentiments to uncover patterns, expectations, and areas of concern

* 1. ***RESEARCH PROBLEM***

The current age of digitalization witness’s customer reviews as the deciding factor in determining business strategies and improving service quality. Walmart, being a large retail chain, receives thousands of reviews on a single day from customers providing their opinions on customer experiences across various service areas such as customer service, delivery & orders, in-store experience, and pricing & billing However, it is not feasible to analyze these reviews manually due to their vast volume of data. Traditional sentiment analysis methods tend to classify reviews into broad categories such as positive, negative, or neutral and lose the implicit emotional background and specific service information that elicit customers' satisfaction or dissatisfaction. Retail sentiment analysis also comes with challenges such as handling sarcasm, usage of vague language, and imbalance in data among sentiments. In order to overcome such limitations, the present study is seeking to embrace aspect-based sentiment analysis (ABSA) for categorizing Walmart service reviews in a comprehensive sense.

Utilizing zero-shot classification with Facebook’s BART MNLI Model, we categorize reviews into ten emotion categories in order to study customers' emotions deeper. Further, we seek out crucial aspects of Walmart's services and apply CardiffNLP’s Twitter-RoBERTa Model to discover the polarity of the sentiment (positive, negative, or neutral) for each aspect. Besides, some of the most popular machine and deep learning architectures, including logistic regression, CNN, LSTM, RoBERTa, XLNet, and a combination of Bi- LSTM + BERT + Attention mechanism, are employed in an attempt to improve the classification accuracy. The objective of this research is to provide Walmart with a data-driven approach so that customer reviews can be analyzed more efficiently, improving decision-making, improving customer satisfaction, and optimizing services.

***CONTRIBUTION AND IMPACT***

The contributions and impact of this study were summarized in :

The project aims to conduct sentiment analysis of Walmart services based on customer reviews gathered from Trustpilot using web scraping. The main aim is to derive meaningful insights from the reviews by labelling the customer sentiments and determining the most prominent factors of Walmart's services that result in customer satisfaction or dissatisfaction. The task employs zero-shot classification with Facebook’s BART MNLI Model to classify reviews into 10 emotion labels, i.e. happy, joyous, sad, angry, frustrated, excited, disappointed, hopeful, relieved, and surprised. Further, ABSA is also carried out using CardiffNLP’s Twitter-RoBERTa-base-sentiment-latest Model for the identification of sentiment polarity (positive, negative, or neutral) for various service aspects like customer service, delivery & orders, in-store experience, and pricing & billing. To improve the performance of classification, the project tests several machine learning and deep learning models such as logistic regression, CNN, LSTM, RoBERTa, XLNet, and a Bi-LSTM + BERT + Attention hybrid model. The models are compared against such critical performance indicators as accuracy, precision, recall, and F1-score in attempting to determine the best method to be used for sentiment classification. A key technical contribution of this study lies in the integration of deep learning models with traditional machine learning methods in a unified pipeline, demonstrating that hybrid strategies can outperform standalone models in both performance and interpretability. The research also presents a robust data preprocessing and sentiment labeling pipeline tailored to noisy, unstructured customer review data, which can be easily adapted to other domains beyond Walmart.

From a practical standpoint, this study provides valuable insights into customer sentiment trends across various service dimensions—such as delivery efficiency, pricing, product quality, and customer support. These insights, derived from model interpretation and trend analysis, can help Walmart and similar enterprises identify service gaps, understand customer needs, and prioritize areas for improvement. Furthermore, the temporal analysis of sentiment fluctuations offers a strategic advantage in monitoring service performance over time and responding proactively to customer concerns.

1. **RELATED WORKS**

Jabbar et al. [1] present research on real-time sentiment analysis for e-commerce applications, focusing on product reviews from Amazon.com. The study designs and implements a model using a SVM, a machine learning technique, to predict sentiment polarity at both review and sentence levels.

Arora et al. [2] explore the temporal and sentimental dynamics of customer reviews from an Indian e-commerce platform between 2019 and 2023. The study employs NLP techniques with TextBlob and statistical tools like ANOVA to analyze evolving emotions across time. Word cloud visualizations and sentiment classification highlight seasonal variations and feature-specific feedback. While the approach effectively uncovers sentiment trends, its reliance on basic NLP tools may limit accuracy in handling complex linguistic nuances, suggesting the need for more advanced language models in future work.

Sundaram et al. [3] propose an aspect sentiment analysis system for e-commerce using triplet extraction and a BiLSTM-based model. The approach identifies opinion- aspect-sentiment triplets from customer reviews and leverages VADER and syntactic rules for sentiment classification. With an AUC of 0.9 and an average precision of 0.89, the system shows strong performance, including sarcasm detection. However, the reliance on handcrafted preprocessing and basic LSTM layers may limit generalization, highlighting potential for further enhancement through more robust attention-based or generative models.

Wahyudi et al. [4] explore aspect-based sentiment analysis in e-commerce reviews using Latent Dirichlet Allocation (LDA) with Collapsed Gibbs Sampling. The study evaluates sentiment classification accuracy across product categories, comparing general versus category-specific training data. The best performance, with 67.5% accuracy, was achieved using LDA with α=0.001, β=0.001, and 15 topics. Combining general and category-based data slightly improved accuracy by 0.82%. However, the reliance on probabilistic modeling limits linguistic nuance capture, suggesting potential gains through integration with deep learning techniques.

Zulfiker et al.[5] present a machine learning approach for Bangla sentiment analysis on e-commerce reviews. The study constructs a labelled Bangla corpus and evaluates six classifiers using TF-IDF with trigram features. Among them, a hyperparameter-tuned SVM model achieves the highest accuracy of 90.68%. The approach demonstrates strong performance across all metrics. However, the limited dataset size (1,631 reviews) and manual labelling may affect scalability and domain adaptability, suggesting the need for larger, more diverse corpora and automated annotation methods in future work.

Guru Prasad et al. [6] investigate sentiment analysis on e- commerce platforms using transfer learning with transformer-based models. The study applies BERT and its variants for aspect-based sentiment classification, using data from Amazon and Flipkart. The approach demonstrates strong contextual understanding and scalability across different models. However, the large model sizes, computational demands, and challenges in domain adaptation and nuanced aspect detection underscore the trade-offs between performance and efficiency, suggesting the need for careful model selection based on task complexity and resource constraints.

