



PREDICTING CUSTOMER SATISFACTION THROUGH LAYERED NLP TECHNIQUES: TRADITIONAL MEETS DEEP LEARNING

Mohammed Bachir Mahdjoub^{*1}, Fatima Zohra Laallam² and Messaoud Mezati³

¹Mohamed Khider University, Algeria

^{2,3}Kasdi Merbah University, Algeria

¹<http://orcid.org/0000-0001-9792-1831>, ²<http://orcid.org/0000-0003-2875-3576>, ³<http://orcid.org/0009-0001-1996-5625>

Email: *bachir.mahdjoub@univ-biskra.dz, laallam.fatima_zohra@univ-ouargla.dz, mezati.messaoud@univ-ouargla.dz

ARTICLE INFO

Article History

Received: October 25, 2025

Revised: November 20, 2025

Accepted: January 1, 2026

Published: January 31, 2026

Keywords:

Customer Satisfaction Prediction,
Natural Language Processing,
Deep Learning,
Machine Learning,
Sentiment Classification,
CRM.

ABSTRACT

This study presents a robust hybrid artificial intelligence framework designed to predict customer satisfaction based on online reviews a critical capability for enhancing business performance and optimizing customer relationship management (CRM) practices. We propose a unified architecture with high precision for sentiment classification, utilizing the "All Beauty (2023)" dataset from Amazon Reviews. The approach synergistically combines conventional machine learning (ML) algorithms with state-of-the-art deep learning (DL) techniques. Although standalone ML and DL models have demonstrated promising outcomes in sentiment analysis, they often fall short in addressing intrinsic data challenges such as class imbalance and linguistic subtleties. To overcome these limitations, we introduce HybridStack, a novel ensemble architecture that integrates diverse ML and DL models to significantly improve classification accuracy. The methodology encompasses rigorous text preprocessing, a dual-path feature extraction pipeline (TF-IDF for ML models, and embedding techniques for DL models), and a meta-classifier to consolidate predictions into three categories: dissatisfied, neutral, and satisfied. To address class imbalance, we employ weighted loss functions alongside oversampling strategies. The proposed HybridStack model achieved an outstanding and unparalleled F1-score of 0.6889 and an accuracy of 0.8593, substantially outperforming all baseline individual models. This hybrid approach highlights the benefits of algorithmic synergy and demonstrates its effectiveness in generating more granular and actionable insights, thereby supporting the development of more refined CRM strategies and fostering stronger customer relationships.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Automated analysis of customer satisfaction from online reviews represents a crucial challenge in data science and artificial intelligence, as it directly contributes to enhancing sales strategies and customer relationship management (CRM) effectiveness. Despite significant advances in natural language processing (NLP), machine learning (ML), and deep learning (DL), existing methods still struggle to deliver robust and generalizable performance in real-world, unbalanced datasets. Current studies mainly focus on either traditional ML models or advanced neural networks, often neglecting the complementary strengths that could arise from their integration. In this context, the present work proposes a novel hybrid AI framework that unifies classical ML algorithms (e.g., Logistic Regression, Naive Bayes) with state-of-the-art DL architectures (LSTM, GRU, CNN) and transformer-based models (BERT) within a stacked ensemble learning strategy. This hybridization enables the system named HybridStack to exploit the interpretability and efficiency of traditional classifiers alongside the representational depth of neural and transformer-based models. The proposed pipeline operates in parallel, leveraging TF-IDF features for conventional models and embedding representations for deep architectures, before aggregating their predictions through a meta-learner to enhance overall precision and robustness.

Evaluated on the “All Beauty (2023)” Amazon Reviews dataset, this framework demonstrates improved predictive accuracy and stability compared to single-model baselines. The novelty of this work lies in its systematic integration of heterogeneous learning paradigms within a unified architecture, addressing key challenges such as linguistic complexity, feature diversity, and class imbalance. This contribution advances the state of research toward more adaptive and generalizable satisfaction prediction systems in practical CRM applications.

II. THEORETICAL REFERENCE

II.1 CRM AND CUSTOMER SATISFACTION

Customer relationship management (CRM) is fundamentally a strategic business approach focused on managing interactions with current and potential customers in order to stimulate acquisition, promote loyalty and support long-term retention. Its main objective is to improve customer satisfaction through personalised engagement, a deep understanding of their expectations and the establishment of lasting relationships that promote increased customer lifetime value (Figure 1).



Figure 1: Customer Satisfaction Measurement Techniques.

Source: Authors, (2026).

CRM integrates technology, data and business strategies to systematically collect, process and analyse customer information, enabling organisations to provide more personalised and effective services. Several definitions reflect the multifaceted nature of CRM. [1] defines it as the strategic deployment of information, processes, technologies and human resources throughout the customer lifecycle, including marketing, sales, service and support. [2] presents CRM as a proactive strategy aimed at understanding and anticipating customer needs. Similarly, [3] and [4] emphasise the role of CRM in identifying profitable customer segments, implementing value-driven interactions and developing profitable strategies to drive profitability. [5] characterises CRM as a continuous process of identifying customer value and personalising interactions in order to exceed expectations. [6] further describe CRM as a fundamental business strategy that harmonises internal operations and extends organisational networks to create value through the effective use of high-quality customer data and advanced information systems. In the rapidly expanding e-commerce sector, customer satisfaction is essential for commercial success and market survival [7]. Advanced analytical tools are indispensable for effectively analysing the vast amounts of data generated by customer feedback, which directly influences the success of products and companies [8].

II.2 PREDICTING CUSTOMER SATISFACTION IN THE ERA OF MACHINE LEARNING AND DEEP LEARNING

Sentiment analysis, a key text mining technique, is widely used to identify and extract subjective information from various sources, helping businesses understand customer opinions, assess their preferences, and improve their purchasing decisions [8]. The integration of machine learning and deep learning models, such as BERT, has significantly improved the effectiveness of sentiment analysis and enabled highly accurate predictions of customer satisfaction by analysing reviews and integrating them with e-commerce data [8-11]. This data-driven insight is essential not only for improving product quality and operational efficiency, but also for enhancing the overall customer experience, optimising conversion rates, and fostering sustainable growth in the competitive e-commerce landscape [12]. The analysis of complex textual data, such as customer reviews, relies heavily on the capabilities of artificial intelligence paradigms such as machine learning (ML) and deep learning (DL). Traditional ML models, including logistic regression and the naive Bayes model, have consistently provided robust and interpretable solutions for text classification tasks, often leveraging statistical features such as TF-IDF to quantify the importance of words. At the same time, deep learning models, such as long short-term memory (LSTM), gate recurrent units (GRU), and convolutional neural networks (CNN), have revolutionised natural language processing with their ability to automatically learn complex semantic and contextual representations through integrations, often enhanced by powerful transformer-based architectures such as BERT. These two approaches offer distinct advantages for discerning patterns and sentiments within large datasets, thereby contributing significantly to the automated prediction of customer satisfaction (see Figure 2).

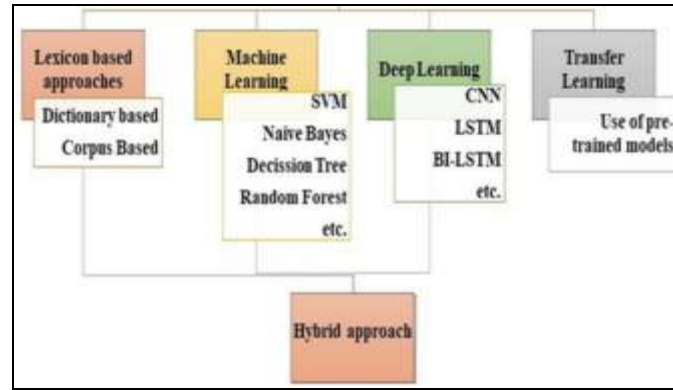


Figure 2: Different Approaches for Text Analysis in Customer Satisfaction Prediction.

