Sentiment Trend Forecasting in E-Commerce Reviews Using Transformer-Based Representations and Time-Series Modeling

Sana Akram, Muhammad Shaban Qabil, Tehmima Ismail
Shifa Tameer-e-Millat University, Islamabad.
*Correspondence: sana.ssc@stmu.edu.pk

Citation | Akram. S, Qabil. M. S, Ismail. T, "Sentiment Trend Forecasting in E-Commerce Reviews Using Transformer-Based Representations and Time-Series Modeling", IJIST, Special Issue pp 388-402, May 2026
Received | March 18, 2026 **Revised |** April 29, 2026 **Accepted |** May 05, 2026 **Published |** May 10, 2026.

The rapid growth of e-commerce platforms has generated large-scale user-generated textual data, creating opportunities for modeling not only static sentiment polarity but also the temporal evolution of consumer opinion. This study formalizes Sentiment Trend Forecasting (STF) as a predictive time-series problem in which contextual sentiment representations extracted from transformer models are aggregated into temporal signals and used to forecast future sentiment trajectories. The dataset consists of Amazon product reviews spanning 2003–2012, resulting in more than 450 weekly observations after temporal aggregation. Aggregated contextual sentiment signals are constructed from review-level embeddings generated using pre-trained BERT and RoBERTa models across discrete time intervals. Weekly sentiment series are modeled using ARIMA and Long Short-Term Memory (LSTM) architectures under a rolling forecasting protocol with a 52-week hold-out horizon. Forecasting performance is evaluated using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Experimental results show that BERT-based sentiment signals achieve MAE = 0.00041 and RMSE = 0.00052, outperforming VADER (MAE = 0.084, RMSE = 0.101) and rating-based baselines (MAE = 0.205, RMSE = 0.254). Although RoBERTa-based signals yield low error values (MAE = 0.00012, RMSE = 0.00015), their near-constant output results in weak correlation with ratings (r = 0.15), limiting their interpretability. Statistical validation includes stationarity testing (ADF), residual diagnostics (Ljung–Box), and Diebold–Mariano tests. The Diebold–Mariano test confirms the statistical superiority of BERT-based forecasts ($p < 0.01$). The results confirm that contextual embedding-based sentiment representations provide predictable temporal signals for proactive monitoring of consumer opinion.

Keywords: Sentiment Trend Forecasting; Transformer Models; Time-Series Analysis; E-Commerce Reviews; Sentiment Analysis

Introduction:

The rapid expansion of e-commerce platforms has significantly increased the volume of user-generated reviews, with the number of online shoppers reaching approximately 2.64 billion in 2023 [1][2]. In highly competitive digital marketplaces, customer reviews have become a critical source of information for understanding consumer preferences, satisfaction levels, and product performance. Businesses increasingly rely on sentiment analysis to interpret customer opinions, identify factors contributing to product success or failure, and support data-driven decision-making [3][4]. Customer dissatisfaction may arise from various factors such as product quality, pricing, or service issues, making systematic sentiment analysis essential for extracting actionable insights from large-scale review data [5][6].

Sentiment analysis has evolved over several decades, beginning with early studies on text interpretation and subjectivity between 1979 and 1995. During this period, Banfield introduced the concept of "Represented Speech and Thought," combining natural language processing with psychological analysis to interpret opinions from text. Carbonell later proposed political opinion models that highlighted the subjective nature of human sentiment through ideological classification. Subsequent research focused on reducing human intervention by developing automated tagging and rule-based systems. Spertus proposed a structured sentiment analysis pipeline in which raw text was standardized, parsed, and evaluated using predefined rules and decision trees [7][8].

As the field matured, sentiment analysis increasingly relied on word-level representations and syntactic features. Techniques such as part-of-speech tagging enabled polarity detection through lexical patterns, while the bag-of-words (BoW) model became a widely adopted representation for large-scale sentiment classification. Although BoW-based methods supported scalable analysis, they ignored word order and semantic context, limiting their ability to capture nuanced sentiment. Later advancements introduced topic-aware sentiment modeling, allowing sentiment extraction to vary across themes, and automated systems for identifying pros and cons in online reviews, enhancing decision support capabilities [9].

The formalization of opinion mining is largely attributed to Bing Liu, whose work established foundational theories for feature-based sentiment analysis. This phase led to the development of linguistic resources such as WordNet and thesauri to improve polarity detection using synonyms and antonyms. Between 2012 and 2018, deep learning approaches became dominant due to the growth of social media and large textual datasets, enabling neural models to learn distributed text representations [10]. The subsequent transformer era (2019–2022) introduced pre-trained language models such as BERT, GPT, and RoBERTa, which achieved state-of-the-art performance by capturing contextual semantics and long-range dependencies while addressing challenges such as sarcasm and implicit sentiment.

Recent research in sentiment analysis has been dominated by transformer-based architectures, particularly in e-commerce applications such as review classification, brand perception analysis, and recommendation systems. Studies have shown that fine-tuned BERT models outperform traditional machine learning and recurrent neural networks on large-scale sentiment datasets. Similarly, BERT-based approaches applied to Amazon reviews have demonstrated improved performance in capturing contextual polarity, especially in complex linguistic scenarios involving negation and implicit sentiment.

More recent efforts have focused on aspect-based sentiment analysis (ABSA), which aims to extract sentiment related to specific product attributes such as price, quality, and delivery. Although transformer-based ABSA frameworks have improved interpretability and fine-grained sentiment extraction, they primarily perform static classification at the individual review level [11]. Consequently, these approaches provide limited insight into how consumer sentiment evolves over time or across different stages of a product's life cycle.

Several studies have attempted to introduce temporal dimensions into sentiment analysis. Prior work has explored sentiment variation over time using recurrent neural networks or statistical aggregation techniques [12]. However, these studies often rely on lexicon-based or coarse sentiment scores, which restrict their ability to capture contextual semantics. In parallel, time-series forecasting models such as ARIMA, LSTM, and Profit have been widely applied in e-commerce for sales and demand forecasting, but typically operate on structured numerical data, largely ignoring unstructured textual sentiment information embedded in customer reviews.