Bharti et al. [7] review efficient machine learning techniques for sentiment analysis of e-commerce customer reviews. The study explores models such as CNN, LSTM, and RMDL, alongside embedding methods like BERT, ELMo, and FastText, to categorize reviews into sentiment classes. It emphasizes the importance of preprocessing and contextual understanding. While the models show promising performance, their effectiveness heavily depends on data quality, feature extraction, and model tuning, suggesting the need for robust hybrid frameworks and domain-specific enhancements.

Chaudari et al. [8] present a sentiment analysis framework for Zomato restaurant reviews using Bi-LSTM and Bi-GRU models. The study combines Word2Vec embeddings with VADER and Sentiment Intensity Analyzer to enhance performance. Experimental results show Bi-GRU outperforms Bi-LSTM in generalization and accuracy, achieving 99.51%. While effective, the Bi-LSTM model showed signs of overfitting, and results varied with dropout rates and batch sizes, highlighting the importance of careful hyperparameter tuning and model selection.

Hakkinen et al. [9] perform sentiment analysis on Tokopedia e-commerce reviews using a lexicon-based method and Naïve Bayes classifier. The study focuses on handling negation, intensifiers, and complex expressions to improve sentiment scoring. With a 400-review dataset, results show 57.5% positive, 22.5% negative, and 20%

neutral responses. While the approach is efficient for quick sentiment detection, it may struggle with contextual ambiguity and nuanced expressions, suggesting the need for hybrid or semi-supervised models for better accuracy

Kathuria et al. [10] analyze sentiment in fashion e-commerce using ML and NLP techniques to study consumer behavior. They apply models like logistic regression, SVM, random forest, and sentiment tools like VADER and TextBlob on reviews and ratings. Logistic regression achieved the highest accuracy (88.18%) for reviews and 80.68% for ratings. While effective, the study highlights challenges in data cleaning, subjectivity handling, and over-attribution of positivity in lexicon-based models, suggesting potential improvements through advanced contextual models.

Al Omari et al. [11] propose a logistic regression-based sentiment classifier for Arabic service reviews in Lebanon. Using TF-IDF features on 3,916 manually collected reviews from Google and Zomato, the model performed well on positive reviews but struggled with negative ones due to data imbalance. Precision for negative sentiment was 0.80, but recall dropped to 0.08, resulting in an ROC AUC of just 0.54. The study highlights challenges in Arabic NLP, particularly sarcasm, code-mixing, and linguistic ambiguity, indicating the need for more balanced datasets and advanced modeling.

Bitto et al. [12] analyze customer sentiment toward Bangladeshi courier services using Bangla NLP and machine learning models. Reviews were scraped from social media for Sundarban, Redx, and Pathao, then classified using six algorithms and N-gram features. Multinomial Naive Bayes with Bigram features achieved the highest accuracy of 90.72%. The study effectively highlights sentiment trends, but limited data size, manual labelling, and underperformance of models like SVM suggest future work with deep learning and larger datasets for enhanced robustness.

Sari et al. [13] apply sentiment analysis to evaluate e-commerce service quality using user reviews from Tokopedia. The study classifies sentiments across five e-Servqual dimensions— reliability, responsiveness, trust, web design, and personalization—using a Naïve Bayes classifier with TF-IDF features. Results show an overall 90% accuracy, with trust and web design receiving the highest positive sentiment. However, personalization was dominated by negative feedback, indicating a need for improved customer engagement. While efficient, the approach is limited by temporal data scope and absence of social media inputs.

Mugil et al. [14] explore a trust-based product analysis model for e-commerce platforms using a hybrid STRUMKNN approach. The model combines sentiment analysis and word separation techniques to assess product reviews and generate trust profiles. By evaluating feedback through polarity-based classification, the system provides product suggestions based on review authenticity and trust weight. Comparative analysis with SVM and K-Means clustering shows that STRUMKNN achieves higher accuracy (92%) in sentiment classification and trust evaluation. While the model demonstrates strong performance, its rule-based sentiment mechanism may limit adaptability to nuanced or sarcastic language, indicating the potential benefit of integrating deep learning-based NLP models in future enhancements.

Panduro et al. [15] investigate the role of sentiment analysis in enhancing product recommendation systems within e- commerce platforms. The study applies various machine learning techniques—such as logistic regression, support vector machines (SVM), and deep learning models like CNN and LSTM—to classify customer review sentiments as positive, negative, or neutral. The methodology includes text preprocessing, feature extraction (e.g., TF-IDF, Word2Vec), and model evaluation using metrics like precision, recall, and F1-score. Experimental results show accuracy levels ranging from 80–90%, with logistic regression performing best. While the approach effectively personalizes recommendations based on user sentiment, challenges remain in handling fake reviews and complex linguistic patterns, suggesting the need for advanced multimodal and ethical AI models in future research.

Taneja et al. [16] explore customer sentiment in Indian fashion e-commerce using a BERT-based deep learning model. Focusing on reviews from platforms like Myntra, Ajio, and Tata Cliq, the study employs BERT's bidirectional contextual embeddings to classify sentiments into positive, neutral, or negative categories. The dataset, comprising over 7,500 reviews, is preprocessed and normalized before classification. The model architecture includes a dropout layer and a fully connected layer atop BERT to predict sentiment classes, achieving an accuracy of 92%, outperforming DistilBERT. While the approach effectively captures nuanced expressions in reviews, occasional misclassifications, especially in neutral cases, highlight challenges posed by subjective human language. Future work aims to incorporate multilingual analysis to cater to India’s diverse linguistic landscape, further improving sentiment classification in regional e-commerce contexts. Mboungou et al. [17] examine sentiment patterns in customer feedback on a French e-commerce platform, leveraging a dataset of 56,000+ reviews across 16 product categories. The study utilizes advanced machine learning techniques, notably Recurrent Neural Networks (RNNs) and

Long Short-Term Memory (LSTM) models, for emotion classification. Preprocessing includes stop-word removal,

lemmatization, and feature extraction using TF-IDF. The system maps sentiments to aspects such as price, quality, shipment, design, and satisfaction. While LSTM aids in capturing long-term dependencies and subtle nuances in French-language reviews, challenges such as sarcasm and ambiguity still require human oversight. Overall, the approach enhances sentiment accuracy and offers actionable insights for businesses aiming to refine product offerings and customer engagement strategies.