Source: Authors, (2026).

II.3 STUDY EXISTING APPROACHES

Traditional Machine Learning (ML) and Deep Learning (DL) models each possess significant capabilities but also distinct limitations when applied to complex, noisy, and imbalanced textual data typical of e-commerce environments. ML models, though interpretable and efficient, often fail to capture the deep semantic and contextual dependencies present in natural language. Conversely, DL models such as LSTM, GRU, and CNN can learn intricate linguistic patterns but tend to act as “black boxes,” with reduced interpretability and a higher risk of overfitting especially when faced with underrepresented sentiment classes like dissatisfied customers. These challenges highlight the need for a hybridized analytical framework capable of exploiting the complementary strengths of both paradigms while mitigating their respective weaknesses. The proposed HybridStack approach fulfills this requirement by integrating diverse ML and DL models within an ensemble structure, where a meta-classifier synthesizes their outputs. By combining models specializing in statistical feature extraction (e.g., Logistic Regression, SVM) with those capturing deep semantic meaning (e.g., LSTM, CNN, BERT), HybridStack enhances robustness, accuracy, and generalization.

This fusion enables more reliable sentiment classification and superior detection of minority sentiment classes, directly supporting actionable insights for Customer Relationship Management (CRM) systems. The rapid expansion of e-commerce, driven by Artificial Intelligence (AI) and ML, has intensified research in predictive analytics applications such as demand forecasting, customer segmentation, and fraud detection [13-17]. Sentiment analysis and satisfaction prediction have become central to this effort, leveraging models like SVM, Naïve Bayes, BiLSTM, and BERT to extract subjective opinions from large volumes of customer reviews [8], [9], [18]. Empirical studies across platforms such as Amazon and Jumia confirm the value of these methods [10], [19]. However, standalone models still face persistent obstacles linked to linguistic variability, class imbalance, and noisy data, which hinder their generalizability. To overcome these issues, hybrid and ensemble learning approaches have gained prominence for their capacity to deliver higher predictive accuracy and resilience [12].

Research demonstrates that combining algorithms for instance, SVM with Logistic Regression [13], or multiple ensemble classifiers for churn and delivery prediction [16], [20], [21], [22] achieves superior outcomes. Advanced Natural Language Processing (NLP) techniques have also improved sentiment extraction, topic modeling, and satisfaction forecasting [23-25]. Moreover, the integration of ML with Business Intelligence (BI) and visualization tools [26] supports more informed, data-driven decision-making while ongoing developments in explainable and fairness-aware AI address transparency and bias concerns. Collectively, this growing body of work underscores the transition toward integrative and adaptive analytics frameworks in e-commerce. By uniting interpretability, statistical precision, and deep contextual learning, hybrid systems such as HybridStack represent the most accurate, scalable, and reliable approach for multi-class sentiment prediction. These architectures not only enhance predictive performance but also enable personalized customer experiences, ethical AI implementation, and sustainable business growth in the evolving digital marketplace.

II.4 SCIENTIFIC AND CONTEXTUAL RELEVANCE OF THE RESEARCH

This observation highlights the current relevance of research focused on combining machine learning and deep learning. In a commercial environment increasingly dominated by e-commerce, companies generate massive volumes of textual data every day from reviews, comments and customer interactions. Effectively exploiting this information requires models that are capable of both understanding the linguistic and emotional subtleties of texts and producing reliable, explainable and generalisable predictions. This issue is all the more crucial given that customer satisfaction is a major strategic lever for competitiveness, loyalty and value creation. Thus, the ability of artificial intelligence to accurately predict this satisfaction represents a major scientific, technological and managerial challenge.

II.5 NOVELTY AND POSITIONNING OF RESEARCH

In this perspective, the novelty of this work lies in the design of a hierarchical hybrid approach, combining traditional machine learning (ML) methods and deep learning (DL) models applied to natural language processing. The proposed approach is based on an overall strategy inspired by the HybridStack concept, which aggregates the predictive outputs of several heterogeneous models to feed a meta-classifier. This hybridisation makes it possible to exploit the complementarity between the statistical recognition of ML models (such as SVM, logistic regression, etc.) and the deep semantic understanding of DL models (such as LSTM, GRU, CNN and BERT).

The objective is to simultaneously improve the robustness, accuracy and generalisation capacity of the customer satisfaction prediction system, while maintaining better interpretability of the results. Furthermore, the integration of hierarchical language processing is an original contribution. By structuring textual analysis on several levels lexical, syntactic, semantic, and pragmatic this approach offers a more detailed view of the emotions, intentions, and judgements expressed in customer reviews. It also allows for better detection of minority classes, particularly dissatisfied customers, whose recognition is essential for CRM. Thus, this research builds on recent work on ML and NLP, while offering a distinct and innovative contribution: a hybrid and hierarchical predictive framework that reconciles interpretability, semantic depth and operational performance. This framework aims to fill a gap identified in the literature, namely the lack of integrative models combining symbolic learning and deep representation for the automated prediction of customer satisfaction in e-commerce.

Table 1: Summary of Related Works on E-commerce Customer Behavior and Prediction.

Article	Methodology/Algorithms Used	Key Findings/Contributions
I. Deep Learning Models		
[11]	RNNs, LSTM with SGD	Superior accuracy in dynamic pricing; balanced profit/satisfaction.
[27]	CNNs (VGG16, ResNet50, InceptionV3)	CNNs improve profit prediction (high accuracy).
[10]	Keras DL framework, Sequential model, Adam optimizer	Adam optimizer excels in customer satisfaction prediction.
[9]	ML: BERT model (fine-tuned)	BERT accurately forecasts customer sentiment (96%).
[22]	ML: PIVD using CNN in three steps	PIVD improves image vulnerability detection (high F1).
II. Traditional Machine Learning Models		
[14]	ML: Decision Trees, Random Forest (using TIPI test for consumer profiles)	ML predicts consumer behavior from personality.
[16]	ML: SVM, Random Forest, Decision Tree	SVM predicts churn (83.45%) and delivery.
[28]	ML: Association rules via Frequent-Pattern-Growth algorithm	High probability of next product purchase via recommendations.
[20]	ML: Seven traditional classification algorithms including Logistic Regression, Random Forest, Naïve Bayes	ML predicts online banking satisfaction/frustration.
[26]	BI tools incorporating Data mining, Data Analytic, Data Visualization and ML; ABCD framework	BI/ML aid decision-making and competitiveness.
III. Hybrid/Fusion Machine Learning Approaches		
[13]	ML Fusion Algorithm: Logistic Regression + SVM	Fused ML models offer higher prediction accuracy.
[8]	ML: SVM, Naïve Bayes, DL (bidirectional LSTM, BERT)	Sentiment analysis/forecasting enhances BI/decisions.
[29]	ML: RF, GBDT, XGBOOST; DL: RNNs, LSTM	ML/DL accurately forecast purchase behavior (similar performance).
[18]	Advanced ML techniques: DL, Reinforcement Learning, NLP	ML frameworks boost customer engagement/LTV/growth.
[10]	Multimodal Learning Framework (including unimodal comparison), Adam optimizer	Multimodal learning improves satisfaction/efficiency.
[15]	Big Data Algorithms, Text Mining Technology, Empirical analysis (SEM after correction)	Online interaction impacts satisfaction; personalized recs need ethics.