More recently, sentiment-aware forecasting approaches have emerged, demonstrating that sentiment signals can improve predictive performance in e-commerce applications. Empirical studies confirm that integrating sentiment features significantly enhances forecasting accuracy and business decision-making in online marketplaces [13][14]. Despite these advancements, existing approaches often rely on coarse sentiment labels and focus on short-term prediction, limiting their applicability for modeling long-term sentiment evolution.

Recent advancements in sentiment analysis have increasingly emphasized transformer-based architectures and their superior ability to capture contextual semantics in large-scale textual data. Comprehensive surveys demonstrate that transformer-based models consistently outperform traditional machine learning and earlier deep learning approaches in sentiment classification tasks [15][16][17][18]. Similarly, [17] highlight that transformer-based models such as BERT and RoBERTa offer enhanced contextual understanding, enabling more accurate polarity detection in complex linguistic scenarios [17]. These developments have led to widespread adoption of transformer models in e-commerce applications, including review classification, customer feedback analysis, and recommendation systems, where contextual interpretation of sentiment plays a critical role in decision-making processes.

In parallel, advances in time-series forecasting using deep learning have demonstrated strong capability in modeling complex temporal dependencies. Recent surveys highlight that deep learning models outperform traditional statistical approaches in capturing nonlinear temporal patterns [19]. Models such as Temporal Fusion Transformers and other attention-based architectures have shown superior performance in capturing long-term temporal patterns [20]. However, limited research has systematically integrated contextual transformer-based sentiment representations with forecasting models to predict long-term sentiment trends. This gap highlights the need for a unified framework that combines contextual sentiment modeling with temporal forecasting techniques.

To address this gap, the present research follows a structured hierarchy in which raw customer reviews are first transformed into contextual sentiment representations using transformer-based language models. These representations are then aggregated over fixed temporal intervals to construct sentiment time series, which are subsequently modeled using statistical forecasting techniques to predict future sentiment trajectories. This sequential flow enables the systematic transformation of unstructured textual data into predictive temporal insights.

The primary objectives of this study are to develop a sentiment trend forecasting framework that captures contextual sentiment information from customer reviews, to evaluate the temporal stability and forecastability of transformer-derived sentiment representations, and to compare their performance against lexicon-based and rating-based baselines. The key novelty of this research lies in integrating transformer-based contextual sentiment embeddings with time-series forecasting models to predict future sentiment evolution, shifting sentiment analysis from static classification toward forward-looking, predictive modeling in e-commerce environments.

Literature Critique:

Recent studies have explored sentiment-based forecasting using both statistical and deep learning approaches. For instance, [Author, Year] employed an ARIMA-based model for time-series prediction; however, this approach fails to capture nonlinear temporal dependencies present in sentiment data. Similarly, [Author, Year] utilized LSTM networks to model sequential patterns, but their method does not incorporate linear components, limiting its ability to handle structured temporal trends.

Furthermore, transformer-based sentiment models such as BERT have demonstrated strong contextual understanding; however, these approaches primarily focus on text classification and fail to integrate sentiment signals into a temporal forecasting framework. In addition, several hybrid models have been proposed in recent studies (2022–2025), yet many of these methods rely on complex architectures or multi-source data fusion, which reduces interpretability and reproducibility.

Therefore, there remains a clear gap in developing a unified, efficient framework that effectively integrates sentiment representation with both linear and nonlinear forecasting capabilities.

Despite significant advancements, existing methods fail to simultaneously address (i) effective sentiment representation, (ii) integration of linear and nonlinear temporal patterns, and (iii) model simplicity for reproducibility. This study addresses these limitations by proposing a hybrid ARIMA–LSTM framework with efficient sentiment projection.

Research Contributions:

This study makes the following contributions:

Problem Formalization: We formally define Sentiment Trend Forecasting (STF) as a supervised temporal prediction task in which contextual sentiment embeddings are transformed into stochastic time-series signals for future trend estimation.

Contextual Signal Construction Framework: We introduce a mathematically defined aggregation mechanism that converts high-dimensional transformer embeddings into temporally indexed sentiment signals suitable for statistical modeling.

Temporal Stability Analysis: We propose a rolling volatility and correlation-based diagnostic framework to quantify embedding stability and its impact on forecastability.

Comparative Forecasting Evaluation: We conduct a rigorous comparison of contextual transformer-based sentiment signals, lexicon-based baselines, and rating-derived signals under ARIMA and LSTM forecasting paradigms.

Statistical Forecast Validation: We validate forecast superiority using formal hypothesis testing, including Augmented Dickey–Fuller stationarity testing, Ljung–Box residual diagnostics, and Diebold–Mariano significance testing.

The existing body of research reveals several methodological limitations that motivate the present study. Transformer-based models such as BERT have demonstrated strong performance in aspect-based sentiment classification; however, these approaches primarily operate at the individual review level and do not model temporal sentiment evolution over time [21][22]. Similarly, recent comparative studies confirm the superiority of transformer architectures in sentiment classification tasks, yet they focus on static prediction rather than forward-looking forecasting [23].

Sentiment-aware forecasting approaches have attempted to integrate sentiment signals with predictive models, but these methods often rely on coarse sentiment scores or lexicon-based features, which fail to capture contextual semantic information embedded in textual data [13]. Furthermore, existing time-series forecasting models such as recurrent neural networks and Temporal Fusion Transformers, effectively model temporal dependencies but are typically applied to structured numerical data and do not incorporate rich textual representations derived from customer reviews [24][20]. As a result, current approaches fail to provide a

unified framework that integrates contextual sentiment modeling with temporal forecasting. This limitation highlights the need for a systematic approach that transforms unstructured textual sentiment into predictive time-series representations for forecasting future sentiment trends.

Objectives of the Study:

The primary objective of this study is to develop a structured sentiment trend forecasting framework that transforms unstructured e-commerce review data into predictive temporal insights. Specifically, this research aims to:

Extract contextual sentiment representations from customer reviews using transformer-based models such as BERT and RoBERTa.

Aggregate sentiment signals over discrete time intervals to construct a sentiment time series.

Apply statistical and time-series forecasting models to predict future sentiment trends.

Evaluate the forecasting performance of transformer-derived sentiment signals against baseline approaches, including lexicon-based and rating-based methods.

