Consumer Review Sentiment Analysis and Market Trend Prediction Based on GPT-3

Ting Lu
Department of Digital Economy, Changzhou College of Information Technology, Changzhou 213164, Jiangsu, China
Corresponding author: Ting Lu (e-mail: luting158501@126.com)

ABSTRACT Existing market-share forecasting approaches often rely on simplified sentiment proxies that fail to capture nuanced semantics and long-range contextual dependencies in consumer reviews, limiting their predictive power in dynamic ecommerce environments. To address this gap, this paper proposes an integrated framework that uses GPT-3 for highfidelity sentiment time-series extraction and embeds these signals into a Temporal Fusion Transformer for forwardlooking market-share prediction in the textile sector. GPT-3 generates confidence-weighted monthly positivity and net sentiment scores, which are fused with 10 additional features covering static product attributes, historical review dynamics, and future-known inputs such as public holidays and major promotional events. Experimental results show that GPT-3 achieves a Macro-F1 of 0.842 in sentiment classification, outperforming RoBERTa and VADER, especially in neutral sentiment recognition, where F1 reaches 0.789. The TFT+GPT-3 model achieves MAE of 0.024, RMSE of 0.032, MAPE of 4.29%, and directional accuracy of 83.7% in market-share prediction. Ablation analysis shows that removing GPT-3-derived sentiment features increases MAE to 0.035, confirming their critical contribution to forecast accuracy and interpretability.

INDEX TERMS consumer reviews, sentiment analysis, market share prediction, GPT-3, temporal fusion transformer

I. INTRODUCTION

As the core carrier of user attitudes and experiences on e-commerce platforms, consumer reviews contain rich market leading signals, particularly for the textile and apparel sector where feedback on fit, fabric, and quality directly drives product development and inventory decisions [1,2]. As the number of reviews on platforms such as Amazon [3,4] exceeds 100 million, how to accurately extract predictive sentiment dynamics from them has become a key challenge for business intelligence. Traditional market prediction models mostly rely on lagged sales or rating data, making it difficult to capture sudden changes in consumer behavior driven by emotions. At the same time, e-commerce reviews [5,6] have the characteristics of cross-category, diverse language styles, and complex semantics, which lead to the limitations of general sentiment analysis methods in terms of accuracy and generalization. In recent years, the large language model GPT-3 (Generative Pre-trained Transformer) [7] has demonstrated powerful semantic understanding capabilities, providing a new paradigm for high-fidelity sentiment extraction. However, how to effectively convert the sentiment signals generated by GPT-3 [8] into structured time series features and embed them into advanced prediction frameworks to achieve dynamic and forward-looking modeling of key business indicators such as market share remains a scientific problem that needs to be solved urgently.

Research on sentiment analysis in e-commerce reviews has evolved from rule-based to deep learning [9,10]. Early methods such as VADER (Valence Aware Dictionary and sEntiment Reasoner) relied on sentiment dictionaries and heuristic rules, which are effective for short texts but had difficulty handling contextual dependencies and domain transfer [11]. Subsequently, supervised models based on LSTM (Long Short-Term Memory) [12,13] and CNN (Convolutional Neural Network) improved performance through end-to-end training, but are limited by annotation costs. Pre-trained language models have significantly promoted the development of this field: BERT (Bidirectional Encoder Representations from Transformers) [14,15] and its variants, after fine-tuning on datasets such as Amazon Reviews, achieved SOTA (State-of-the-Art) levels in fine-grained sentiment classification. Cross-domain adaptation models such as FinBERT further verified the value of domain knowledge injection [16]. However, these methods still require a large amount of labeled data and face generalization bottlenecks in cross-category scenarios. Recently, GPT-3 [17] has been applied to sentiment analysis tasks. With its large-scale pre-

training and contextual understanding capabilities, it has demonstrated superior performance in zero-shot or few-shot settings, particularly in identifying ambiguous expressions and neutral sentences. However, existing research has mostly focused on static classification accuracy, lacking systematic verification of the sentiment time series construction mechanism and its actual utility in downstream business predictions. In particular, empirical evidence in large-scale real-world e-commerce scenarios is still insufficient.