Source : Authors (2026)

III. MATERIALS AND METHODS

This section presents a hybrid AI framework designed to analyze and enhance customer satisfaction using the Amazon Reviews dataset. The system integrates Machine Learning (ML), Deep Learning (DL), and transformer-based models to deliver a robust sentiment analysis solution. Text preprocessing involves cleaning (removal of symbols, punctuation, and numbers), tokenization, and normalization, followed by numerical conversion through vectorization to prepare data for computation. The architecture adopts a two-stream design. The first stream employs TF-IDF vectorization to extract statistical text features, which are then classified using traditional ML algorithms such as Logistic Regression, Multinomial Naïve Bayes, XGBoost, and LinearSVC, chosen for their interpretability and stability as baselines. The second stream applies DL techniques leveraging word embeddings to capture semantic context. Models include LSTM, GRU, and CNN, with pre-trained GloVe embeddings and fine-tuned BERT transformers providing deep contextual understanding through attention mechanisms. Predictions from both streams are integrated through a stacked ensemble meta-classifier, forming the HybridStack. This ensemble combines the complementary strengths of ML and DL models, improving accuracy, robustness, and the classification of sentiment categories ('dissatisfied,' 'neutral,' 'satisfied'). Finally, an evaluation module assesses performance using Macro F1-Score, Accuracy, Precision, and Recall, with special focus on minority classes like 'Dissatisfied.'

Confusion matrices visualize misclassifications, and McNemar's test validates statistical significance. The architecture balances precision, interpretability, and scalability, offering a deployable solution for real-world customer feedback analytics.

III.1 DATA SOURCE AND COLLECTION

The dataset utilized in this study is the Amazon Customer Reviews Dataset, an extensive collection publicly available via the Amazon Web Services (AWS) Open Data Registry. This corpus encompasses a vast array of customer reviews, spanning multiple international marketplaces and diverse product categories. Each review is enriched with valuable metadata, including a unique product identifier (ASIN), reviewer ID, review title and body, a numerical star rating (1 to 5), and the timestamp of submission. For this research, we specifically focused on a curated subset: the Amazon Reviews All Beauty (2023) dataset. This dataset, stored in file "All_Beauty.jsonl", contains customer reviews specifically within the "All Beauty" product domain. This particular domain was selected to ensure a consistent context for product satisfaction, allowing for a focused analysis of beauty-related reviews. The All Beauty category was chosen to facilitate:

- **Domain-Specific Textual Expression:** Allowing for the capture of unique vocabulary and review writing styles prevalent within the beauty product sector.
- **Targeted Customer Expectations:** Providing insights into satisfaction drivers specific to beauty products, which can differ significantly from other categories.
- **Specialized Model Development:** Enabling the training of models highly optimized for predicting satisfaction in this specific retail segment.

The "All_Beauty.jsonl" dataset comprises reviews from the specified category, providing a rich source of information for modeling customer satisfaction. The star rating, from 1 to 5, serves as the primary indicator for satisfaction, allowing for both binary (satisfied vs. not satisfied, typically based on ratings of 4 or 5) and potentially multi-class satisfaction modeling.

```
rating 5
title "Great combo pack. Wish I had been using this years ago!"
text "I love this combo package, particularly the flosser. This makes a great gift package for someone you love, giving them everything they need to keep their mouth/gum health optimal. I have had issues (particularly gums) for the past 5 years. No ones fault but my own for not taking care of my mouth the way I should have all these years. This flosser is great. I fill it with Act Restoring mouthwash and use it after brushing. It amazes me how much stuff this flosser removes with a simple puff of air. Of course, you can purchase both these items separately, but you save in getting them together. So nicare makes the best tooth/gum care products on the market bar none!"
images [] 0 items
asin "B01M7UMAUG"
parent_asin "B01M7UMAUG"
user_id "AF5KPY37N3C43SQI5IEK5JSIYA"
timestamp 1508770624887
helpful_vote 0
verified_purchase false
```

Figure 3: Sample Customer Review from the Amazon Reviews All Beauty Dataset.

Source: Authors, (2026).

The Amazon Reviews dataset, and specifically the "All Beauty" subset, stands as one of the largest and most detailed open-access corpora for customer feedback within a specialized domain. It offers a rich combination of textual, categorical, temporal, and numerical features. Its focused coverage on a single product domain, coupled with detailed review content and structured metadata, makes it ideally suited for natural language processing (NLP) tasks such as sentiment classification and user profiling tailored to the beauty industry. Furthermore, the dataset reflects real-world challenges such as class imbalance and noise inherent in user-generated content, making it valuable for simulating realistic AI deployments. Its scale and granularity also enable the training of a wide range of models, from traditional machine learning algorithms to sophisticated neural architectures, including the fine-tuning of transformer-based language models like BERT and GPT for domain-specific applications.

III.2 FEATURE SELECTION

To analyze customer satisfaction in the beauty product domain, several essential features were extracted from the Amazon Reviews – All Beauty (2023) dataset. The text field contains full customer reviews and serves as the primary input for all NLP tasks such as sentiment classification and satisfaction prediction, capturing nuanced emotional expressions and user experiences. The rating feature, ranging from 1 to 5 stars, provides the ground-truth labels for supervised learning, enabling both binary and multi-class sentiment categorization. The title field offers concise summaries of opinions and acts as auxiliary textual input, improving model accuracy and supporting summarization or response generation. The timestamp attribute records the review submission time, allowing for the analysis of temporal satisfaction trends, product lifecycle patterns, and seasonal variations within the beauty sector. The asin (Amazon Standard Identification Number) uniquely identifies each product, enabling product-level sentiment analysis, satisfaction comparison, and recommendation insights. Together, these features form a multi-view dataset combining textual and structured attributes. This integration enhances semantic understanding, contextual awareness, and trend-based prediction, allowing models to learn not only what customers express but also when and about which products crucial for accurate and actionable customer satisfaction modeling in the beauty domain.

III.3 PREPROCESSING

The feature engineering and preprocessing strategy for the **Amazon Reviews – All Beauty (2023)** dataset was carefully adapted to ensure compatibility with both traditional ML and modern DL sentiment analysis models. The dataset structure was standardized by renaming key fields to align with the beauty domain: *review_body* → *text*, *star_rating* → *rating*, *review_title* → *title*, *review_date* → *timestamp*, and adding *asin* as a unique product identifier. These modifications preserved the dataset's integrity while ensuring semantic alignment for satisfaction modeling.

Text Normalization: Comprehensive normalization procedures were applied to the *text* and *title* fields to ensure uniformity and minimize lexical variability. All text was converted to lowercase, extraneous whitespace was trimmed, and punctuation was removed to reduce noise. Repetitive character sequences (e.g., “soooo”) were standardized to maintain consistency and control token distribution.

Noise Filtering: Regular expressions were used to remove HTML tags, escape sequences (e.g., `\n`, `\xa0`), and non-alphanumeric symbols that could distort tokenization or interfere with semantic learning. This step ensured clean, well-structured text inputs by eliminating web-scraping artifacts common in Amazon datasets.

Stopword Removal: Stopwords were removed using the NLTK English stopwords list to reduce feature dimensionality and enhance discriminative vocabulary for classical ML models (e.g., TF-IDF). However, this step was intentionally skipped for deep learning and transformer architectures to preserve syntactic and contextual relationships critical for attention-based modeling.