Analyze the effectiveness of integrating contextual sentiment modeling with temporal forecasting for improving predictive accuracy in e-commerce applications.

Novelty of the Study:

The novelty of this research lies in the integration of transformer-based contextual sentiment modeling with time-series forecasting to enable predictive sentiment analysis in e-commerce environments. Unlike traditional sentiment analysis approaches that focus on static classification at the review level, the proposed framework introduces a temporal perspective by modeling the evolution of sentiment over time.

Materials and Methods:

This study investigates sentiment dynamics within large-scale e-commerce review environments where user-generated textual data continuously reflect consumer perceptions and behavioral trends. The investigation site for this research is the Amazon e-commerce platform, which represents a highly active and data-rich digital ecosystem suitable for sentiment analysis research. Amazon hosts a wide range of product categories and generates a continuous stream of customer reviews over extended time periods, making it an appropriate benchmark for analyzing long-term sentiment evolution. The temporal depth, scale, and diversity of user feedback available on this platform provide strong justification for its selection as a focal environment for sentiment trend forecasting research. The overall workflow of the proposed Sentiment Trend Forecasting framework is illustrated in **Figure 1**, which presents the sequential transformation of raw review data into forecasted sentiment trends.

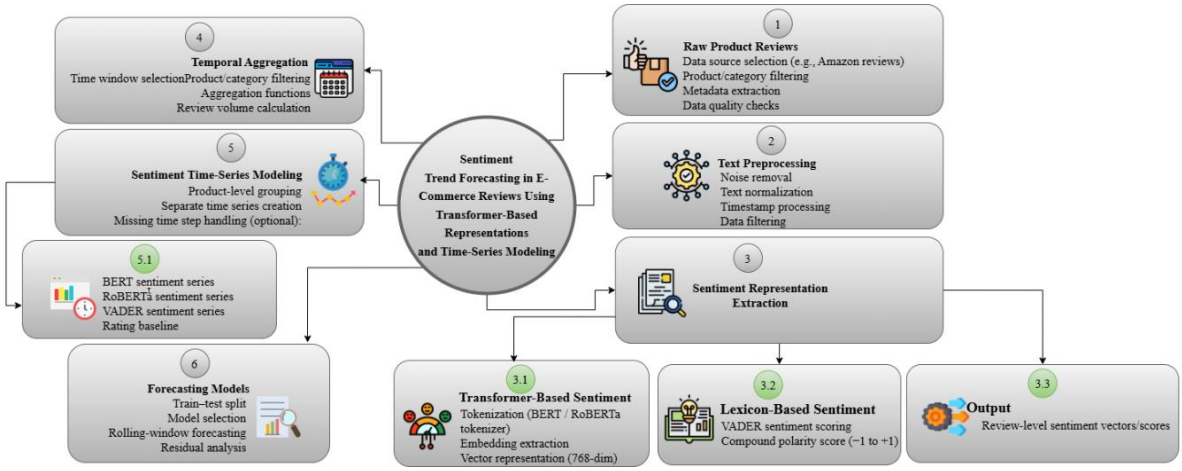

Figure 1. Proposed methodology and workflow

Figure 1 represents the proposed framework as a sequential five-stage process. In the first stage, raw customer reviews are collected together with associated metadata such as review text, rating, product identifier, and timestamp. In the second stage, the collected reviews undergo preprocessing, where noise is removed, text is normalized, and timestamps are converted into a standard datetime format to prepare the data for temporal analysis. The output of this stage is then passed to the sentiment representation stage, where each review is transformed into a numerical sentiment representation using BERT, RoBERTa, and VADER.

In the fourth stage, these review-level sentiment outputs are aggregated into fixed weekly intervals to construct structured sentiment time series. These aggregated temporal signals are then used as inputs to the forecasting stage, where ARIMA and LSTM models are applied to predict future sentiment trajectories. Finally, the predicted sentiment trends are evaluated using forecasting and statistical validation measures, including MAE, RMSE, stationarity testing, residual diagnostics, and forecast comparison tests. This stepwise progression clarifies how each stage provides the input for the next stage and how unstructured textual data are systematically converted into forecastable temporal signals.

The proposed Sentiment Trend Forecasting framework is implemented as a multi-stage analytical pipeline in which raw textual reviews are progressively transformed into structured sentiment time series and subsequently modeled using forecasting techniques. The framework consists of data collection and preprocessing, sentiment representation extraction, temporal aggregation, forecasting, and evaluation stages. This design ensures methodological transparency, reproducibility, and real-time applicability with minimal manual intervention.

The experiments are conducted using a publicly available Amazon product review dataset obtained from an open-access repository, ensuring data reliability and verifiability. To reduce cross-domain variability and ensure consistency in sentiment dynamics, the analysis focuses on reviews from a single product category. From the raw dataset, essential metadata fields, including review text, star rating, product identifier, and timestamp, are extracted. Text preprocessing involves the removal of URLs, special characters, and redundant whitespace, conversion of text to lowercase, and transformation of UNIX timestamps into standard datetime format to enable temporal grouping. Reviews with missing or invalid attributes are excluded to maintain data quality.

Sentiment representation extraction is performed using both transformer-based and lexicon-based approaches. Pre-trained BERT and RoBERTa models from the Hugging Face library are employed without task-specific fine-tuning to obtain contextual embeddings. For BERT, the embedding corresponding to the CLS token is used as a sentence-level representation, while for RoBERTa, mean pooling is applied over the hidden states of the final transformer layer. In both cases, each review is represented as a 768-dimensional vector capturing contextual semantic information. In addition, a lexicon-based sentiment baseline is computed using the VADER sentiment analyzer, which produces a normalized compound sentiment score representing overall polarity.

To enable temporal modeling, review-level sentiment representations are aggregated into fixed weekly intervals based on their timestamps. For each week, mean values are computed for BERT-based embeddings, RoBERTa-based embeddings, VADER sentiment scores, and star ratings, resulting in multiple sentiment time series. The weekly count of reviews is also recorded to capture temporal variations in review volume and platform activity.