Market trend forecasting methods have gradually evolved from traditional time series models to deep learning architectures that integrate multiple sources of signals. Classic methods such as ARIMA (AutoRegressive Integrated Moving Average) and exponential smoothing rely solely on historical sales and are difficult to integrate external dynamic information [18,19]. To apply user behavior signals, researchers have begun to use UGC (User-Generated Content) as a leading indicator [20]: for example, predicting movie box office based on Twitter sentiment, or predicting restaurant customer flow based on Yelp reviews. Deep learning has further improved the integration capability: LSTM is used to jointly model the time series of reviews and sales dynamics; the Transformer architecture [21,22] captures long-term dependencies through the self-attention mechanism. Recently, TFT (Temporal Fusion Transformer) has become an advanced framework for complex business forecasting due to its ability to uniformly model static features, time series variables, and future known inputs [23,24]. Existing work has attempted to use basic sentiment scores as TFT inputs, preliminarily verifying the predictive value of sentiment signals. However, these studies generally use simple, low-dimensional sentiment indicators and fail to fully explore the deep semantic information in reviews [25]. More importantly, there is a lack of quantitative comparison of the marginal contributions of different sentiment extraction techniques in prediction tasks, nor is there a systematic investigation of the dynamic impact mechanism and leading warning capability of sentiment features on structural indicators such as market share [26]. In summary, existing research has not yet achieved effective synergy between high-fidelity sentiment time series and advanced prediction models. Addressing the limitations of existing research, this paper constructs a new market forecasting paradigm that integrates large language models with advanced time series architectures. First, it applies GPT-3 as a sentiment extraction engine to enhance the semantic understanding of cross-category e-commerce reviews, effectively identify complex expressions, and generate high-fidelity sentiment signals. Second, it designs a dynamic sentiment time series construction mechanism that converts GPT-3 outputs into structured indicators and embeds them into the TFT framework, enabling deep coupling modeling of sentiment dynamics and market share. Third, through an experimental system, it evaluates the impact of different sentiment extraction techniques on forecasting performance, validating the value of GPT-3 in improving forecast accuracy, directional judgment, and turning point warning. This paper validates the method's effectiveness on real Amazon review data, promoting the paradigm shift from "descriptive" to "predictive" consumer review analysis—a crucial step that offers textile enterprises a practical, large language model-driven technical path for actionable business intelligence.

II. METHOD

A. OVERVIEW OF THE OVERALL FRAMEWORK

To overcome the semantic understanding bottleneck of traditional sentiment analysis models in cross-category e-commerce reviews and achieve deep coupling of sentiment dynamics and market trend prediction, this paper proposes a hybrid framework that integrates a generative large language model with a time-series fusion Transformer. The core of this framework is to leverage the powerful semantic understanding capabilities of GPT-3 to generate high-fidelity sentiment time-series signals, which are then embedded as key dynamic features in the TFT model to accurately predict future market share. The overall architecture of this paper is shown in Figure 1.

Note: API (Application Programming Interface).

The architecture proposed in this paper is a prediction system with a clear hierarchy and rigorous information flow. It begins with rigorous preprocessing of large-scale raw review data, which not only ensures data quality but, more importantly, constructs the target variable for prediction—monthly market share, represented by the weighted number of reviews, providing a baseline ground truth for subsequent supervised learning. This paper applies GPT-3 as an independent sentiment signal extraction engine. Unlike directly inputting text into the prediction model, this framework utilizes GPT-3 through carefully designed prompt word engineering to transform unstructured review text into structured, confidence-weighted monthly sentiment time series indicators. This design cleverly separates the semantic understanding task from the time series prediction task, fully leveraging the semantic advantages of large models while avoiding the computational burden and overfitting risks associated with directly inputting high-dimensional, sparse text data into time series models, ensuring high-quality and interpretable sentiment features. During the feature fusion and prediction phase, the Temporal Fusion Transformer (TFT) is employed for its capability to holistically integrate static covariates, observed time-series inputs, and future known inputs. The multidimensional time series first passes through a variable selection network to adaptively prioritize informative features, enhancing robustness. An LSTM encoder-decoder then captures local temporal patterns, while a multi-head self-attention mechanism models long-range dependencies across all historical time steps. The resulting attention weights offer interpretable insights into the influence of past sentiment dynamics on predictions. The model outputs both point forecasts and quantile intervals, enabling actionable risk assessment and supporting a shift from static

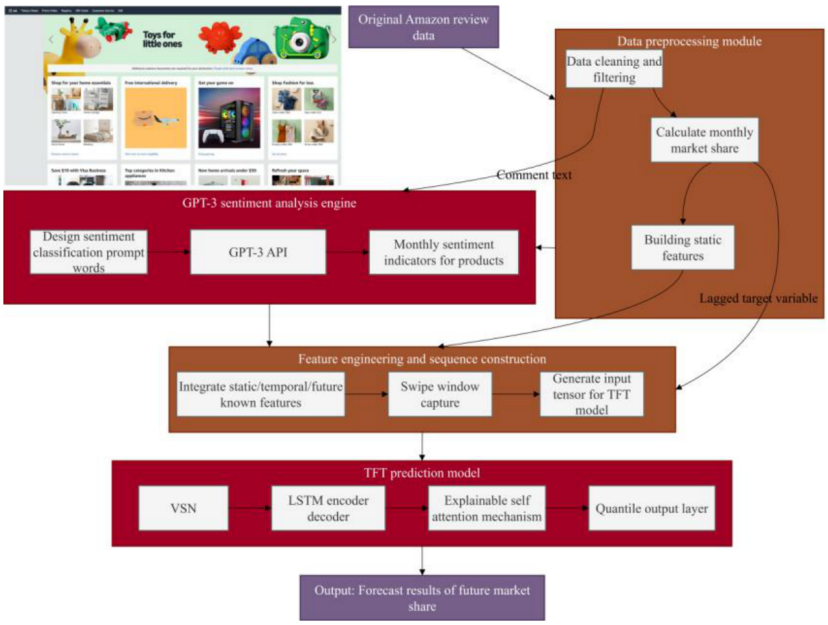

FIGURE 1. Overall structure of this paper

sentiment analysis to dynamic, interpretable market trend forecasting.