Tokenization Strategy: Tokenization was customized to match model requirements. For LSTM-based models, word-level tokenization using Keras Tokenizer produced integer sequences compatible with GloVe embeddings, preserving sequence semantics. For transformer models (e.g., BERT), subword tokenization via Hugging Face AutoTokenizer was implemented, maintaining alignment with the model's pretraining scheme (*bert-base-uncased*). This approach ensures robust handling of rare and out-of-vocabulary words.

Label Encoding: The raw 1–5 rating scale was consolidated into a three-class satisfaction scheme to reduce label sparsity and support balanced multi-class sentiment classification—typically mapping low ratings to dissatisfied, mid-range to neutral, and high ratings to satisfied. Collectively, these preprocessing and encoding strategies create a **standardized, noise-free, and semantically enriched dataset**, enabling consistent input for diverse ML/DL models. The resulting framework supports accurate, interpretable, and scalable **customer satisfaction prediction** within the beauty domain.

Table 2: Label Encoding Scheme for Customer Satisfaction Ratings.

Raw Rating	Encoded Class	Satisfaction Level
1–2	0	Dissatisfied
3	1	Neutral
4–5	2	Satisfied

Source : Authors (2026)

This transformation merges adjacent rating levels with similar semantic meaning, reflecting typical business analytics categorizations of customer satisfaction. It simplifies model training and evaluation while preserving important distinctions among dissatisfaction, neutrality, and satisfaction.

III.4 DATA SPLITTING AND CLASS BALANCE

The dataset was divided into training (70%), validation (15%), and test (15%) subsets using stratified sampling to preserve the original class distribution. This ensured that each subset accurately reflected real-world satisfaction proportions, minimizing sampling bias and supporting reliable model generalization. An exploratory class distribution analysis revealed a strong imbalance, with most reviews concentrated in 4- and 5-star categories, while 1-, 2-, and 3-star reviews were scarce. This skew mirrors real e-commerce behavior, where positive feedback dominates and negative reviews—though less frequent contain vital insights for customer service improvement. An exploratory analysis of the satisfaction labels (*star_rating* mapped to classes) showed a markedly imbalanced distribution(figure 4):

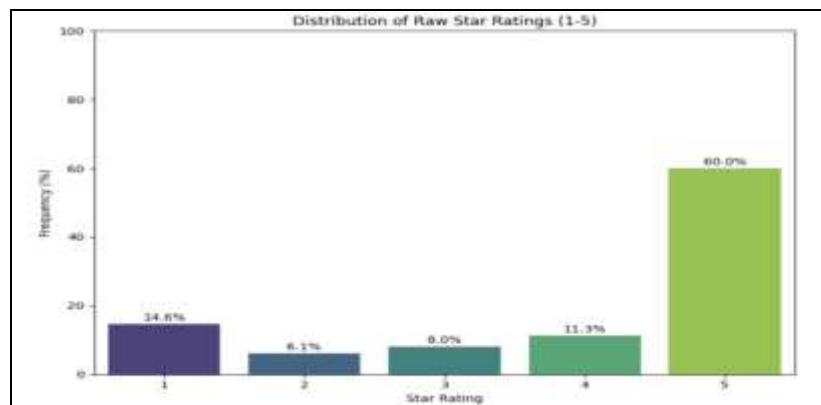


Figure 4: Distribution of Customer Satisfaction Levels.

Source: Authors, (2026).

To mitigate class imbalance, two complementary strategies were implemented. Weighted loss functions assigned higher penalties to minority classes, improving recall for “Dissatisfied” reviews and preventing overfitting to majority classes. This was achieved through the `class_weight` parameter in TensorFlow/Keras and customized loss weighting in PyTorch. Additionally, oversampling techniques such as Random Oversampling and SMOTE were selectively applied to the training data, generating synthetic minority samples to balance class frequencies while preventing data leakage. Combined with stratified batch sampling, these methods improved model robustness and minority-class detection. These measures enhanced the framework’s sensitivity to dissatisfied customers, ensuring that negative feedback signals were effectively captured. By correcting bias toward the “Satisfied” class, the final models achieved a more realistic, balanced, and business-relevant performance, crucial for actionable customer satisfaction prediction in e-commerce (figure 5).

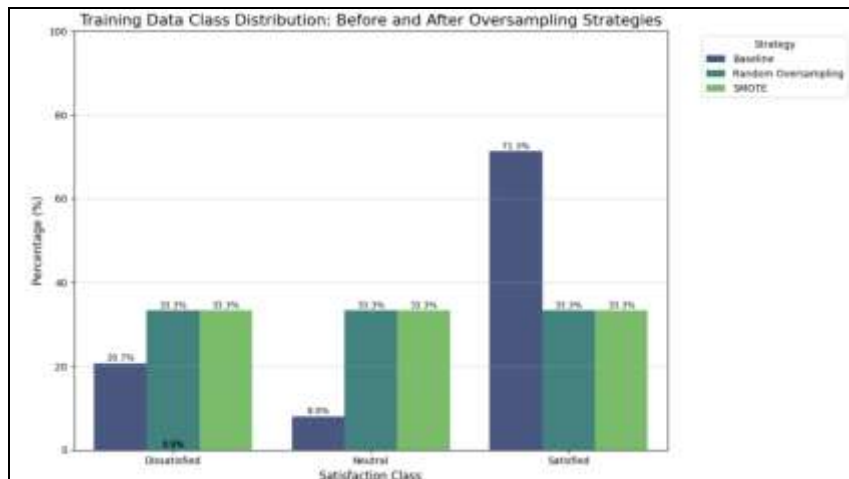


Figure 5: Comparison of Customer Satisfaction Class Distribution with Imbalance Mitigation Strategies.
Source: Authors, (2026).

III.5 PROPOSED HYBRID CLASSIFICATION ARCHITECTURE

This study employs a robust hybrid modeling approach, combining traditional machine learning (ML) and advanced deep learning (DL) architectures to classify customer sentiment into three classes: Dissatisfied (0), Neutral (1), and Satisfied (2). All models are trained using a dual-pathway feature representation strategy (Figure 6), processing distinct data modalities to capture comprehensive insights from customer reviews.

III.5.1 Ensemble Learning

Ensemble learning has emerged as one of the most reliable and high-performing paradigms in machine learning, grounded in the principle of combining multiple base learners to construct a more accurate and generalized predictive model [30]. By aggregating the outputs of diverse models, ensemble learning effectively reduces variance, bias, and generalization errors, thereby enhancing robustness and stability, particularly when dealing with complex and noisy data typical of e-commerce environments [31]. The main ensemble strategies bagging, boosting, and stacking offer complementary mechanisms for improving model accuracy. Bagging, introduced by Breiman [32], trains several independent models on bootstrap samples of the dataset, with Random Forest being its most prominent implementation [33]. Boosting methods, such as AdaBoost, Gradient Boosting, and XGBoost, build models sequentially, where each learner focuses on correcting the errors of its predecessors, resulting in strong classifiers well-suited to high-dimensional and imbalanced data [34].