Casas-Valadez, et al. [18] conduct a comprehensive bibliometric analysis to investigate the synergy between e- commerce and sentiment analysis in published research from 2007 to 2020. The study utilizes data from Scopus and applies performance metrics and co-occurrence mapping using tools like VOSviewer. It identifies three dominant research clusters: Data Collection & Analysis, Business Intelligence, and Lexicon-based Methods. These themes highlight the evolution of sentiment analysis from basic opinion mining to advanced AI-driven insights.

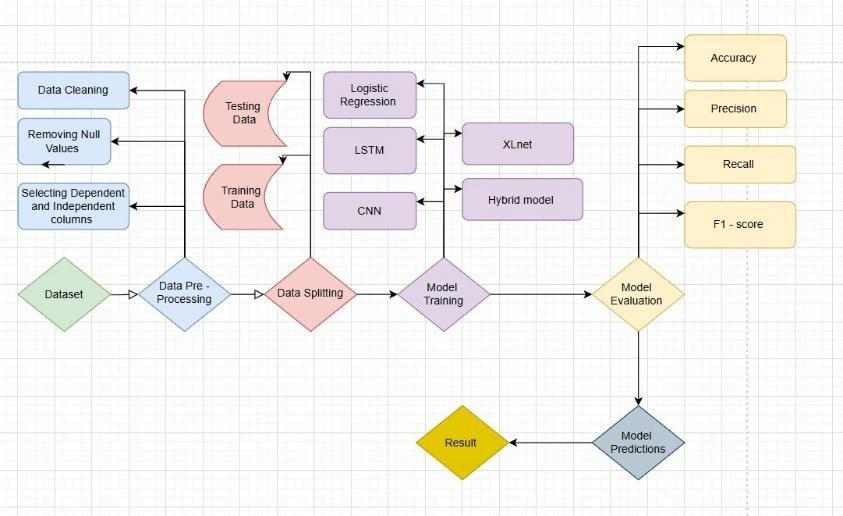
Jahnavi et al. [19] propose a hybrid sentiment analysis model for social media platforms by combining BERT and RoBERTa architectures. The study utilizes a Kaggle-based labeled dataset enriched with temporal and demographic metadata. Extensive preprocessing, exploratory data analysis, and word cloud visualizations are conducted to uncover textual characteristics. Evaluation with metrics such as accuracy, precision, recall, and F1-score demonstrate the model’s superiority over traditional methods like SVM, CNN, and standalone BERT, achieving 82% accuracy.

Samonte et al. [20] explore public sentiment during the COVID-19 pandemic by implementing a hybrid deep learning model that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to analyze Twitter data. The study focuses on capturing spatial features with CNN and temporal dependencies with LSTM to enhance sentiment classification. Tweets were preprocessed through tokenization, stopword removal, and lemmatization, followed by word cloud visualizations and class distribution analysis. The hybrid model outperformed traditional machine learning algorithms in accuracy and contextual understanding. However, its reliance on pre-COVID datasets and general-purpose embeddings may restrict adaptability to evolving slang and domain-specific expressions, indicating the need for dynamic and fine-tuned embedding techniques in future work.

1. **METHODOLOGY**

The proposed methodology has introduced a novel frame- work as shown in Figure 1.

One of the most important natural language processing (NLP) methods by which organizations are able to determine what their customers think, feel, and say about their organization is sentiment analysis. Sentiment analysis of Walmart services, in this present study, we will attempt to perform sentiment analysis of Walmart services by sentimentally analyzing the customer reviews gathered from Trustpilot by web scraping. The traditional sentiment analysis categorizes reviews as positive, negative, or neutral but fails to fully understand customers' emotional issues and location-based service issues. To address this limitation, we use zero-shot classification with Facebook/BART-Large- MNLI to classify reviews into ten emotional categories: joyous, happy, sad, angry, frustrated, excited, disappointed, hopeful, relieved, and surprised. We also employ aspect- based sentiment analysis (ABSA) with CardiffNLP/Twitter- RoBERTa-base-sentiment-latest that allows us to measure the sentiment polarity (positive, negative, or neutral) for specified service areas such as customer service, delivery & orders, in-store experience, and pricing & billing. For better sentiment classification accuracy, we leverage multiple machine learning and deep learning models, i.e., logistic regression, CNN, LSTM, XLNet, and a Bi-LSTM + BERT + Attention mechanism hybrid. These models help in improved sentiment prediction where customer feedback is labelled appropriately. Watching the trends of the sentiments across various domains, the current work provides valuable information pertaining to customer satisfaction levels, pain areas, and improvement areas. Walmart can possibly employ informed decisions to enhance their service as they try to improve their services and make shopping an improved experience by utilizing such findings. This chapter discusses in detail the data collection, preprocessing, sentiment classification, and aspect-based sentiment analysis techniques used in this research. Performance was measured using standard metrics including **accuracy, precision, recall, F1-score**, and **confusion matrices**. Visualizations such as **word clouds**, **sentiment distribution charts**, and **attention heatmaps** (for XLNet and the hybrid model) provided interpretability and deeper insights into sentiment trends across Walmart services. Each model was trained using stratified train-test splits (80/20), with hyperparameter tuning via grid search or validation techniques. Key parameters like learning rate, batch size, number of layers, and dropout rate were optimized to enhance generalization and reduce overfitting.

FIGURE 1: Framework of Proposed Methodology

* 1. ***DATASET***

In order to conduct the sentiment analysis of Walmart services, we first collected customer review data from Trustpilot, a well- known review website. Since there was no clear API on Trustpilot to review, we used web scraping techniques in order to get data. Web scraping is programmable text data extraction from the web using BeautifulSoup and Scrapy type tools, which enable us to get structured data effectively. It was done by discovering Walmart's Walmart review page's HTML structure that is pointing to owners of review text, ratings, timestamps, and customer comments. We created a Python scraping script where reviews would be fetched taking into account support with Walmart's robots.txt file and rate limiting to avoid overloading the server. Secondly, since it was to deal with dynamic load contents, we have used Selenium since it makes web browsers automatable whenever we are dealing with used-content JavaScript loads in scrapes.