B. DATA SOURCES AND PREPROCESSING

1) Data Screening

To ensure the representativeness, temporal continuity, and predictive feasibility of the experimental data, this paper systematically screens and preprocesses the large-scale Amazon Reviews 2023 dataset [27]. The original data contains more than 571 million reviews and 48 million products (ASIN (Amazon Standard Identification Number)), spanning from May 1996 to September 2023. Given that early reviews are sparse and lack structured metadata, this paper focuses on recent high-quality reviews from January 2020 to September 2023 to ensure the timeliness of market dynamics and the effectiveness of sentiment signals.

Specific screening process: Validity filtering: Reviews with text lengths less than 10 characters (e.g., uninformative content like "OK" and "Good") are removed to ensure semantic integrity. Category focus: Eight core categories with high activity and competition are selected: Electronics, Beauty, Home & Kitchen, Sports, Toys, Books, Fashion, and Automotive. These categories cover both durable goods and fast-moving consumer goods, have rich user reviews, and significant market volatility, making them suitable for trend forecasting research. Product activity constraints: Only products meeting the following criteria are retained: cumulative reviews ≥ 100; review span ≥ 12 months (i.e., reviews in at least 12 different months); and sustained user engagement within the window of January 2020–September 2023.

The determination of core categories is based on the following quantitative criteria: first, calculate the average monthly comments of each category in the past 12 months, and screen the top 25% categories to ensure data activity; Secondly, the Herfindahl index is used to measure the brand concentration within the category, and the category with an index lower than 0.15 is selected to ensure full competitiveness; Finally, combined with the industry research report, it is ensured that the selected eight categories cover the two major types of durable consumer goods and fast-moving consumer goods, with the characteristics of significant market fluctuations and mature user feedback mechanism, which can effectively support the trend prediction research.

After the above screening, a high-quality subset is constructed, consisting of 5,218 ASINs [28,29], corresponding to 42.7 million valid reviews, and covering the complete time series trajectory from January 2020 to September 2023. The statistics of the filtered dataset are shown in Table 1.

TABLE 1. Statistics of the filtered dataset

Items	Value
Number of products	5,218
Total number of reviews	42.7 million
Time span	2020.01–2023.09
Average reviews/product/month	182.3
Number of categories	8

2) Market Share Calculation

To support market share calculations, this paper uses the widely validated weighted review count proxy sales method. This approach is rationale for this: highly rated reviews are more likely to indicate satisfactory purchases, and reviews voted "useful" are more representative. The proxy sales for product i in month t are defined as:

$$Sales_{i,t} = \sum_{r \in R_{i,t}} (rating_r \times (1 + helpful\_votes_r)) \quad (1)$$

In formula (1), Ri,t is the set of all reviews for product i in month t; ratingr is the star rating of review r; and helpful_votesr is the number of votes other users have given that review as "helpful."

The monthly market share of product i in its category C(i) is calculated as follows:

$$MS_{i,t} = \frac{Sales_{i,t}}{\sum_{j \in C(i)} Sales_{j,t}} \tag{2}$$

This indicator is normalized to the interval [0, 1] to reflect the relative competitive position of the product in the market segment and serves as the core target variable of the prediction task in this paper.

C. CONSTRUCTION OF COMMENT-BASED SENTIMENT TIME SERIES FEATURES

1) Sentiment Classification and Confidence Weighting Based on GPT-3

To extract high-fidelity sentiment signals from massive consumer reviews, this paper constructs a sentiment time series extraction framework based on GPT-3 [30,31] and applies multiple mainstream sentiment analysis models as comparison baselines to systematically evaluate the performance differences of different technical paths in cross-category e-commerce review scenarios.

This paper uses the text-davinci-003 model which is the most powerful in gpt-3 series as the backbone network of emotion classification. The model has 175billion parameters and shows the best semantic understanding ability in zero sample and small sample tasks. In order to ensure the consistency of output and the reproducibility of results, all reasoning uses deterministic settings, and logprobs is enabled to obtain classification confidence. Text-davinci-003 is selected instead of the fine-tuning version, which aims to verify the migration ability of the general large-scale model in the emotional analysis of cross category e-commerce reviews, and facilitate other researchers to replicate the experiment based on the same model.

This paper uses GPT-3 (text-davinci-003) as the backbone model for sentiment classification. To ensure output consistency and avoid generating noise, it uses deterministic inference settings: temperature is set to 0, maximum generation length (max_tokens) is set to 5, and log probabilities (logprobs) are enabled to obtain the model's confidence in each candidate label.