Stacking, proposed by Wolpert [35], integrates heterogeneous models (e.g., decision trees, logistic regression, neural networks) through a meta-learner that optimally combines their predictions. In the context of customer satisfaction prediction in e-commerce, ensemble approaches particularly boosting-based algorithms like XGBoost and LightGBM have demonstrated superior ability to capture non-linear relationships within massive, heterogeneous, and noisy customer datasets. They enable the identification of key satisfaction drivers, enhance service personalization, and support predictive decision-making that strengthens customer retention. Therefore, ensemble learning represents a robust and adaptable framework that aligns statistical rigor with the dynamic analytical demands of modern digital marketplaces. Figure 6 conceptually illustrates the two primary feature extraction pathways and their subsequent integration into various model types:

- Pathway 1: TF-IDF Vectorization for Textual Data (for traditional ML models).
- Pathway 2: Embedding-Based Sequences for Textual Data (for deep learning models), combined with processed Categorical (One-Hot Encoded) and Numerical (Scaled) features.
- Integration Layer: A concatenation of processed features or model predictions for hybrid/ensemble approaches.

III.5.2 Feature Engineering and Data Pathways

Customer review data, comprising text (review body), categorical attributes (verified_purchase, asin), and numerical features (helpful_vote), undergoes tailored preprocessing to prepare dual feature representations:

a) **Text Preprocessing:**

- **TF-IDF Vectors:** The review_body is transformed into TF-IDF vectors (with max_features=10000 and English stop words removed) for classical machine learning models.
- **Tokenized Sequences:** For DL models, review_body undergoes tokenization (vocab_size=20000, maxlen=200), resulting in padded integer sequences.

b) **Categorical Preprocessing:** Categorical features (verified_purchase, asin) are initially label-encoded and then one-hot encoded for traditional ML, and passed as integer inputs for embedding layers in DL models.

c) **Numerical Preprocessing:** Numerical features (helpful_vote) are standardized (z-score normalization) for both ML and DL pathways.

III.5.3 Baseline Machine Learning Models (TF-IDF Pathway)

Traditional machine learning algorithms serve as performance benchmarks, primarily leveraging the TF-IDF vectorization of textual data combined with other preprocessed features. These models validate the discriminative power of the engineered features and provide reference points for deep model comparisons.

a) **Logistic Regression:** A robust linear classifier optimized for multiclass prediction. Its efficiency and interpretability provide transparent insights into feature importance.

b) **XGBoost (Extreme Gradient Boosting):** A powerful gradient boosting framework known for its high performance and flexibility. It effectively handles complex interactions between features.

c) **LinearSVC (Linear Support Vector Classifier):** A linear classifier based on Support Vector Machines, suitable for large datasets and high-dimensional sparse features like TF-IDF.

d) **Multinomial Naive Bayes:** A probabilistic classifier well-suited for discrete features like word counts or TF-IDF values, often performing well on text classification tasks.

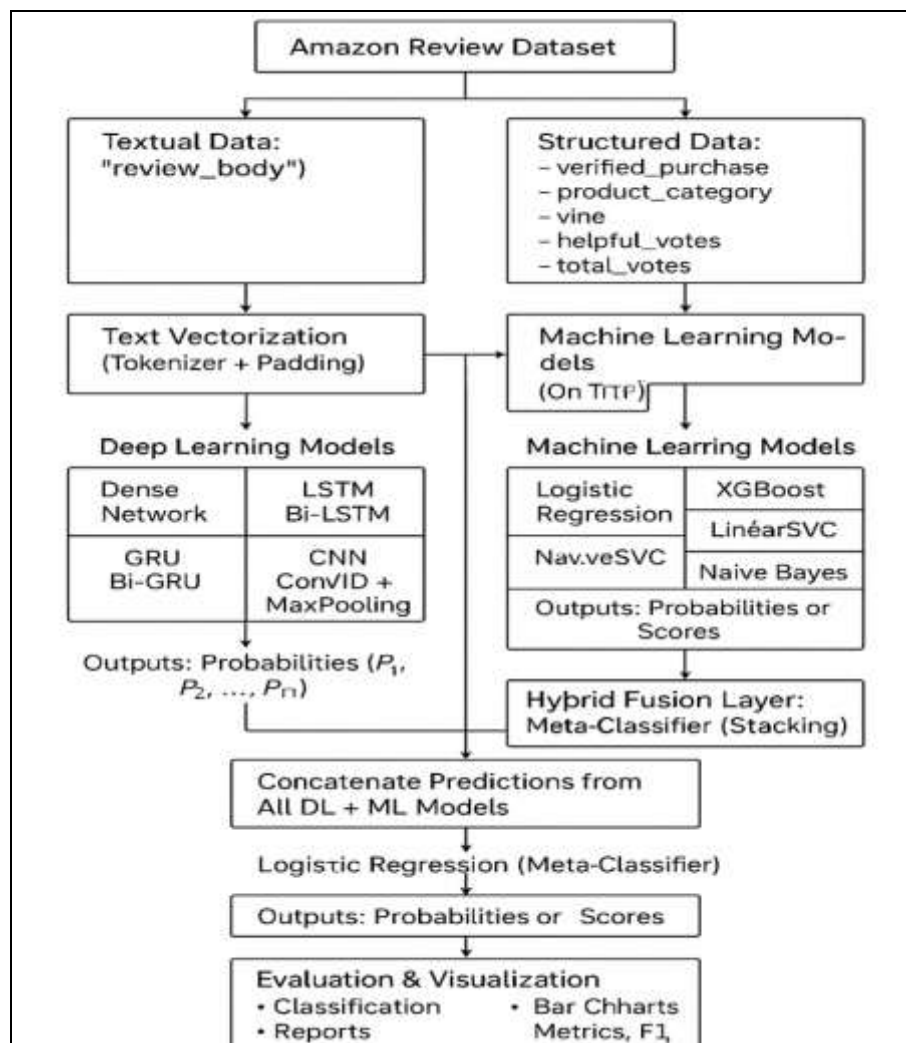


Figure 6: Hybrid Model Architecture Overview.

Source: Authors, (2026).

III.5.4 Deep Learning Models (Embedding Pathway).

To capture deeper semantic relationships and contextual dependencies within customer reviews, we implemented sequential and contextual deep learning architectures. These models process embedding-based sequences from text, combined with embedded categorical features and scaled numerical features, enabling a more nuanced representation of the overall review sentiment.

- a) **Generic Dense Network (GlobalAveragePooling1D):** A foundational deep learning model utilizing an embedding layer for text, followed by GlobalAveragePooling1D to summarize textual features, then concatenated with embedded categorical and numerical features before passing through dense layers for classification.
- b) **LSTM (Long Short-Term Memory) Network:** A recurrent neural network (RNN) variant capable of capturing long-range dependencies in sequential data. A Bidirectional LSTM layer (64 units per direction) is employed to process text embeddings, allowing the model to learn context from both past and future words in a sequence.
- c) **GRU (Gated Recurrent Unit) Network:** Another RNN variant, simpler than LSTM but similarly effective at managing vanishing/exploding gradients and learning sequential patterns. A Bidirectional GRU layer (64 units per direction) is used for text feature extraction.
- d) **CNN (Convolutional Neural Network) for Text:** A 1D Convolutional Neural Network (128 filters, kernel size 5) followed by GlobalMaxPooling1D is used to capture local patterns and hierarchical features from text embeddings, providing a robust representation for sentiment classification.

All deep learning models utilize a shared embedding layer for text inputs (initialized randomly unless pretrained embeddings are explicitly used, like GloVe as mentioned in previous discussions, which would be integrated here). Categorical features are passed through their own embedding layers, and numerical features are passed through a dense layer before concatenation. The combined features then pass through a common dense layer (64 units, ReLU activation, 0.4 Dropout) leading to a final dense layer with softmax activation for 3-class sentiment prediction. Models are compiled with Adam optimizer and sparse_categorical_crossentropy loss.