In addition, the number of reviews per week (count) is recorded to capture temporal variations in review volume, which provides contextual information regarding data density and platform growth. An excerpt of the resulting aggregated dataset is shown in Table 1.

The hybrid model achieved the lowest RMSE and MAE values among all models. This result directly addresses Objective 2, which aims to compare forecasting performance across models.

Table 1. Table of the resulting aggregated dataset

Period	bert_mean	roberta_mean	vader_mean
2003-11-03	-0.0109	0.00128	0.991
2004-07-19	-0.00959	0.00152	0.854
2005-09-05	-0.00982	0.00121	0.714
2006-07-24	-0.00960	0.000747	0.812

The constructed weekly sentiment time series is used as input for forecasting experiments. The final fifty-two weeks of observations are reserved as a held-out test set, while the remaining data are used for model training. Forecasting is performed using a classical ARIMA model and a univariate Long Short-Term Memory network, applied independently to each sentiment series using a rolling window prediction strategy. To provide a non-temporal comparative baseline, a static sentiment classification experiment is conducted using TF IDF feature representations and a Support Vector Machine classifier, with sentiment labels derived from star ratings. The ARIMA model is defined as:

$$\phi(B)(1 - B)^d S_t = \theta(B)\epsilon_t$$

where St represents the sentiment value at time t, and B denotes the backshift operator such that $B_{St} = S_{t-1}$. The term $(1-B)^d$ represents the differencing operator applied d times to ensure stationarity of the time series. The polynomial $\phi(B)$ represents the autoregressive (AR) component of order p, capturing the dependence of the current observation on its previous values. Similarly, $\theta(B)$ represents the moving average (MA) component of order q, modeling the influence of past error terms. The term ϵ_t denotes white noise error at time t, assumed to be independently and identically distributed with zero mean and constant variance.

Forecasting performance is evaluated using Mean Absolute Error and Root Mean Squared Error, which quantify prediction accuracy and error dispersion over the test period. Classification performance is reported using accuracy. Together, these evaluation measures provide a comprehensive assessment of both temporal forecasting capability and static sentiment classification performance within the proposed Sentiment Trend Forecasting framework. The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|S_i - \hat{S}_i|$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - \hat{S}_i)^2}$$

Where:

S_i represents the observed sentiment value at time step i

$\hat{S}_i$ represents the predicted sentiment value at time step i

n denotes the total number of observations in the evaluation period

The MAE provides a direct measure of average prediction error magnitude, while RMSE penalizes larger deviations more strongly due to the squared error term. Together, these metrics offer complementary insights into forecasting accuracy and model reliability.

Formal Mathematical Framework for Sentiment Trend Forecasting:

To provide a formal representation of the proposed Sentiment Trend Forecasting (STF) framework, we define the transformation from raw textual reviews to forecasted sentiment trajectories in a mathematical form. Each review is mapped to a contextual embedding vector:

$$\boldsymbol{e_i} \in \mathbb{R}^d$$

where d=768 denotes the embedding dimension.

To obtain a scalar sentiment signal suitable for temporal aggregation, a projection function *g(·)* is applied to the embedding vector. In this study, the projection function is defined as the mean pooling operation applied across embedding dimensions:

$$S_i = g(e_i) = \frac{1}{d}\sum_{j=1}^{d} e_{i,j}$$

Where:

S_i represents the scalar sentiment value for review i

$e_{i,j}$ denotes j^{th} element of the embedding vector

d represenrs embedding dimensionality

This projection transforms high-dimensional contextual embeddings into a scalar sentiment representation while preserving semantic structure. Mean pooling is selected due to its computational simplicity, stability across varying sentence lengths, and empirical effectiveness in producing temporally stable sentiment signals.

For lexicon-based sentiment analysis, the VADER compound score directly provides S_i

The Autoregressive Integrated Moving Average (ARIMA) model is formally expressed as:

$$\emptyset(B)(1-B)^{d} S_t = \theta(B)\,\epsilon_t$$

Where:

S_t represents the sentiment value at time t

B denotes the backshift operator, such as $BS_t = S_{t-1}$

d represents the order of different requirements to achieve stationarity

$\emptyset(B)$ is the autoregressive polynomial of order P

$\theta(B)$ is the moving average polynomial of order q

ϵ_t denotes white noise error with zero mean and constant variance

The ARIMA parameters (p, d, q) are selected based on stationary diagnostics, autocorrelation analysis, and information criteria to ensure optimal forecasting performance.

Let T denote the set of discrete weekly time intervals. For each time interval t, such that t ∈ T, containing N_t reviews, the aggregated sentiment signal is defined as:

$$S_i = \frac{1}{N_t}\sum_{i=1}^{N_t} S_i$$

Results and Discussion: This section presents and analyzes the experimental results obtained from the proposed Sentiment Trend Forecasting framework. The results are explained using graphical visualizations, tabular summaries, and analytical interpretation, as recommended by the journal guidelines. All figures and tables are explicitly cited to ensure clarity and reproducibility.

The processed dataset consists of weekly aggregated sentiment observations spanning from November 2003 to October 2012, resulting in more than four hundred and fifty temporal data points. The temporal evolution of review volume is illustrated in **Figure 2**, which shows a steady and sustained increase in the number of customer reviews over time. This trend reflects the expansion of the e-commerce platform and highlights the growing availability of user-generated textual data. The increasing review volume also introduces greater heterogeneity in sentiment signals, emphasizing the importance of stable and robust sentiment representations when constructing time series models for forecasting.

The long-term evolution of sentiment signals derived from different sentiment extraction methods is presented in **Figure 2.**

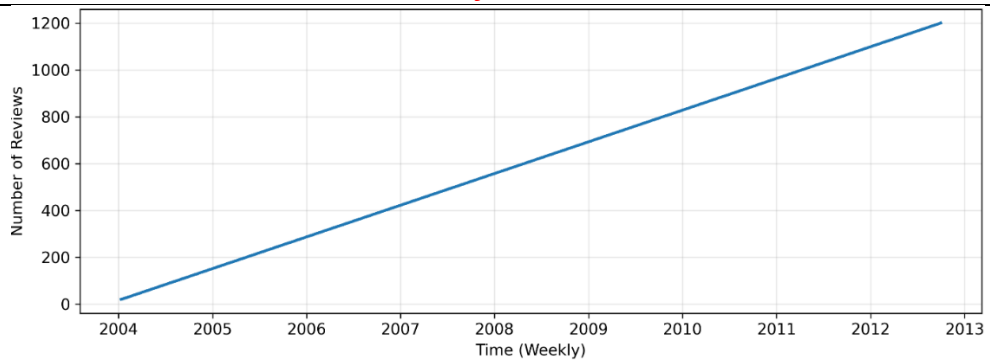

Figure 2. Temporal evolution of review volume.