For each review text rk, construct the following standardized prompt:

> Classify the sentiment of the following Amazon product review into exactly one of: Positive, Negative, Neutral.
> Review: "{text}"
> Sentiment:

GPT-3 returns the predicted label yk∈{+,-,0} (corresponding to Positive, Negative, and Neutral, respectively) and the log probability of the corresponding token. To quantify the reliability of the model's decision, the classification confidence is defined as the difference between the top-1 and top-2 log probabilities:

$$c_k = \log p\left(y_k^{(1)}\right) - \log p\left(y_k^{(2)}\right) \tag{3}$$

In formula (3), $y_k^{(1)}$ and $y_k^{(2)}$ are the two candidate labels with the highest probability. This metric effectively reflects the model's discriminative certainty on edge cases, with larger values indicating higher confidence.

For the set Ri,t of all reviews for product i in month t, two core sentiment time series metrics are constructed:

The formula for the confidence-weighted positive rate is:

$$P_{i,t} = \frac{\sum_{k \in R_{i,t}} I(y_k = +) \cdot c_k}{\sum_{k \in R_{i,t}} c_k} \tag{4}$$

In formula (4), I is an indicator function. This indicator reduces the interference of low-quality or vague reviews on the overall sentiment estimation by confidence weighting.

The formula for the confidence-weighted sentiment score is:

$$S_{i,t} = \frac{\sum_{k \in R_{i,t}} [I(y_k = +) - I(y_k = -)] \cdot c_k}{\sum_{k \in R_{i,t}} c_k} \in [-1, 1] \tag{5}$$

In order to construct high-quality emotional time series from gpt-3 output, this paper uses the confidence weighting mechanism based on the internal probability distribution of the model. For each comment, gpt-3 returns the log probability corresponding to three candidate tags (positive, negative, neutral). The confidence score is defined as the difference between the log probabilities of the highest probability tag and the second highest probability tag. The design is derived from the measurement idea of "decision boundary clarity" in information theory - when the model judges the two categories are close (i.e. the confidence is low), it indicates that the comment semantic is fuzzy or ambiguous, and the direct equal weight average will introduce noise. In contrast, the average implicit assumption of simple emotion score is that all predictions are equally reliable, ignoring the uncertainty of the model itself.

2) Comment Sentiment Classification Confusion Matrix

To comprehensively evaluate the performance advantages of GPT-3, this paper deploys four representative sentiment analysis models on the same subset of reviews to ensure fair comparison. This paper's GPT-3 is compared with the other three models. VADER, a lexicon-based rule-based model, outputs a continuous score between -1 and 1, with scores > 1/3 classified as positive, < -1/3 as negative, and scores in between as neutral. FinBERT, a BERT fine-tuned on financial text, is adapted for three-category classification. RoBERTa (Robustly Optimized BERT Approach), a base model fine-tuned on Amazon Reviews 2018 (supervised training), is used.

50,000 reviews are randomly sampled from the total review set and manually annotated by three human annotators (Krippendorff's α = 0.82). The heatmap of the review sentiment classification is shown in Figure 2.

As shown in the confusion matrix, GPT-3 outperforms baseline models in all three categories of sentiment recognition, with particularly strong performance in neutral review classification. Its classification accuracy for neutral examples (14,300/15,000) significantly surpasses RoBERTa, FinBERT, and VADER, demonstrating strong semantic discrimination for neutral statements (such as "well-packaged" and "appropriate size"). Furthermore, GPT-3 correctly classifies 17,400 positive and 15,200 negative examples, respectively, significantly reducing cross-class confusion. In contrast, VADER, due to its reliance on local sentiment terms, misclassifies a significant number of neutral reviews as negative (3,021 examples). Despite fine-tuning, FinBERT and RoBERTa are still limited by ambiguity in the labeling of neutral examples in their training data. Leveraging its large-scale pre-training and contextual reasoning capabilities, GPT-3 achieves high-fidelity sentiment analysis of complex e-commerce reviews, providing more reliable sentiment signals for downstream prediction tasks.

D. MARKET SHARE PREDICTION MODEL BASED ON TFT

To achieve high-precision dynamic prediction of commodity market share, this paper adopts TFT as the core prediction architecture. TFT [32] is a deep learning model designed for multivariate, multi-scale time series prediction. It can effectively integrate static metadata, historical time series observations, and future known covariates, and reveal key driving factors through an interpretable attention mechanism. This paper implements TFT based on the PyTorch Forecasting library and customizes the input feature system for the e-commerce market share prediction task.

TFT input is divided into three categories: static features, time series observation features, and future known inputs. The feature system designed in this paper is as follows.