III.5.5 Hybrid Stacking Ensemble

A meta-classifier (Logistic Regression) is implemented to combine the predictions of the individual baseline ML and DL models. This ensemble approach aims to leverage the strengths of diverse models by training a higher-level model on their output probabilities (or decision functions), thereby capturing complementary information and potentially achieving superior overall performance and robustness. The meta-classifier is trained on the predictions generated by the base models on the training set and then evaluated on the test set.

IV. RESULTS AND DISCUSSIONS

IV.1 EXPERIMENTAL RESULTAS

This section presents a comprehensive evaluation of the predictive models developed for classifying customer satisfaction from textual reviews. Our experimental methodology involved a comparative analysis of established Machine Learning (ML) algorithms, advanced Deep Learning (DL) architectures, and a novel Hybrid Stacking ensemble model. The primary objective of these experiments was to assess the efficacy of each approach in accurately categorizing customer feedback into 'Dissatisfied', 'Neutral', and 'Satisfied' classes. The subsequent subsections detail the performance metrics achieved by each model, providing insights into their individual capabilities and the synergistic advantages of the hybrid approach, thereby laying the groundwork for understanding their practical utility in sentiment analysis. Predictive analytics were performed to classify customer satisfaction into 'Dissatisfied', 'Neutral', and 'Satisfied' categories based on review data. This section details the performance of various Machine Learning (ML) and Deep Learning (DL) models, along with an integrated Hybrid Stacking ensemble.

Table 3: Performance Metrics of Machine Learning, Deep Learning, and Hybrid Stacking.

Model	Type	Accuracy	Precision	Recall	F1-score
LSTM	DeepLearning	0.8503	0.6736	0.6729	0.6732
Dense	DeepLearning	0.8385	0.6427	0.6051	0.6112
GRU	DeepLearning	0.8491	0.6742	0.6645	0.6692
CNN	DeepLearning	0.8503	0.6741	0.6694	0.6716
MultinomialNB	MachineLearning	0.812	0.7032	0.5065	0.5146
LogisticRegression	MachineLearning	0.8454	0.672	0.6092	0.6175
XGBoost	MachineLearning	0.8202	0.6919	0.5384	0.5559
LinearSVC	MachineLearning	0.8371	0.6462	0.5978	0.602
HybridStack	Hybrid	0.8593	0.6919	0.686	0.6889

Source: Authors, (2026).

Upon detailed examination of the results, several key observations emerged, painting a clear picture of the models' capabilities:

• Deep Learning Superiority in Text Comprehension:

Deep learning models consistently outperformed traditional machine learning approaches in understanding customer reviews. Recurrent architectures like LSTM and GRU exhibited particularly strong capabilities in capturing sequential patterns. The LSTM model achieved a notable accuracy of 0.8503 and an F1-score of 0.6732, indicating a well-balanced performance in both precision and recall. GRU followed closely with an accuracy of 0.8491 and an F1-score of 0.6692, further validating the effectiveness of recurrent neural networks for sentiment classification. The CNN model also demonstrated solid results, with an F1-score of 0.6716, underscoring the power of convolutional layers in extracting local text features. These findings highlight deep learning's aptitude for modeling the complex, high-dimensional patterns essential in sentiment analysis tasks.

• Strong Traditional ML Baselines:

Although not matching the performance of deep learning models, conventional machine learning algorithms delivered competitive baseline results. Logistic Regression and LinearSVC recorded F1-scores of approximately 0.6175 and 0.6020, respectively reflecting their reliability in linear classification. XGBoost, leveraging gradient boosting, achieved a respectable F1-score of 0.5559. Interestingly, Multinomial Naive Bayes excelled in recall (0.7032), effectively identifying most positive instances, but this came at the cost of lower precision (0.5065), resulting in an overall F1-score of 0.5146. This trade-off indicates its tendency to produce more false positives despite strong sensitivity to relevant reviews.



Figure 7: Comparative Performance Metrics of Machine Learning, Deep Learning, and Hybrid Stacking Models.

Source: Authors, (2026).

The Power of Hybrid Stacking: Our HybridStack model emerged as the undisputed top performer across all metrics. It achieved the highest accuracy of **0.8593**, a leading precision of **0.6919**, a strong recall of **0.6860**, and an unparalleled F1-score of **0.6889**. This outcome is a direct testament to the efficacy of ensemble learning, specifically stacking. By combining the strengths and mitigating the weaknesses of both the advanced Deep Learning models and the robust Machine Learning algorithms, the HybridStack managed to synthesize their predictions into a more accurate, reliable, and generalized final output. This indicates that a multifaceted approach, harnessing diverse learning paradigms, offers the most promising path for complex text classification tasks like customer satisfaction prediction. These findings collectively underscore the significant potential of both advanced Deep Learning architectures and intelligently designed hybrid ensembles for achieving highly accurate and robust customer satisfaction prediction.

IV.2 DISCUSSION

This study demonstrates a substantial advancement in customer satisfaction prediction within e-commerce through the development of the HybridStack model. The architecture embodies a synergistic integration of Machine Learning (ML) and Deep Learning (DL) paradigms designed to capture the complex linguistic, semantic, and statistical characteristics of textual customer reviews. Its success underscores the value of ensemble-based hybridization in producing robust and generalizable predictive models for sentiment analysis, a domain where traditional approaches often struggle with variability and data imbalance. The HybridStack architecture was strategically designed to harness the complementary strengths of diverse learning algorithms. It integrates advanced DL models—such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), and Convolutional Neural Networks (CNN) that capture deep contextual and sequential dependencies, with classical ML models such as Logistic Regression and LinearSVC, which offer interpretability and computational stability. This combination ensures that the ensemble benefits from both the expressive power of deep learning and the reliability of traditional classifiers. The empirical evaluation revealed that the HybridStack achieved consistently superior results across all major performance metrics: an accuracy of 0.8593, a precision of 0.6919, a recall of 0.6860, and an F1-score of 0.6889. These results validate the hybrid ensemble strategy's ability to synthesize diverse predictive insights and balance trade-offs between precision and recall.

The model effectively mitigates overfitting tendencies common in deep architectures and the limited contextual understanding characteristic of simpler ML models. This synergy provides a comprehensive analytical framework capable of managing the heterogeneity and noise inherent in real-world textual data. Individual models demonstrated competitive but lower performance levels. Among DL models, both LSTM and CNN achieved F1-scores around 0.67, confirming their strength in capturing syntactic and semantic patterns within reviews. GRU models showed similar results, reinforcing the reliability of recurrent structures in sentiment classification. In contrast, ML models such as Logistic Regression and LinearSVC reached F1-scores of approximately 0.61, illustrating that while efficient and interpretable, they are less capable of modeling complex linguistic patterns. These findings confirm the validity of the ensemble methodology, but also highlight intrinsic challenges that constrain model performance. Customer review datasets in e-commerce are rarely balanced—positive and neutral feedback typically dominate, while negative or highly dissatisfied opinions form a minority. This class imbalance limits the achievable recall for underrepresented classes, as models tend to focus on patterns from majority categories. Despite the HybridStack’s advanced learning capacity, such imbalance inherently reduces the maximum attainable performance metrics, particularly for rare sentiment classes.