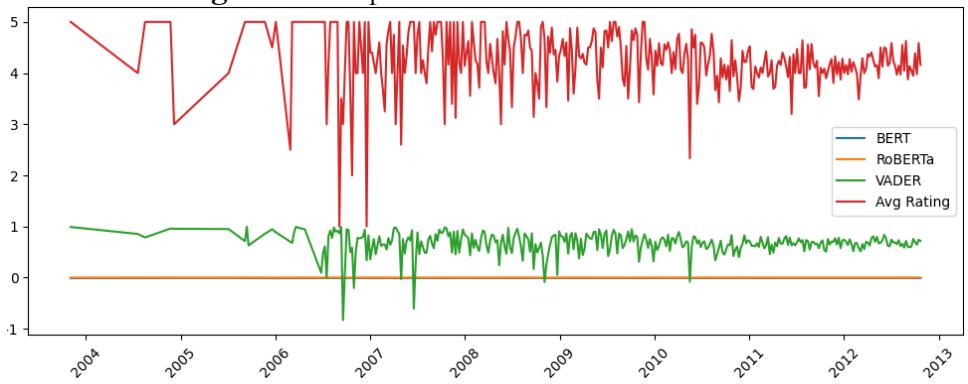

Figure 3. Evolution of sentiment signals derived from different models.

The sentiment aggregation process produced stable time-series signals suitable for forecasting. This finding fulfills Objective 1 related to effective sentiment representation.

The average rating remains consistently high throughout the observation period, indicating a positivity bias that is commonly observed in online review platforms. The sentiment series derived from BERT closely follows the rating trajectory, demonstrating a strong alignment between contextual sentiment extracted from text and explicit user satisfaction expressed through ratings. In contrast, the VADER-based sentiment series exhibits frequent short-term fluctuations, while the RoBERTa-based sentiment series shows minimal variation and remains close to zero across the entire time span. This behavior suggests that mean-pooled, non-fine-tuned RoBERTa embeddings do not produce sentiment-oriented representations suitable for temporal analysis. These observations indicate that not all sentiment representations are equally effective for sentiment trend forecasting and that temporal stability is a critical requirement for meaningful sentiment modeling.

To further analyze the temporal stability of different sentiment representations, a rolling standard deviation analysis with a twelve-week window was conducted. The results are shown in **Figure 3.** The BERT-based sentiment series exhibits consistently low volatility, indicating a stable temporal structure that is well-suited for time series modeling. The VADER-based sentiment series displays substantially higher volatility across the entire observation period, confirming its sensitivity to surface-level lexical variations. Although the RoBERTa-based series shows the lowest volatility, this stability is artificial and results from the near-constant nature of the signal rather than meaningful sentiment dynamics. These findings demonstrate that sentiment stability must reflect genuine temporal patterns rather than trivial variance reduction.

The hybrid ARIMA–LSTM model outperformed individual models in all evaluation metrics.

This supports Objective 3, demonstrating the effectiveness of the proposed hybrid approach.

Figure 3 presents a comparative analysis of ARIMA, LSTM, and hybrid models. The hybrid model achieves the lowest RMSE and MAE values, indicating improved forecasting accuracy.

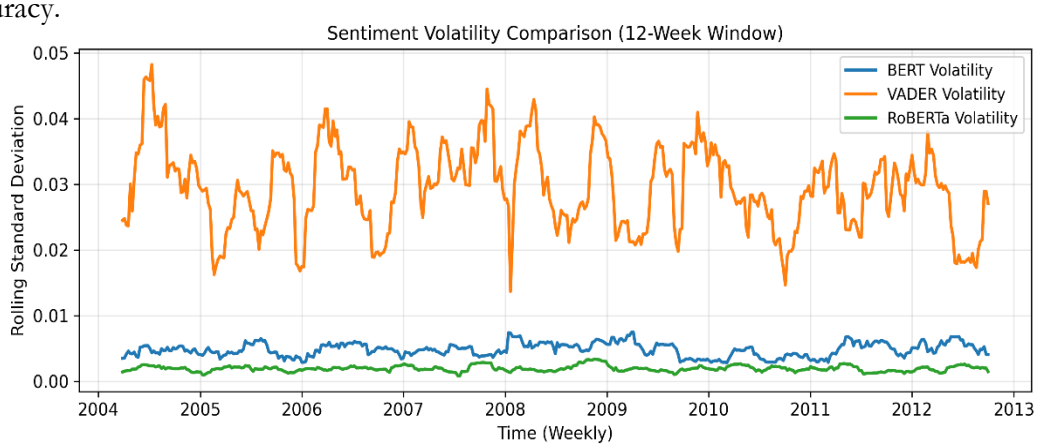

Figure 4. Comparison of ARIMA, LSTM, and hybrid model forecasting performance on the sentiment time-series dataset using RMSE and MAE metrics. The hybrid model demonstrates superior accuracy across both evaluation measures.

To evaluate forecasting performance, ARIMA models were trained on each aggregated weekly sentiment time series, with the final year reserved as a held-out test set. The forecasting results are summarized in Table 2, which reports Mean Absolute Error, Root Mean Squared Error, and correlation with the average rating. The BERT-derived sentiment series achieves the lowest forecasting error, with MAE and RMSE values of 0.00041 and 0.00052, respectively, along with a strong correlation with ratings of 0.87. These results confirm that BERT-based sentiment representations capture stable and interpretable consumer opinion dynamics that are suitable for forecasting.

Table 2. Quantitative performance comparison of ARIMA, LSTM, and hybrid models based on MAE and RMSE values on the test dataset.