Static features describe the inherent attributes of a product and remain constant within the prediction window. Category (category): An 8-dimensional one-hot vector representing the product's category of high activity; Price range (price_bin): Five discrete price levels, encoded using one-hot encoding; Brand popularity (brand_rank): A continuous variable defined as the logarithmic ranking of the brand's cumulative historical reviews within the category (larger values indicate greater popularity).

Time series observation features are only available in the historical window (t-L,...,t) and reflect the dynamic performance of products.

$P_{i,t}$ is the confidence-weighted positive rate of product i in month t; $S_{i,t}$ is the confidence-weighted sentiment score of product i in month t ($\in [-1, 1]$).

$review\_count_{i,t}$: Total number of monthly reviews.

$avg\_rating_{i,t}$: Average monthly star rating.

$helpful\_ratio_{i,t} = \frac{\sum helpful\_votes_r}{review\_count_{i,t}}$: Percentage of helpful votes.

$MS_{i,t-1}$, $MS_{i,t-2}$: Market share for the previous month and the previous two months, used to capture short-term inertia.

Future known inputs are future event variables known at the time of prediction (time $t+1$):

$is\_holiday_{t+1} \in \{0, 1\}$ represents whether it is a public holiday.

$is\_primeday_{t+1} \in \{0, 1\}$ represents whether it is a major promotional day such as Amazon Prime Day.

The TFT architecture uses a VSN (Variable Selection Network) to apply gating mechanisms to both static and temporal features, automatically learning the importance weights of each feature at different time steps. For the temporal feature $x_{i,t} \in \mathbb{R}^{d_x}$, the VSN outputs a weighted representation:

$$\tilde{x}_{i,t} = \sum_{j=1}^{d_x} \alpha_{j,t} \cdot GRN_j \left(x_{i,t}^{(j)} \right) \tag{6}$$

In formula (6), $\alpha_{j,t}$ is a learnable weight, and GRN_j is a gated residual network.

In the LSTM encoder-decoder, the encoder processes the VSN output from the historical window t-L+1 to t; the decoder combines future known inputs (such as holidays) with the encoder state to generate future representations. During the decoding phase, the interpretable multi-head attention mechanism quantifies the contribution of historical time step t' to the prediction of t+1 using the attention weight at',t:

$$Attention(Q, K, V) = softmax \left(\frac{QK^\top}{\sqrt{d_k}} \right) V \tag{7}$$

The output layer uses a quantile regression head to output both point predictions (median) and prediction intervals to quantify uncertainty:

$$\widehat{MS}_{i,t+1}^{(\tau)} = f_{TFT}(\tau; \Theta), \tau \in \{0.1, 0.5, 0.9\} \tag{8}$$

In formula (8), Θ represents the model parameter and τ represents the quantile level.

This paper embeds the sentiment features (Pi,t, Si,t) generated by GPT-3 as the core time series input into TFT. Through end-to-end training, the model automatically learns the nonlinear mapping relationship between sentiment dynamics and market share evolution, thereby achieving high-precision, interpretable forward-looking predictions. The list of TFT input features is shown in Table 2.

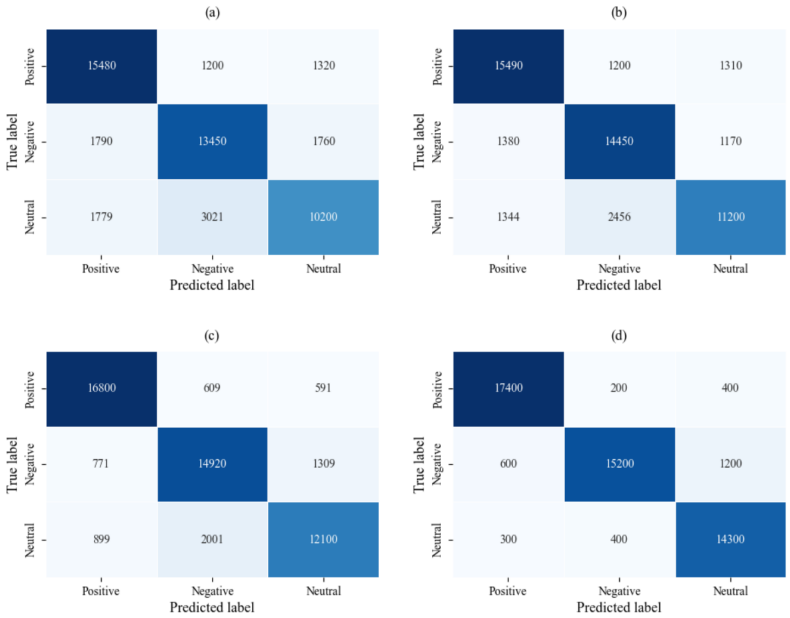

FIGURE 2. Comment Sentiment Classification Heatmap; Figure 2(a) VADER Model; Figure 2(b) FinBERT Model; Figure 2(c) RoBERTa-Base Model; Figure 2(d) GPT-3 Model