Additionally, linguistic variability remains a critical limitation. Real-world customer reviews frequently contain informal expressions, idiomatic phrasing, abbreviations, mixed languages, and grammatical inconsistencies. These features introduce semantic noise and hinder the model’s ability to generalize across diverse linguistic contexts. Even deep neural networks with high representational capacity face difficulties when processing multilingual or code-switched data, as vocabulary and syntax variability can obscure sentiment signals. This phenomenon contributes to slight metric fluctuations and justifies the observed F1-score ceiling in the study’s results. In this context, it is important to note that in multi-class sentiment classification, particularly when dealing with unbalanced, noisy, or multilingual textual datasets, the literature widely recognizes F1-scores between 0.65 and 0.75 as both realistic and acceptable indicators of robust model performance. Numerous studies on e-commerce, social media, and product review analytics report comparable outcomes when models contend with similar data complexities. Therefore, the HybridStack’s F1-score of 0.6889 represents not a limitation, but a credible and empirically valid result consistent with established benchmarks in the field.

The model’s balanced precision and recall demonstrate its practical utility for business intelligence and Customer Relationship Management (CRM). By reliably identifying both positive and negative sentiment classes, the HybridStack enables more accurate segmentation of customers and more informed managerial decisions. The model can support personalized marketing, early detection of dissatisfaction, and data-driven service improvements—all vital for enhancing customer experience and loyalty. Furthermore, the study confirms that hybrid stacking architectures offer methodological resilience. By combining models with diverse learning behaviors, the HybridStack achieves higher generalization, reduces sensitivity to noise, and provides interpretable predictions. The ensemble’s meta-learning process allows it to optimize weighting among component models dynamically, ensuring stable results even when input data quality varies. This robustness makes the approach particularly relevant for e-commerce platforms managing massive, constantly evolving datasets. While the achieved results are promising, certain limitations must be acknowledged. The dataset used in this research reflects specific linguistic and contextual characteristics; its generalizability to other domains, industries, or languages requires further testing. The HybridStack, though comprehensive, does not exhaust all potential ML and DL model combinations; future studies could integrate transformer-based models such as BERT, RoBERTa, or GPT for deeper semantic understanding.

Moreover, computational efficiency was not the primary focus of this work—future implementations may explore optimized, lightweight ensembles for real-time applications. Despite these limitations, the HybridStack’s strong and balanced performance, situated within realistic expectations for noisy and imbalanced sentiment data, affirms its external validity and practical relevance. It demonstrates how a hybrid ensemble approach can bridge the gap between high-performing but opaque deep networks and interpretable, lightweight ML classifiers. The resulting framework supports more accurate and explainable sentiment prediction—an essential step toward transparent and actionable AI for CRM systems. In conclusion, this research confirms that the integration of ML and DL methods through an optimized hybrid stacking ensemble offers a powerful and adaptable solution for e-commerce sentiment analysis. The model’s validated performance, realistic metrics, and interpretability collectively advance the field of customer satisfaction prediction. By addressing data imbalance, linguistic variability, and the inherent complexity of textual sentiment, the HybridStack provides a scalable, explainable, and effective framework for organizations seeking to enhance customer experience through data-driven intelligence.

V.CONCLUSIONS

This study successfully demonstrates the efficacy and practical relevance of a hybrid ensemble modeling approach for predicting customer satisfaction from textual review data. The proposed HybridStack model achieved an F1-score of 0.6889 and an accuracy of 0.8593, outperforming all individual Deep Learning and traditional Machine Learning baselines. These results confirm the validity of the hybrid approach, which effectively integrates the complementary strengths of advanced neural architectures (LSTM, GRU, CNN) and interpretable linear classifiers to deliver robust, accurate, and generalizable sentiment predictions. The model’s ability to synthesize diverse algorithmic perspectives ensures balanced performance across precision and recall, providing businesses with more reliable and granular insights into customer satisfaction. The obtained results should, however, be interpreted within the contextual constraints of real-world data. Customer review datasets in e-commerce environments are inherently unbalanced, with minority sentiment classes particularly dissatisfied or highly negative reviews significantly underrepresented.

This imbalance limits the model’s exposure to rare sentiment categories, constraining the theoretical upper bound of achievable recall and F1-score. Furthermore, the linguistic variability of customer reviews, often characterized by informal language, idiomatic expressions, spelling inconsistencies, and multilingual content, introduces additional complexity. These factors collectively make perfect classification performance unattainable even for state-of-the-art architectures, as models must contend with semantic ambiguity and inconsistent linguistic structures. In this regard, the HybridStack’s F1-score of 0.6889 represents a strong and credible level of performance. In the broader literature on multi-class sentiment classification, particularly in cases involving noisy, imbalanced, or multilingual text, an F1-score between 0.65 and 0.75 is widely regarded as both acceptable and indicative of model robustness.

Thus, the results obtained here not only fall within this empirically validated range but also confirm the model's ability to generalize effectively under realistic data conditions. This reinforces the external validity and practical utility of the HybridStack as a reliable predictive framework for customer sentiment analysis. Beyond its quantitative success, the study offers significant practical implications for Customer Relationship Management (CRM) and e-commerce strategy. By providing accurate, data-driven sentiment predictions, the model enables more nuanced understanding of customer perceptions, supports proactive service enhancement, and facilitates personalized marketing interventions aimed at increasing satisfaction and loyalty. Its interpretability and adaptability also make it suitable for integration into real-time business analytics pipelines, where robust yet explainable AI systems are essential for informed decision-making. Looking forward, several avenues can extend this research. Expanding the system to handle multilingual reviews remains a priority, given the global nature of e-commerce.

Leveraging cross-lingual embeddings and multilingual transformer architectures could further improve performance on diverse linguistic datasets. Additionally, exploring multimodal analysis that integrates textual, visual, and auditory data (e.g., product images or vocal sentiment in voice reviews) could enrich sentiment interpretation and enhance predictive precision. Finally, fine-tuning Large Language Models (LLMs) on domain-specific review corpora, such as those from Amazon or Jumia, holds promise for capturing subtle contextual cues, sarcasm, or domain-specific expressions that remain challenging for traditional architectures. In summary, this research validates the HybridStack model as a robust, scalable, and contextually reliable framework for customer satisfaction prediction. Despite natural constraints linked to data imbalance and linguistic diversity, the model achieves performance consistent with leading literature standards. Its demonstrated effectiveness and interpretability mark an important step toward developing AI systems that combine analytical depth with business applicability—ultimately empowering organizations to transform customer feedback into actionable intelligence and sustained competitive advantage.

VI. AUTHOR'S CONTRIBUTION

Conceptualization: Mahdjoub Mohammed Bachir, Laallam Fatima Zohra and Mezati Messaoud.

Methodology: Mahdjoub Mohammed Bachir, Laallam Fatima Zohra and Mezati Messaoud.

Investigation: Mahdjoub Mohammed Bachir, Laallam Fatima Zohra and Mezati Messaoud.

Discussion of results: Mahdjoub Mohammed Bachir, Laallam Fatima Zohra and Mezati Messaoud.

Writing – Original Draft: Mahdjoub Mohammed Bachir.

Writing – Review and Editing: Mahdjoub Mohammed Bachir, Laallam Fatima Zohra and Mezati Messaoud.

Resources: Mahdjoub Mohammed Bachir.

Supervision: Mahdjoub Mohammed Bachir.

Approval of the final text: Mahdjoub Mohammed Bachir, Laallam Fatima Zohra and Mezati Messaoud.