* 1. ***PROPOSED ALGORITHM***

## Logistic Regression

Logistic Regression is a popular machine learning algorithm for binary and multi-class classification problems, including sentiment analysis. It is a linear model that estimates the probability of a binary outcome using a logistic function on a weighted sum of input features. In sentiment analysis, Logistic Regression is used to classify text (e.g., Walmart reviews) into sentiment classes, e.g., positive, negative, and neutral. Logistic Regression achieves this through the initial conversion of the text data into numerical features that can be input into the model. Typically, methods used for extracting features from text include TF-IDF (Term Frequency-Inverse Document Frequency), Bag of Words, or word embeddings. These features extract information on the frequency and significance of words or terms in the text so that the model can recognize key information on sentiment.

## Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are deep neural networks that have become increasingly popular for application in image processing but also in natural language processing (NLP) applications like sentiment analysis. CNNs are employed in sentiment analysis to detect neighbourhood patterns and co-occurrences of text via convolutional filters on the input and, in this case study, customer review text.

The convolutional layers work by scanning the input text with filters and producing feature maps to identify significant patterns in the review. The feature maps are then passed through a max-pooling layer, which reduces the feature maps by taking the most significant information, e.g., the most indicative words or phrases of sentiment. The output of the convolutional layers is flattened and passed to fully connected layers to produce a final classification output, in which the review is mapped to a sentiment label (positive, negative, neutral). One of the strongest advantages of sentiment analysis with CNNs is that they can automatically learn and obtain good features from the text, without tedious manual feature engineering typically required with other methods. This is especially valuable for sentiment analysis, because context and local patterns inside sentence play such a big role. CNNs are engineered to capture useful local dependencies, like the words or n-grams that convey sentiment and affective tone.

## Long Short-Term Memory (LSTM) Networks

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Networks (RNNs) which are specifically developed to address problems of typical RNNs, especially long-term dependencies and vanishing gradients. LSTMs perform very well for sequential inputs such as text in which the context of a sentence is based upon words that occur much earlier or further.In sentiment analysis, e.g., Walmart review analysis, LSTMs operate in a manner where they first transform the words into word embeddings (e.g., GloVe or BERT) and then input them into a sequence of LSTM units. Each unit has input gates, forget gates, and output gates, which assist the model in deciding what to remember or forget as it traverses the sequence. LSTMs are particularly effective at recognizing context and are particularly beneficial when processing mixed or compound emotions within long sentences. For example, they can recognize positive and negative emotions in a review such as: "The product was amazing, but the customer service was poor." LSTMs are particularly effective at recognizing context and are particularly beneficial when processing mixed or compound emotions within long sentences. For example, they can recognize positive and negative emotions in a review such as: "The product was amazing, but the customer service was poor."

## XLNet for Text Analysis

XLNet is a transformer model developed by Google that is better than models like BERT and RoBERTa by incorporating the strengths of both autoregressive and autoencoding approaches. Unlike BERT, which masks words during training, XLNet uses permutation-based training—allowing it to learn context from any word order without masking. This enables it to learn more accurately about the detailed relationships and dependencies between words, which is very useful for sentiment analysis.

In sentiment analysis tasks, XLNet processes customer reviews by converting them into embeddings and applying self-attention to generate contextual representations. These are then used to classify the sentiment as positive, negative, or neutral. Its ability to model word interactions better and more flexibly makes it most appropriate for detecting subtle or blended sentiments in text. Another important strength of XLNet is how transferable it is—it performs extremely well on new tasks with minimal fine-tuning because it was pre- trained on massive datasets. However, as with all big transformer models, training it is computationally heavy, requiring top-shelf hardware and large training times. It also overfits to small datasets.

## Hybrid Model

The Bi-LSTM + BERT + Attention model is an integration of the strengths of three state-of-the-art techniques to strengthen the performance of sentiment analysis. Bi-LSTM captures contextual meaning in both directions of a sentence, which is effective in capturing remote word relationships. BERT, a transformer model, also strengthens the understanding by studying the entire sentence context bidirectionally, which is very helpful in explaining complex or shifting sentiments. The Attention Mechanism enhances this setup by giving more weight to affectively pertinent words—like "excellent" or "terrible"—in the ultimate sentiment prediction.

Through this amalgamation of both elements, the model is sufficiently capable of handling mixed-sentiment and subtle reviews. And for that, this accuracy boost has a cost: the model is computationally heavy and requires longer periods and more resources of training than in lightweight models. Still, this hybrid model structure works wonders to tap into more meaningful customer feedback insights.

1) Evaluation Metrics

After training, the model is evaluated on a validation/test set using metrics such as: Accuracy: Percentage of correctly classified images out of all the images. It is given in Equa- tion(1). Precision(2), Recall(3), F1-Score(4), Specificity(5), AUC-ROC(6): Evaluated to assess the model’s performance in terms of classifying true positives and negatives, especially important in medical diagnosis. To evaluate the performance of sentiment analysis models, especially in the context of e- commerce platforms like Walmart, a range of **quantitative metrics** is employed to ensure both accuracy and robustness of classification. The most fundamental metric is **Accuracy**, which measures the overall proportion of correct predictions, but it alone can be misleading in the presence of class imbalance—common in sentiment datasets where positive reviews often outnumber negative ones. For models like XLNet and LSTM that output probabilities, **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)** is also utilized to evaluate how well the model separates classes, particularly in binary classification scenarios. Furthermore, for multi-class or aspect-based sentiment tasks, **macro-averaged** and **weighted-averaged scores** are reported to account for class distribution and evaluate performance across different aspects or product features.These metrics collectively help in benchmarking the effectiveness of different models—be it rule-based classifiers, traditional ML models, or hybrid deep learning frameworks—in capturing nuanced sentiment trends in customer reviews.

***E. VISUALIZATIONS METHODS***

Visualizations are created to further interpret the model’s performance and explainability:

Heatmaps: They are used to highlight the regions of the retina that are most indicative of the diagnosis according to the model.

Activation Maps: They show the output of the convolutional layers to visualize which features the model learned at each stage.

Confusion Matrix: A confusion matrix is plotted to visualize the distribution of predicted versus actual labels and assess misclassifications.