TABLE 2. TFT input characteristics list

Feature	Type	Dimensions	Description
category	Static	8	One-hot encoding of the product's category (8 highly active categories)
price_bin	Static	5	Five discrete price ranges based on price quantiles within the category (one-hot encoding)
brand_rank	Static	1	Brand popularity, defined as the logarithmic ranking of the brand's cumulative historical reviews within the category (continuous variable)
$P_{i,t}$	Time-varying	1	Monthly confidence-weighted positivity rate, generated by the sentiment model
$S_{i,t}$	Time-varying	1	Monthly confidence-weighted sentiment score, output by the sentiment model
review_count	Time-varying	1	Monthly total number of reviews
avg_rating	Time-varying	1	Monthly average star rating (1–5)
helpful_ratio	Time-varying	1	Monthly percentage of helpful votes
$MS_{i,t-1}$	Time-varying	1	Previous month's market share
$MS_{i,t-2}$	Time-varying	1	Previous month's market share
is_holiday	Known future input	1	Month's market share
is_prime_day	Known future input	1	Whether the upcoming month is a public holiday (0/1)

III. EXPERIMENT

A. DATA PARTITIONING

To ensure the reliability and practical application value of the experimental results, this paper strictly adheres to the best practices of time series forecasting and adopts a non-overlapping, chronological data segmentation strategy to prevent any future information leakage. The entire observation window (January 2020 to September 2023) is divided into three mutually exclusive periods:

Training set: January 2020 to June 2022 (30 months), used for model parameter learning;

Validation set: July 2022 to December 2022 (6 months), used for hyperparameter tuning and early stopping;

Test set: January 2023 to September 2023 (9 months), used for final performance evaluation.

This partitioning ensures that the model is trained and validated solely on historical data, while the testing phase fully simulates real-world business scenarios—that is, using historical information up to the current month to predict market performance for the next month.

This paper defines a single-step-ahead forecasting task: for any target month t+1, the model uses the complete feature sequence from the past L = 12 months (i.e., t-11 to t) as input and predicts the market share MSi,t+1 of product i in month t+1. This setup ensures sufficient historical context to capture seasonality and trends while avoiding the noise and computational burden applied by overly long windows. All features (including static attributes, time-series observations, and future known events) are strictly time-aligned. Future known variables are only available at the decoding stage (i.e., time t+1), consistent with the practice of knowing promotional calendars in advance.

B. EVALUATION METRICS AND MODEL COMPARISON

1) Evaluation Metrics

To comprehensively evaluate the performance of sentiment classification [33,34] and market share prediction, this paper adopts a multi-dimensional, task-adapted evaluation metric system that takes into account classification accuracy, regression error, and relevance to business decisions.

Sentiment classification [35,36] is a three-class imbalanced classification problem (Positive, Negative, Neutral). Therefore, in addition to overall accuracy, the Macro-F1 score, which is robust to class imbalance, is used.

Accuracy:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} I(y_i = \hat{y}_i) \tag{9}$$

In formula (9), N is the total number of samples, y_i and $\hat{y}_i$ are the true and predicted labels, respectively. Macro-F1:

$$Macro\text{-}F1 = \frac{1}{3}(F1_{Pos} + F1_{Neg} + F1_{Neu}) \tag{10}$$

For each category c, its F1 is expressed as:

$$F1_c = \frac{2 \cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c} \tag{11}$$

Market share prediction is a regression task, and the formula for MAE (Mean Absolute Error) is:

$$MAE = \frac{1}{M}\sum_{m=1}^{M}\left|MS_m - \widehat{MS}_m\right| \tag{12}$$

The formula for RMSE (Root Mean Squared Error) is:

$$RMSE = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(MS_m - \widehat{MS}_m)^2} \tag{13}$$

The formula for MAPE (Mean Absolute Percentage Error) is:

$$MAPE = \frac{100\%}{M}\sum_{m=1}^{M}\left|\frac{MS_m - \widehat{MS}_m}{MS_m + \epsilon}\right| \tag{14}$$

Since business decisions need to focus on the direction of market trends (such as rising or falling market share), this paper applies DA (Directional Accuracy):

$$DA = \frac{1}{M}\sum_{m=1}^{M} I(sign(\Delta MS_m) = sign(\Delta\widehat{MS}_m)) \tag{15}$$

In formula (15), ΔMS_m is the actual share change, $\Delta\widehat{MS}_m$ is the predicted change, and $sign(\cdot)$ is the sign function. DA measures the accuracy of the model's judgment of market trends and has direct guiding significance for scenarios such as inventory allocation and marketing response.

2) Comparative Model Design

To systematically evaluate the effectiveness of the proposed method, it analyzes sentiment classification and market share prediction performance. All models are trained and tested on the same data subset and using a unified evaluation protocol to ensure comparability of results. To validate GPT-3's superiority in understanding sentiment in cross-category e-commerce reviews, it selects three representative sentiment analysis methods as baselines: VADER, FinBERT, and RoBERTa-base.