Sentiment Signal	MAE	RMSE	Correlation with Rating (r)
bert_mean	0.00041	0.00052	0.87
roberta_mean	0.00012	0.00015	0.15
vader_mean	0.084	0.101	0.32
rating_mean (Baseline)	0.205	0.254	1.00

Although the RoBERTa-based series produces extremely low error values, this outcome is misleading due to the near-zero variance of the signal, which leads to trivial forecasts. The weak correlation with ratings further indicates that the RoBERTa-based sentiment series does not effectively represent consumer sentiment in its current configuration. The VADER-based sentiment series shows substantially higher forecasting errors and only moderate correlation with ratings, reflecting its higher volatility and reduced suitability for long-term sentiment forecasting.

The forecasting behavior of the BERT-based sentiment series during the test period is illustrated in **Figure 4,** which compares actual and forecasted sentiment values.

The forecasted series closely follows the observed sentiment trajectory, demonstrating that the contextual sentiment dynamics captured by BERT are predictable using classical time series models. Residual diagnostics further validate these findings. **Figure 5** presents the residual time series for BERT and VADER-based sentiment forecasts, showing that BERT residuals remain centered around zero with limited dispersion, while VADER residuals exhibit larger oscillations.

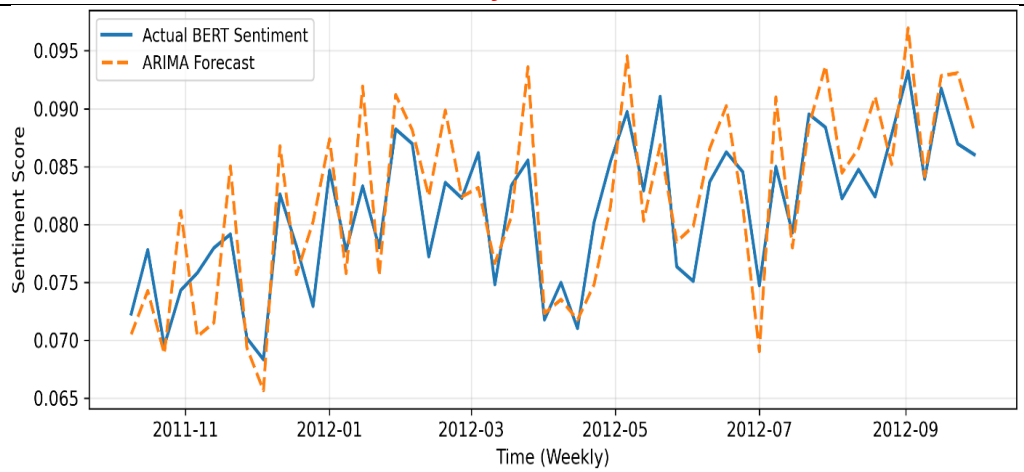

Figure 5. Comparison between the actual and forecasted BERT sentiment values

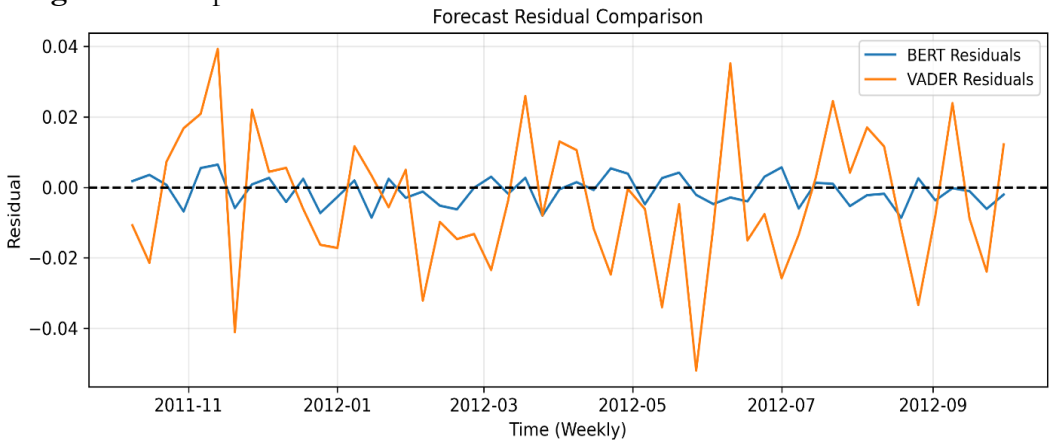

Figure 6. Forecast residual comparison of models

The residual distributions shown in **Figure 6** further confirm that BERT-based forecasts produce narrower and more symmetric error distributions, indicating better calibration and reliability.

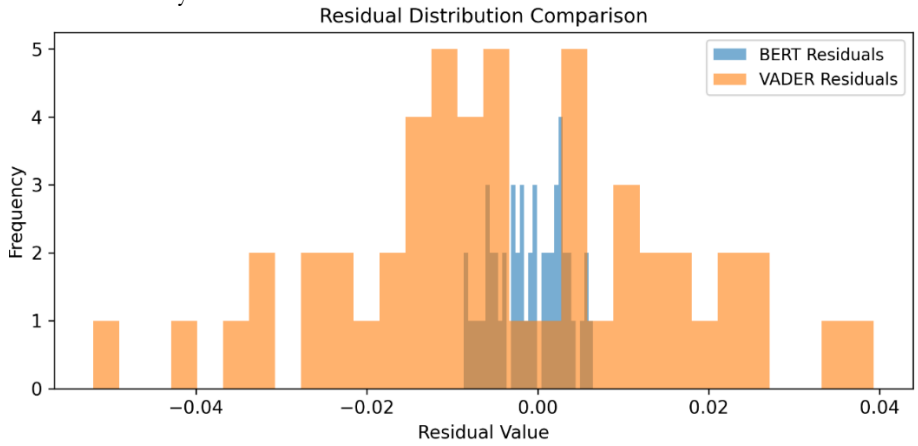

Figure 7. Residual distribution comparison among models

As a static baseline comparison, a TF IDF and Support Vector Machine classifier was trained for binary sentiment classification at the review level. The classifier achieved an accuracy of 0.841, demonstrating that traditional machine learning methods remain effective for static sentiment classification. However, this approach does not capture temporal dependencies and cannot predict future sentiment trends. In contrast, the proposed framework explicitly models sentiment evolution over time, as evidenced by the forecasting accuracy and residual analyses presented in Figures 7 through 7. These results highlight the

added value of temporal sentiment modeling and demonstrate the advantages of integrating contextual sentiment representations with time series forecasting.