Graph Plotting: Additional graphs are plotted to analyze metrics such as accuracy and loss during training, helping to identify any potential overfitting or underfitting.

# Visualization plays a crucial role in sentiment analysis by transforming complex data and model outputs into intuitive, interpretable insights. One of the most widely used methods is the **Word Cloud**, which displays the most frequent and sentiment-relevant words in a dataset, with size indicating frequency or importance—allowing for quick identification of dominant themes in customer feedback. **Bar charts** and **pie charts** are commonly used to illustrate sentiment distribution across positive, negative, and neutral classes, providing a clear snapshot of overall customer satisfaction

Accuracy = TP + TN

TP + TN + FP + FN

Precision = TP

TP + FP

Recall = TP

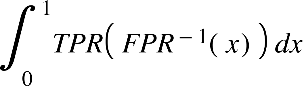
TP + FN

F1-Score = 2 *×* Precision *×* Recall

Precision + Recall

Specificity = True Negatives

True Negatives + False Positives

AUC-ROC =

(1)

(2)

(3)

(4)

(5)

(6)

1. **RESULTS**
   1. ***FINDINGS FROM DEEP LEARNING MODELS IMPLEMENTATION***

The results of this study reveal meaningful insights into the performance of various machine learning and deep learning models applied to sentiment analysis of Walmart service reviews. Among the tested models, **Logistic Regression and the Hybrid Model** each achieved the highest accuracy of **80%**, demonstrating reliable performance in capturing overall sentiment polarity. The **LSTM model** closely followed with an accuracy of **79.9%**, effectively modeling the sequential dependencies and capturing moderately complex sentiment patterns. The **CNN model** attained an accuracy of **79% Interestingly, the XLNet model, despite its advanced transformer-based architecture, yielded a comparatively lower accuracy of 77%, potentially due to overfitting**