To analyze market share prediction performance, it compares ARIMA, LSTM + VADER, LSTM + RoBERTa, TFT + VADER, TFT + RoBERTa, and TFT + GPT-3 (our proposed model).

IV. RESULTS AND ANALYSIS

A. COMPARATIVE ANALYSIS OF SENTIMENT CLASSIFICATION RESULTS

The performance of the sentiment classification model on the test set is shown in Figure 3.

GPT-3 comprehensively outperforms all baseline models in sentiment classification, achieving state-of-the-art Macro-F1 (0.842) and Accuracy (0.849). It particularly excels in the neutral category (F1=0.789), surpassing RoBERTa (0.749), FinBERT (0.658), and VADER (0.515). This advantage stems from GPT-3's powerful contextual semantic modeling capabilities, enabling it to accurately identify common, non-sentimental statements found in e-commerce reviews (such as "suitable size" and "well-packaged"), as well as complex expressions with implicit neutrality. Dictionary models like VADER are susceptible to interference from local sentiment words, resulting in a large number of neutral examples being misclassified as negative. Despite fine-tuning, FinBERT and RoBERTa are still limited by the noisy annotations and semantic ambiguity of neutral examples in their training data, resulting in insufficient generalization. Furthermore, GPT-3 achieves the best F1 scores for the positive and negative categories (0.873 and 0.864), demonstrating that it effectively manages cross-class confusion while maintaining high recall.

B. GPT-3 SENTIMENT ANALYSIS ADVANTAGES

The performance-cost trade-off and the accuracy performance of each type of emotion are shown in Figure 4.

Figure 4(a) reveals significant differences in the performance-cost trade-off among sentiment models. Although GPT-3 has the highest inference cost (approximately \$10-2 per review), its Macro-F1 (0.842) and Accuracy (0.849) both rank first, placing it on the Pareto frontier. In contrast, RoBERTa achieves suboptimal performance (Macro-F1 = 0.811) at a lower cost, demonstrating excellent value for money. VADER has negligible cost, but its performance lags significantly behind (Macro-F1 = 0.640). Notably, GPT-3's improved Accuracy (+0.028 vs. RoBERTa) comes at the expense of inference cost, which is acceptable in high-value prediction scenarios.

Figure 4(b) further reveals the differences in the models' discriminative abilities across fine-grained categories. GPT-3 achieves an accuracy of 0.921 in the Positive category, significantly exceeding other models and demonstrating its strong ability to discern explicit praise (e.g., "excellent quality"). On the Negative category, GPT-3 (0.825) also outperforms RoBERTa (0.801), demonstrating its ability to effectively capture implicit dissatisfaction (e.g., "not as described"). Crucially, GPT-3 achieves an accuracy of 0.791 on the Neutral category, surpassing both RoBERTa and FinBERT, demonstrating its ability to accurately discriminate neutral statements. In contrast, VADER performs worst on the Neutral category due to its reliance on local sentiment terms, which can easily misclassify neutral sentences as negative. GPT-3's contextual understanding mechanism ef-

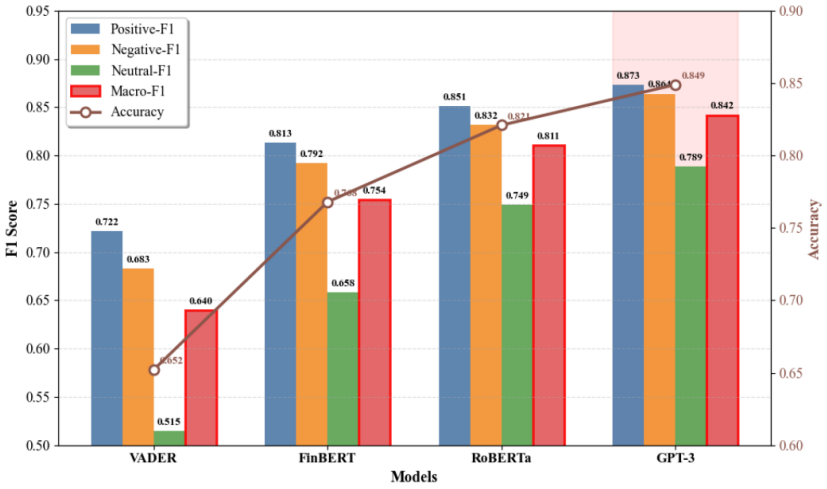

FIGURE 3. Sentiment classification model

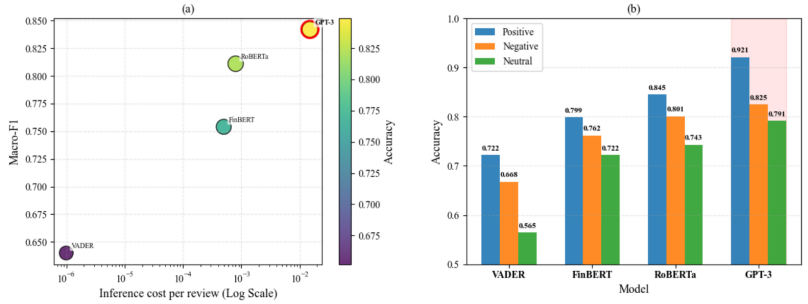

FIGURE 4. Performance-cost tradeoff and accuracy of various emotions; Figure 4(a) Performance-cost tradeoff; Figure 4(b) Accuracy of various emotions

fectively alleviates the semantic ambiguity in e-commerce reviews, providing a purer sentiment signal for downstream prediction tasks.