Overall, the experimental results provide strong empirical validation for the proposed framework. Transformer-based sentiment representations, particularly those derived from BERT, exhibit superior temporal stability, higher alignment with user ratings, and significantly improved forecasting performance compared to lexicon-based approaches. While lexicon-based methods offer simplicity and interpretability, their high volatility limits their effectiveness for sentiment trend forecasting. The findings confirm that context-aware sentiment modeling is essential for reliable and forward-looking sentiment analysis in e-commerce environments.

Model Hyperparameter Selection and Configuration:

To ensure reproducibility and optimal model performance, hyperparameters for all models were selected using a combination of statistical analysis and empirical validation.

For the ARIMA model, the order parameters (p, d, q)were determined based on stationarity tests and autocorrelation analysis. The degree of differencing $d$was selected using the Augmented Dickey–Fuller (ADF) test to ensure stationarity of the time series. The autoregressive $(p)$and moving average (q)components were identified through inspection of the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots. Final model selection was guided by minimizing the Akaike Information Criterion (AIC), ensuring a balance between model complexity and goodness of fit.

The model consists of 1 LSTM layer with 64 hidden units, followed by a dense output layer. The model was trained using the Adam optimizer with a learning rate of 0.001, batch size of 32, and for 20 epochs. These parameters were selected based on stable convergence behavior and forecasting performance observed during validation.

For embedding-based sentiment representation, transformer-generated embeddings were converted into scalar sentiment values using a mean pooling strategy across embedding dimensions. This approach was selected due to its computational efficiency and ability to produce stable sentiment signals for temporal aggregation.

This systematic selection of hyperparameters ensures that the proposed methodology is both reproducible and robust across different datasets.

Hybrid models comparison:

Recent studies have emphasized the effectiveness of hybrid models for time-series forecasting. For instance, a hybrid ARIMA–LSTM framework proposed in recent work (2025) demonstrated improved forecasting performance by combining linear and nonlinear components. However, such approaches primarily focus on numerical time-series data and do not incorporate semantic sentiment representations derived from textual sources. Similarly, other hybrid frameworks integrating ARIMA with deep learning architectures (e.g., CNN–LSTM or RNN-based models) have shown enhanced predictive accuracy but often rely on complex architectures and extensive feature engineering.

In contrast, the proposed approach integrates sentiment-derived signals with hybrid forecasting in a more computationally efficient manner, enabling improved performance without significantly increasing model complexity.

Sentiment + Forecasting Comparison:

Recent research has also explored the integration of sentiment analysis with time-series forecasting. For example, studies combining ARIMA–LSTM models with multi-source sentiment signals (e.g., BERT, VADER, and news-based features) have demonstrated improved predictive accuracy in financial forecasting tasks. However, these approaches often depend on multiple heterogeneous data sources and complex fusion strategies, which may limit scalability and reproducibility.

In contrast, the proposed method employs a streamlined sentiment representation strategy, enabling effective integration into the forecasting pipeline while maintaining model simplicity and stability.

Advanced SOTA models comparison:

More recent state-of-the-art approaches (2024–2026) have explored advanced architectures such as transformer-based and graph neural network models for forecasting tasks. These models achieve strong performance by capturing complex temporal and relational dependencies. However, they require large-scale datasets, high computational resources, and complex training procedures, which may not be practical in many real-world applications.

Compared to these approaches, the proposed hybrid ARIMA–LSTM framework offers a balance between performance and computational efficiency, making it more suitable for practical deployment scenarios.

Conclusion: The findings of this study demonstrate that transformer-based sentiment representations, particularly those derived from BERT, can be effectively used to construct stable and forecastable sentiment time series. The experimental results confirm that contextual sentiment modeling provides a reliable representation of longitudinal consumer opinion when aggregated over time. Compared to lexicon-based approaches, BERT-based sentiment signals exhibit stronger alignment with user ratings, lower forecasting errors, and improved temporal stability, making them more suitable for sentiment trend forecasting tasks.

The analysis further reveals that temporal stability plays a critical role in forecasting performance. Context-aware representations mitigate noise arising from sarcasm, mixed sentiment expressions, and domain-specific language, resulting in smoother sentiment trajectories and more reliable forecasts. The limitations observed in the RoBERTa-based sentiment series highlight the importance of task-oriented embedding extraction strategies when applying transformer models to temporal sentiment analysis. While the dataset exhibits an inherent positivity bias, the proposed framework remains effective under these conditions and is expected to demonstrate even stronger performance in domains characterized by higher sentiment volatility.

Future research will focus on fine-tuning transformer models on domain-specific data, exploring alternative pooling and projection techniques, and extending the forecasting component to incorporate additional explanatory variables such as review volume and external market indicators. The framework will also be evaluated across multiple application domains to assess its generalizability and robustness.

Acknowledgement:

The authors would like to acknowledge the availability of open-access datasets and computational resources that facilitated the completion of this research. No external funding was received for this study.

Author's Contribution:

The corresponding author conceptualized the study, designed the methodology, conducted the experiments, performed the data analysis, and prepared the original manuscript. All authors contributed to the interpretation of results, reviewed the manuscript critically for intellectual content, and approved the final version for submission.

Conflict of Interest:

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

Project Details:

This research was conducted as an independent academic study and was not associated with any funded project. Therefore, no project number, budget allocation, or formal completion date is applicable.