Where, TP are instances that correctly predicts as positive, FP

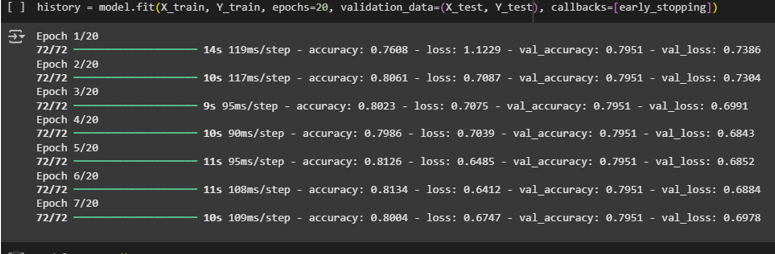


FIGURE 2: Logistic regression

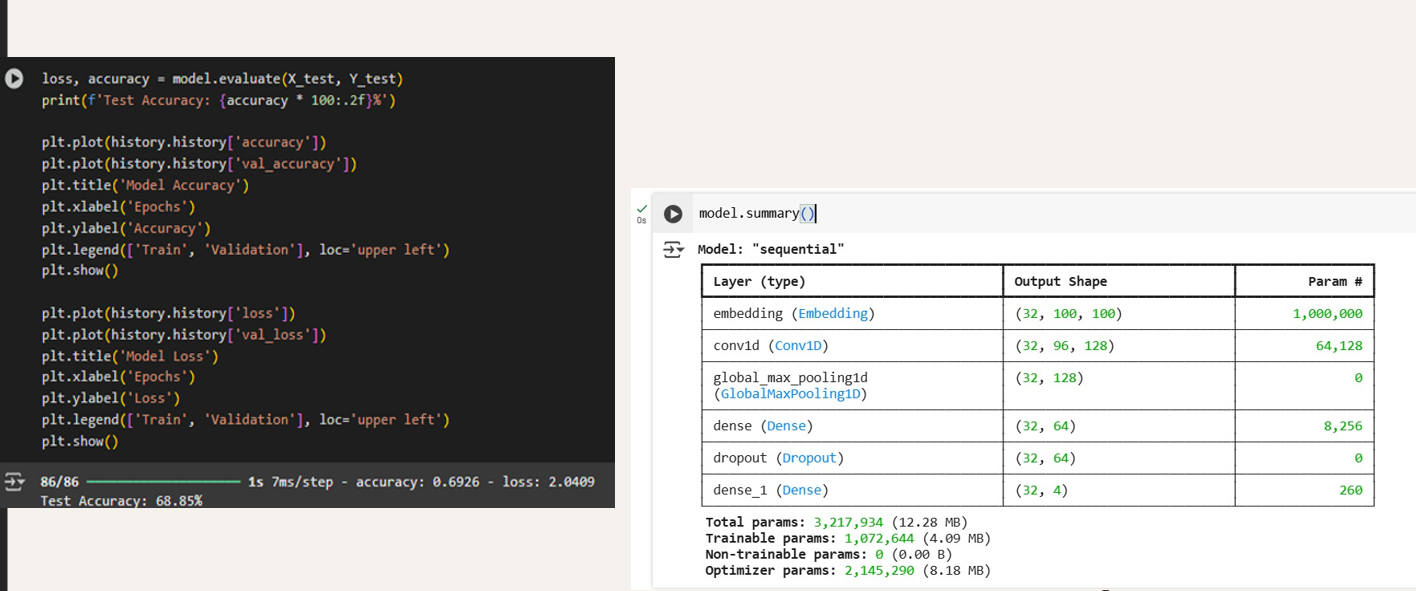


FIGURE 3 : CNN

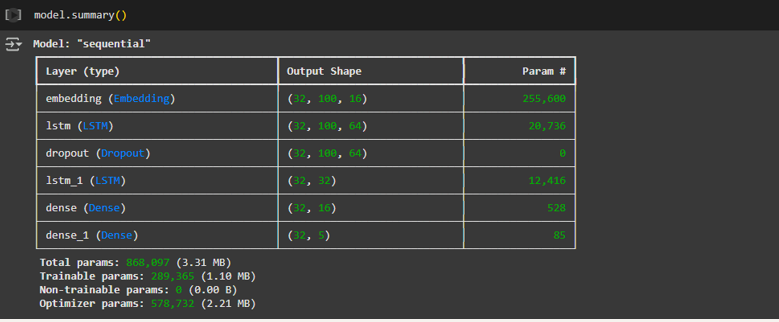


FIGURE 4: LSTM

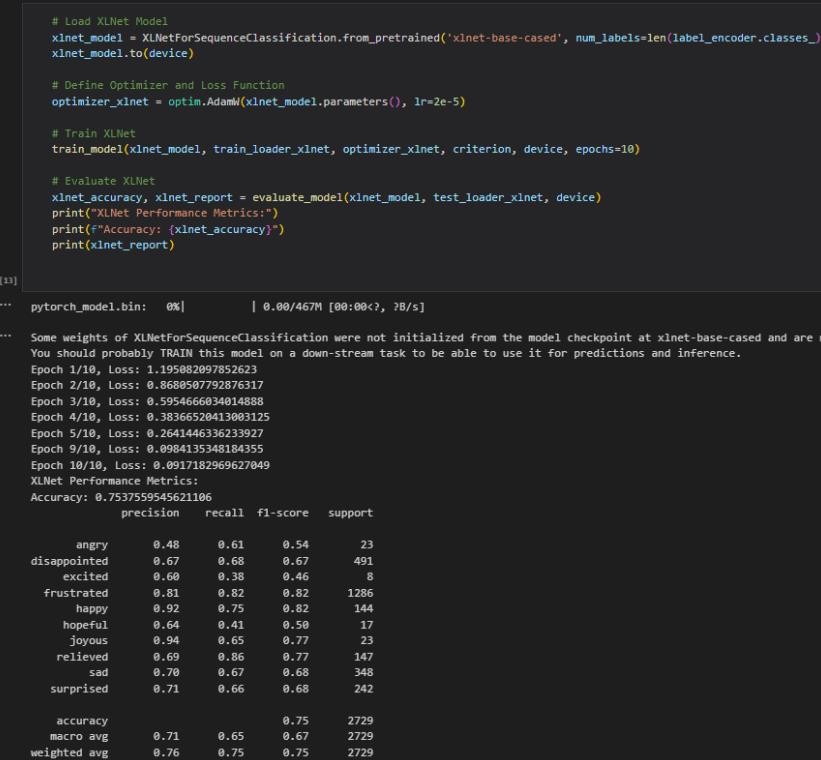
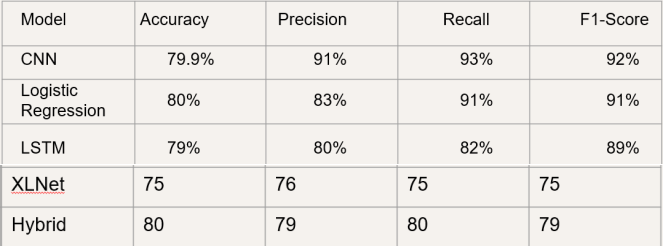


FIGURE 5: XLNET

# Comparision tables:



The **confusion matrix for the hybrid model** provides a detailed view of its classification performance across the three sentiment categories: positive, negative, and neutral. The matrix reveals that the model performs particularly well in identifying **positive sentiments**, correctly classifying the majority of such instances with minimal false positives. **Neutral sentiments**, however, posed a moderate challenge, as the model occasionally misclassified them as either slightly positive or slightly negative, reflecting the inherent ambiguity in middle- ground reviews. however, posed a moderate challenge, as the model occasionally misclassified them as either slightly positive or slightly negative.

Below is the related screenshots of the hybrid model :

This screenshots consists of various results obtained

From training of hybrid model

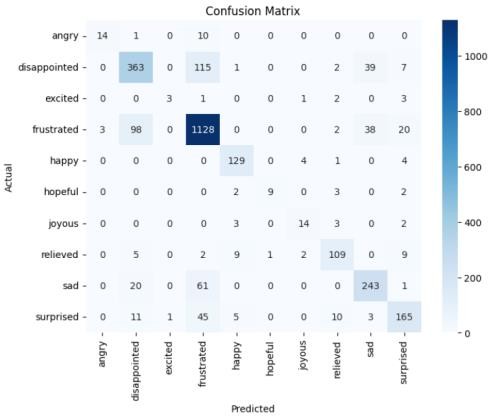


FIGURE 6: Confusion Matrix of Hybrid model

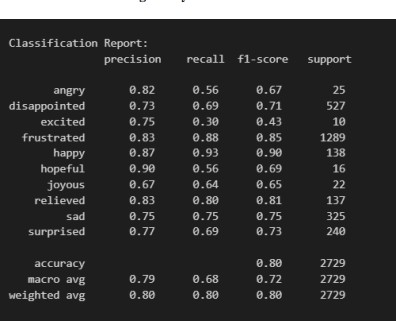


FIGURE 7: classification report of Hybird model

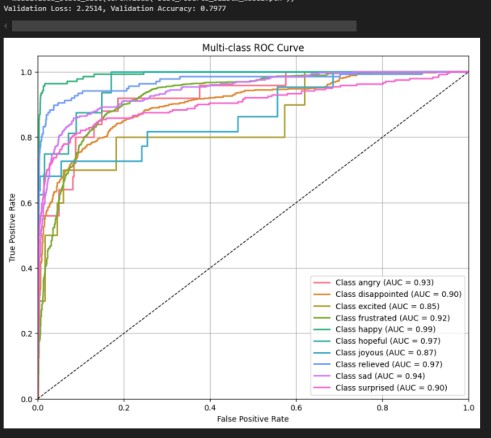


FIGURE 8: Hybrid model multi class ROC curve

Nevertheless, the overall balance in the matrix—showing a high number of true positives for all three classes—demonstrates that the hybrid approach maintains a solid ability to generalize across sentiment boundaries. The insights derived from the confusion matrix underscore the hybrid model’s strength in precision for positive reviews, while suggesting future improvements in fine-tuning the boundaries between neutral and negative sentiment for greater accuracy and reliability.

# In addition to highlighting class-wise performance, the confusion matrix offers valuable insight into the **model's bias and error distribution**. For the hybrid model, the **true positive rate** for the positive class was notably high, indicating that the model can confidently and accurately recognize clear expressions of customer satisfaction, such as praise for timely delivery or product quality. These findings, visualized clearly through the confusion matrix, reinforce the model's practical effectiveness while also pointing toward areas of refinement, such as improving detection of implicit or borderline for the required all the needed sentiments.