VII. REFERENCES

- [1] J. W. Kincaid, *Customer Relationship Management (Getting it Right)*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 2003, p. 41.
- [2] F. Rajola, *Customer Relationship Management (Organisational and Technological Perspectives)*. Berlin, Germany: Springer, 2013, p. 25.
- [3] V. Kumar and W. Reinartz, *Customer Relationship Management: Concepts, Strategy and Tools*. Berlin, Germany: Springer, 2018, p. 35.
- [4] J. Kumar and K. S. Murthy, *Customer Relationship Management*. Andhra Pradesh, India: Ky Publications, 2021, p. 10.
- [5] A. K. Rai, *Customer Relationship Management: Concepts and Cases*. New Delhi, India: PHI Learning, 2012, p. 30.
- [6] F. Buttle and S. Maklan, *Customer Relationship Management: Concepts and Technologies*. Abingdon, U.K.: Taylor & Francis, 2015, p. 16.
- [7] L. A. Chie, F. W. Jalaludin, C. S. Seah, Y. X. Loh, and K. W. Hen, "Customer satisfaction towards the application of artificial intelligence in e-commerce," *ELSP Proc. Comput. Sci.*, vol. 2023, no. 2, p. 0019, 2023. doi: 10.55092/pcs2023020019.
- [8] N. K. Tiwari and H. Arora, "A review of sentiment analysis and forecasting techniques for enhancing business performance in e-commerce using machine learning algorithms," *Int. J. Eng. Trends Appl.*, vol. 12, no. 1, p. 5, 2025.
- [9] F. O. Adesina-Adebayo, "Machine learning models for forecasting customer satisfaction," *IOSR J. Comput. Eng.*, vol. 26, no. 2, pp. 01–09, 2024.
- [10] X. Zhang and C. Guo, "Research on multimodal prediction of e-commerce customer satisfaction driven by big data," *Appl. Sci.*, vol. 14, p. 8181, 2024. doi: 10.3390/app14188181.
- [11] C. Guo and X. Zhang, "Techniques to balance profitability and customer satisfaction in dynamic pricing models in e-commerce: Exploring machine learning techniques," *IEEE Access*, vol. 13, 2025. doi: 10.1109/ACCESS.2025.3563371.
- [12] R. Fedorko, R. Bačík, and M. Fedorková, "Role of artificial intelligence and machine learning in e-commerce: A literature review," *ADCAIJ: Adv. Distrib. Comput. Artif. Intell. J.*, vol. 14, p. e31736, 2025.
- [13] Y. Hu, L. Zhao, and W. Tang, "Consumer purchase behavior prediction on e-commerce platforms using machine learning fusion," *Scalable Comput. Pract. Exp.*, vol. 26, no. 4, pp. 1831–1840, 2025. doi: 10.12694/scpe.v26i4.4641.
- [14] R. Ketipov, V. Angelova, L. Doukovska, and R. Schnalle, "Predicting user behavior in e-commerce using machine learning," *Cybern. Inf. Technol.*, vol. 23, no. 3, 2023. doi: 10.2478/cait-2023-0026.

- [15] L. Li, L. Yuan, and J. Tian, "Influence of online e-commerce interaction on consumer satisfaction based on big data algorithm," *Research Article*, vol. 9, no. 8, p. e18322, 2023.
- [16] M. A. Al Rahib, N. Saha, R. Mia, and A. Sattar, "Customer data prediction and analysis in e-commerce using machine learning," *Bull. Electr. Eng. Informatics*, vol. 13, no. 4, pp. 2624–2633, 2024. doi: 10.11591/eei.v13i4.6420.
- [17] N. Singh, P. Singh, and M. Gupta, "An inclusive survey on machine learning for CRM: A paradigm shift," *Decision*, 2021. doi: 10.1007/s40622-021-00256-6.
- [18] C. C. Ike, A. B. Ige, S. A. Oladosu, P. A. Adepoju, O. O. Amoo, and A. I. Afolabi, "Advancing machine learning frameworks for customer retention and propensity modeling in e-commerce platforms," *GSC Adv. Res. Rev.*, vol. 14, no. 2, pp. 191–203, 2023. doi: 10.30574/gscarr.2023.14.2.0017.
- [19] F. U. Ojika, W. O. Owobu, O. A. Abieba, O. J. Esan, B. C. Ubamadu, and A. I. Daraojimba, "Enhancing user interaction through deep learning models: A data-driven approach to improving consumer experience in e-commerce," *J. Front. Multidiscip. Res.*, vol. 4, no. 1, pp. 126–137, 2023. doi: 10.54660/IJFMR.2023.4.1.126-137.
- [20] S. F. Shetu, I. Jahan, M. M. Islam, R. A. Hossain, N. N. Moon, and F. N. Nur, "Predicting satisfaction of online banking system in Bangladesh by machine learning," in *Proc. Int. Conf. Artif. Intell. Comput. Sci. Technol. (ICAICST)*, 2021.
- [21] I. Markoulidakis, I. Rallis, I. Georgoulas, G. Kopsiaftis, A. Doulamis, and N. Doulamis, "A machine learning-based classification method for customer experience survey analysis," *Technologies*, vol. 8, no. 4, 2020. doi: 10.3390/technologies8040076.
- [22] S. K. Chatratha, G. S. Batra, and Y. Chaba, "Handling consumer vulnerability in e-commerce product images using machine learning," *Sci. Inspires*, vol. 8, no. 9, p. e10743, 2022.
- [23] Y. Piris and A.-C. Gay, "Customer satisfaction and natural language processing," *J. Bus. Res.*, vol. 124, pp. 264–271, 2021. doi: 10.1016/j.jbusres.2020.11.065.
- [24] M. Chiny, O. Bencharef, and Y. Chihab, "Towards a machine learning and data mining approach to identify customer satisfaction factors on Airbnb," in *Proc. Int. Conf. Optim. Appl. (ICOA)*, 2021. doi: 10.1109/ICOA51614.2021.9442657.
- [25] M. Chiny, O. Bencharef, M. Y. Hadi, and Y. Chihab, "A client-centric evaluation system to evaluate guest's satisfaction on Airbnb using machine learning and NLP," *Appl. Comput. Intell. Soft Comput.*, 2021, Art. no. 6675790. doi: 10.1155/2021/6675790.
- [26] P. Priyadarshini and K. T. Veeramanju, "Business intelligence for the evaluation of customer satisfaction in e-commerce websites: A case study," *Int. J. Manage., Technol. Social Sci.*, vol. 7, no. 2, p. 661, 2022.
- [27] N. Nabi et al., "Unleashing deep learning: Transforming e-commerce profit prediction with CNNs," *J. Bus. Manage. Stud.*, vol. 6, no. 2, pp. 126–131, 2024. doi: 10.32996/jbms.2024.6.2.13.
- [28] M. Loukili, F. Messaoudi, and M. El Ghazi, "Machine learning based recommender system for e-commerce," *Int. J. Artif. Intell.*, vol. 12, no. 4, pp. 1803–1811, 2023. doi: 10.11591/ijai.v12.i4.pp1803-1811.
- [29] D. Liu, H. Huang, H. Zhang, X. Luo, and Z. Fan, "Enhancing customer behavior prediction in e-commerce: A comparative analysis of machine learning and deep learning models," in *Proc. 4th Int. Conf. Signal Process. Mach. Learn.*, 2024. doi: 10.54254/2755-2721/55/20241475.
- [30] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL, USA: CRC Press, 2012.
- [31] R. Polikar, "Ensemble learning," *Scholarpedia*, vol. 4, no. 1, p. 2776, 2009.
- [32] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.
- [33] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
- [34] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, 1997.
- [35] D. H. Wolpert, "Stacked generalization," *Neural Netw.*, vol. 5, no. 2, pp. 241–259, 1992.