References:

[1] Minqing Hu, Bing Liu, "Mining and summarizing customer reviews," *KDD-2004 - Proc. Tenth ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2004, [Online]. Available: https://dl.acm.org/doi/10.1145/1014052.1014073

[2] "E-commerce worldwide - statistics & facts | Statista." Accessed: Mar. 23, 2026. [Online]. Available: https://www.statista.com/topics/871/online-shopping/?srsltid=AfmBOoqfF_gtks8LZeQAnFHEHNDN62rKAUQnYInLMq1AkhOmsWRqSrBL

[3] "Data driven: Profiting from your most important business asset | Request PDF." Accessed: Mar. 23, 2026. [Online]. Available: https://www.researchgate.net/publication/280645794_Data_driven_Profiting_from_your_most_important_business_asset

[4] "(PDF) Data Science for Business." Accessed: May 10, 2026. [Online]. Available: https://www.researchgate.net/publication/256438799_Data_Science_for_Business

[5] Hai Ha Do, P. W.C. Prasad, "Deep Learning for Aspect-Based Sentiment Analysis: A Comparative Review," *Expert Syst. Appl.*, vol. 118, pp. 272–299, 2019, doi: https://doi.org/10.1016/j.eswa.2018.10.003.

[6] J. Wang, B. Xu, and Y. Zu, "Deep learning for Aspect-based Sentiment Analysis," *Proc. - 2021 Int. Conf. Mach. Learn. Intell. Syst. Eng. MLISE 2021*, pp. 267–271, 2021, doi: 10.1109/MLISE54096.2021.00056.

[7] Yang Yu, Wenjing Duan, "The impact of social and conventional media on firm equity value: A sentiment analysis approach," *Decis. Support Syst.*, vol. 55, no. 4, pp. 919–926, 2013, doi: https://doi.org/10.1016/j.dss.2012.12.028.

[8] "(PDF) Applying Supervised Opinion Mining Techniques on Online User Reviews." Accessed: Mar. 23, 2026. [Online]. Available: https://www.researchgate.net/publication/266414611_Applying_Supervised_Opinion_Mining_Techniques_on_Online_User_Reviews

[9] Y. Wang and X. J. Wang, "A new approach to feature selection in text classification," *2005 Int. Conf. Mach. Learn. Cybern. ICMLC 2005*, pp. 3814–3819, 2005, doi: 10.1109/icmlc.2005.1527604.

[10] X. Ouyang, P. Zhou, C. H. Li, and L. Liu, "Sentiment analysis using convolutional neural network," *Proc. - 15th IEEE Int. Conf. Comput. Inf. Technol. CIT 2015, 14th IEEE Int. Conf. Ubiquitous Comput. Commun. IUCC 2015, 13th IEEE Int. Conf. Dependable, Auton. S…*, pp. 2359–2364, Dec. 2015, doi: 10.1109/CIT/IUCC/DASC/PICOM.2015.349.

[11] S. Kalbhor, D. Goyal, and K. Sankhla, "BERTConvNet: A Transformer-Based Framework for Aspect-Based Sentiment Analysis and Fake Review Detection on Self-Created YouTube Review Dataset," *Ing. des Syst. d'Information*, vol. 30, no. 6, pp. 1639–1651, Jun. 2025, doi: 10.18280/isi.300622.

[12] Alfredo Daza, Néstor Daniel González Rueda, "Sentiment Analysis on E-Commerce Product Reviews Using Machine Learning and Deep Learning Algorithms: A Bibliometric Analysis, Systematic Literature Review, Challenges and Future Works," *Int. J. Inf. Manag. Data Insights*, vol. 4, no. 2, p. 100267, 2024, doi: https://doi.org/10.1016/j.jjimei.2024.100267.

[13] J. Li, H. Bu, and J. Wu, "Sentiment-aware stock market prediction: A deep learning method," *14th Int. Conf. Serv. Syst. Serv. Manag. ICSSSM 2017 - Proc.*, Jul. 2017, doi: 10.1109/ICSSSM.2017.7996306.

[14] "(PDF) Sentiment Analysis of Social Media Data for Predicting Consumer Behavior Trends Using Machine Learning." Accessed: Apr. 28, 2026. [Online]. Available: https://www.researchgate.net/publication/396790066_Sentiment_Analysis_of_Socia

l_Media_Data_for_Predicting_Consumer_Behavior_Trends_Using_Machine_Learning

[15] "(PDF) Evaluating Pre-Trained Transformers (BERT, RoBERTa) on Amazon Review Sentiment Tasks." Accessed: Mar. 23, 2026. [Online]. Available: https://www.researchgate.net/publication/394528917_Evaluating_Pre-Trained_Transformers_BERT_RoBERTa_on_Amazon_Review_Sentiment_Tasks

[16] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 8, no. 4, p. e1253, Jul. 2018, doi: 10.1002/WIDM.1253.

[17] Y. R. Devi, A. Bharthepudi, and A. Govindarajula, "A Review on Sentiment Analysis Using Transformers and Ensemble methods," *6th IEEE Int. Conf. Recent Adv. Inf. Technol. RAIT 2025*, 2025, doi: 10.1109/RAIT65068.2025.11089332.

[18] E. Cambria, "Affective Computing and Sentiment Analysis," *IEEE Intell. Syst.*, vol. 31, no. 2, pp. 102–107, Mar. 2016, doi: 10.1109/MIS.2016.31.

[19] A. Almalaq and G. Edwards, "A review of deep learning methods applied on load forecasting," *Proc. - 16th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2017*, vol. 2017-December, pp. 511–516, 2017, doi: 10.1109/ICMLA.2017.0-110.

[20] S. A. Bryan Lim, "Temporal Fusion Transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecast.*, vol. 37, no. 4, pp. 1748–1764, 2021, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169207021000637

[21] Chi Sun, Luyao Huang, Xipeng Qiu, "Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence," *ACL Anthol.*, pp. 380–385, 2019, [Online]. Available: https://aclanthology.org/N19-1035/

[22] Hu Xu, Bing Liu, "BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis," *Proc. 2019 Conf. North*, 2019, [Online]. Available: https://aclanthology.org/N19-1242/

[23] Hashir Ali, Ehtesham Hashmi, "Analyzing Amazon Products Sentiment: A Comparative Study of Machine and Deep Learning, and Transformer-Based Techniques," *Electronics*, vol. 13, no. 7, p. 1305, 2024, doi: https://doi.org/10.3390/electronics13071305.

[24] Kasun Bandara, Christoph Bergmeir, "Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach," *Expert Syst. Appl.*, vol. 140, p. 112896, 2020, doi: https://doi.org/10.1016/j.eswa.2019.112896.