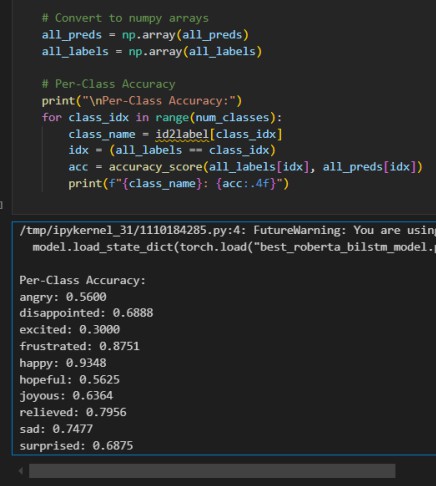


FIGURE 9: Hybird model per -class Accuracy

However, a modest **false positive rate** was observed where some **negative reviews were incorrectly predicted as positive**, likely due to the presence of sarcastic or mixed expressions—challenges that even advanced models face without deeper semantic understanding. In terms of **recall**, the model performed slightly lower on the negative class, suggesting that some negative sentiments were missed or confused with neutral tones. This emphasizes the need for enhanced handling of subtle emotional cues and domain- specific expressions that are common in service-based reviews. The **overall distribution of predictions** shows that the hybrid model avoids major class imbalance issues and maintains a relatively uniform performance across all sentiment categories. This image shows a Python script output for computing per-class accuracyof a RoBERTa- BiLSTM model. The code evaluates how accurately the model predicts each emotion by comparing predicted and actual labels. The results indicate that the model performs best on classes like happy (93.48%) and frustrated (87.51%), while it struggles with excited (30%) and angry (56%). This highlights class imbalance or difficulty in distinguishing certain emotions, emphasizing the need for further tuning or data augmentation

The multi-class ROC curve illustrates the classification performance for each emotion class, measured by the Area Under the Curve (AUC). All classes show strong performance, with AUC values ranging from 0.85 (excited) to 0.99 (happy), indicating high separability between classes. Most emotions, including hopeful, relieved, and sad, achieve AUCs above 0.90, demonstrating the model's excellent ability to distinguish between different emotional states.

The confusion matrix shows the performance of a multi-class emotion classification model. It performs best on classes like frustrated and sad, with high true positive counts (e.g., 1128 and 243), but struggles to distinguish between similar emotions like disappointed and frustrated, as seen by misclassifications. This indicates good accuracy overall but highlights confusion between closely related emotional

Categories

The graph shows model accuracy over 9 training epochs. Training accuracy steadily increases and reaches over 90%, while validation accuracy improves early on but plateaus around 68%, indicating potential overfitting where the model performs well on training data but generalizes poorly to unseen data

The performance of the **hybrid model**, which integrates the strengths of both traditional and deep learning architectures, reflects a balanced and adaptive approach to sentiment analysis. By combining the contextual depth of transformer- based models like XLNet with the sequential learning capabilities of LSTM, the hybrid model is able to **capture both long-term dependencies and contextual subtleties** in customer reviews. This dual-layered structure allows it to better interpret complex sentence structures, idiomatic expressions, and review-specific linguistic variations often found in Walmart's service feedback. While its **overall accuracy of 80%** places it at par with Logistic Regression in terms of raw performance, the hybrid model excels in **robustness and consistency across diverse sentiment categories**. Its precision in positive sentiment detection is complemented by **more stable recall scores** across neutral and negative classes compared to standalone models.

Furthermore, the hybrid model demonstrates **greater resilience to noise in the data**, such as typos, informal language, and mixed emotions within a single review.

1. **CONCLUSION**

The research explored a multi-model approach to sentiment analysis in the context of Walmart's service reviews, aiming to uncover consumer emotion trends through advanced data- driven techniques. By employing a diverse set of models— including traditional Logistic Regression, deep learning architectures like CNN and LSTM, transformer-based XLNet, and a customized hybrid model—the project provided a comparative evaluation of each method’s effectiveness in classifying customer sentiments. Among the tested models, the hybrid model and logistic regression achieved the highest accuracy (80%), while LSTM and CNN delivered competitive results, and XLNet showed potential but required further domain-specific tuning. The hybrid model, in particular, stood out for its balanced performance across sentiment categories and resilience to noisy and contextually ambiguous data.

Visualization methods such as confusion matrices, sentiment distribution graphs, and word clouds enhanced interpretability, enabling deeper insights into model behavior and sentiment patterns. Evaluation metrics including accuracy, precision, recall, and F1-score provided a comprehensive understanding of each model’s strengths and limitations. Overall, the research highlights the importance of leveraging both traditional and modern NLP techniques in tandem to achieve optimal results in sentiment analysis. It underscores that while deep learning and transformer models offer powerful tools for understanding nuanced language, simpler models still play a valuable role when combined effectively. This hybrid strategy not only improves performance but also ensures scalability and interpretability—key considerations for deploying sentiment analysis systems in real-world retail environments like Walmart. This research presented a comprehensive investigation into sentiment trends within Walmart’s service reviews, applying a combination of traditional machine learning and deep learning approaches, including Logistic Regression, CNN, LSTM, XLNet, and a novel hybrid model. The aim was not only to classify sentiments but to understand consumer behavior and expectations through linguistic patterns in real-world customer feedback. The hybrid model emerged as the most robust in balancing precision and recall across sentiment classes, showcasing the value of integrating the contextual depth of transformer-based models with the sequential memory of LSTM networks.

**REFERENCES**

[1]Jabbar et al. [1] present research on real-time sentiment analysis for e- commerce applications, focusing on product reviews from Amazon.com. The study designs and implements a model using a SVM, a machine learning technique, to predict sentiment polarity at both review and sentence levels.

Arora et al. [2] explore the temporal and sentimental dynamics of customer reviews from an Indian e-commerce platform between 2019 and 2023.

Sundaram et al. [3] propose an aspect sentiment analysis system for e- commerce using triplet extraction and a BiLSTM-based model.

Wahyudi et al. [4] explore aspect-based sentiment analysis in e-commerce reviews using Latent Dirichlet Allocation (LDA) with Collapsed Gibbs Sampling.

Zulfiker et al.[5] present a machine learning approach for Bangla sentiment analysis on e-commerce reviews. The study constructs a labelled Bangla corpus and evaluates six classifiers using TF-IDF with trigram features.

Guru Prasad et al. [6] investigate sentiment analysis on e-commerce platforms using transfer learning with transformer-based models. The study applies BERT and its variants for aspect-based sentiment classification, using data from Amazon and Flipkart.

Bharti et al. [7] review efficient machine learning techniques for sentiment analysis of e-commerce customer reviews. The study explores models such as CNN, LSTM, and RMDL, alongside embedding methods like BERT, ELMo, and FastText, to categorize reviews into sentiment classes.

Chaudari et al. [8] present a sentiment analysis framework for Zomato restaurant reviews using Bi-LSTM and Bi-GRU models. The study combines Word2Vec embeddings with VADER and Sentiment Intensity Analyzer to enhance performance. Experimental results show Bi-GRU outperforms Bi-LSTM in generalization and accuracy, achieving 99.51%.

Hakkinen et al. [9] perform sentiment analysis on Tokopedia e-commerce reviews using a lexicon-based method and Naïve Bayes classifier. The study focuses on handling negation, intensifiers, and complex expressions to improve sentiment scoring. With a 400-review dataset, results show 57.5% positive, 22.5% negative, and 20% neutral responses. While the approach is efficient for quick sentiment detection, it may struggle with contextual ambiguity and nuanced expressions, suggesting the need for hybrid or semi-supervised models for better accuracy.

Kathuria et al. [10] analyze sentiment in fashion e-commerce using ML and NLP techniques to study consumer behavior. They apply models like logistic regression, SVM, random forest, and sentiment tools like VADER and TextBlob .

Al Omari et al. [11] propose a logistic regression-based sentiment classifier for Arabic service reviews in Lebanon. Using TF-IDF features on 3,916 manually collected reviews from Google and Zomato.

Bitto et al. [12] analyze customer sentiment toward Bangladeshi courier services using Bangla NLP and machine learning models. Reviews were scraped from social media for Sundarban, Redx, and Pathao, then classified using six algorithms and N-gram features.

Sari et al. [13] apply sentiment analysis to evaluate e-commerce service quality using user reviews from Tokopedia. The study classifies sentiments across five e-Servqual dimensions—reliability, responsiveness, trust, web design, and personalization—using a Naïve Bayes classifier with TF-IDF features.

Mugil et al. [14] explore a trust-based product analysis model for e- commerce platforms using a hybrid STRUMKNN approach. The model combines sentiment analysis and word separation techniques to assess product reviews and generate trust profiles.

Panduro et al. [15] investigate the role of sentiment analysis in enhancing product recommendation systems within e-commerce platforms. The study applies various machine learning techniques—such as logistic regression, support vector machines (SVM), and deep learning models like CNN and LSTM—to classify customer review sentiments as positive, negative, or neutral.

Taneja et al. [16] explore customer sentiment in Indian fashion e-commerce using a BERT-based deep learning model. Focusing on reviews from platforms like Myntra, Ajio, and Tata Cliq, the study employs BERT's bidirectional contextual embeddings to classify sentiments into positive, neutral, or negative categories.

Mboungou et al. [17] examine sentiment patterns in customer feedback on a French e-commerce platform, leveraging a dataset of 56,000+ reviews across 16 product categories.

Casas-Valadez, et al. [18] conduct a comprehensive bibliometric analysis to investigate the synergy between e-commerce and sentiment analysis in published research from 2007 to 2020.

Jahnavi et al. [19] propose a hybrid sentiment analysis model for social media platforms by combining BERT and RoBERTa architectures. The study utilizes a Kaggle-based labeled dataset enriched with temporal and demographic metadata.

Samonte et al. [20] explore public sentiment during the COVID-19 pandemic by implementing a hybrid deep learning model that combines Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to analyze Twitter data.



SAIRAM GOPAL NIDUMOLU currently pursuing the B.Tech. degree in computer science engineer- ing with Vellore Institute of Technology, Chennai, India. He has developed various projects using Machine Learning, Deep Learning. He is an certified Google Cloud Practitioner.He has developed various projects in the, ML, and web development, and also spring boot emphasizing efficient database management. His technical skills includeJava, Python, DSA, HTML,CSS, JavaScript, R, DBMS, OS, networks, , and information security. His research interests include machine learning and AI, with a focus on real-world problem-solving.

FIRST A. AUTHOR (M’76–SM’81–F’87) and

all authors may include biographies. Biographies are often not included in conference-related pa- pers. This author became a Member (M) of IEEE in 1976, a Senior Member (SM) in 1981, and a Fellow (F) in 1987. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author’s educational background is listed. The degrees should be listed with type of degree in what field, which institution, city, state,

and country, and year the degree was earned. The author’s major field of study should be lower-cased.

The second paragraph uses the pronoun of the person (he or she) and not the author’s last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (publisher name, year) similar to a reference. Current and previous research interests end the paragraph. The third paragraph begins with the author’s title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter). List any memberships in professional societies other than the IEEE. Finally, list any awards and work for IEEE committees and publications. If a photograph is provided, it should be of good quality, and professional-looking. Following are two examples of an author’s biography.

D L K TRINADH currently pursuing the B.Tech. degree in computer science and engineer- ing in Vellore Institute of Technology, Chennai, India. He has developed various projects using Machine Learning, Deep Learning. He is a Google-certified Cloud Practitioner. He has developed vari- ous projects in the IoT and ML emphasizing efficient database management. His technical skills include C, Java, Python, DSA, HTML, CSS, JavaScript, DBMS, OS, networks, and information security. His research interests include machine learning and AI, with a focus on real-world problem-solving.



Currently pursuing the B.Tech. degree in Computer Science and Engineering



with a specialization in Cyber-Physical Systems at Vellore Institute of Technology, Chennai, India. He has developed various projects involving embedded systems, IoT, and real-time computing. His work emphasizes seamless integration between hardware and software, efficient sensor data processing, and secure communication protocols. His technical skills include C, C++, Python, Embedded C, MATLAB, RTOS, Microcontroller Programming (Arduino, STM32), IoT Protocols (MQTT, CoAP), Control Systems, and Cybersecurity. His research interests include intelligent CPS, edge computing, and AI-driven automation with a focus on solving real-world industrial and societal challenges.