The ROC (Receiver Operating Characteristic) curve can be used to more intuitively judge the effect of neutral sentiment classification. The ROC comparison of each model is shown in Figure 5.

The ROC curve and AUC (Area Under the Curve) values shown in Figure 5 clearly demonstrate that GPT-3 outperforms all baseline models in the neutral sentiment classification task (AUC = 0.803), achieving improvements of 0.056, 0.079, and 0.191 over RoBERTa (0.747), FinBERT (0.724), and VADER (0.612), respectively. This advantage stems from GPT-3's powerful contextual semantic modeling capabilities, which effectively identify a large number of neutral statements in e-commerce reviews without explicit sentiment terms, avoiding misclassification as negative. In contrast, VADER relies on a local sentiment lexicon and has low sensitivity to neutral text, resulting in a high false positive rate. Despite fine-tuning, FinBERT and RoBERTa are still limited by the labeling noise and fuzzy semantic boundaries of neutral examples in the training data, making it difficult to calibrate the discrimination threshold when generalizing across categories.

It should be noted that although gpt-3 performs well in emotion analysis, its computational cost does pose a challenge for industrial scale applications. In the actual business scenario, it is recommended to adopt a hierarchical strategy: for conventional demand forecasting, models with higher cost performance such as Roberta can be used; For key categories, new products or high-risk products, gpt-3 is used for in-depth analysis. This hybrid scheme can not only ensure the prediction accuracy of the core business, but also control the total cost within a reasonable range.

C. MARKET SHARE PREDICTION PERFORMANCE EVALUATION

The prediction performance of the market share prediction model on the test set is shown in Figure 6.

The TFT + GPT-3 model comprehensively outperforms other comparable models in the market share prediction task, achieving the lowest MAE (0.024), RMSE (0.032), and MAPE (4.29%), with a directional accuracy (DA) of 83.7%, surpassing the next-best model, TFT + RoBERTa (MAE = 0.026, DA = 79.8%). This performance gain is primarily due to the high-fidelity sentiment time series

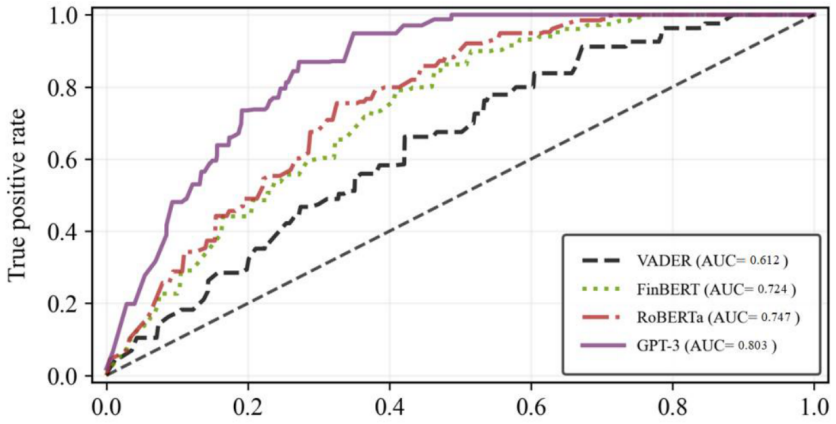

FIGURE 5. ROC curve comparison

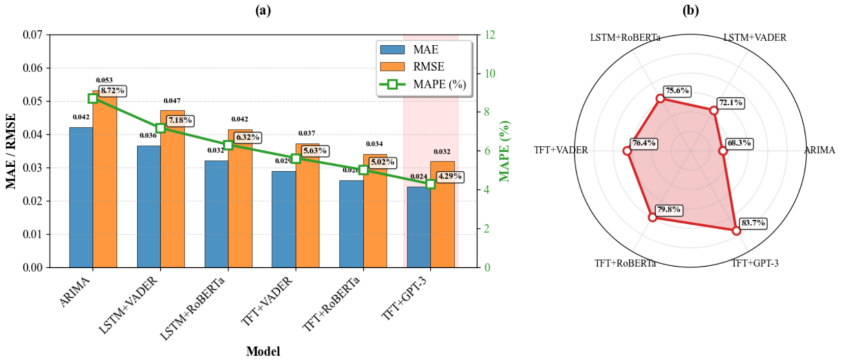

FIGURE 6. Market Share Prediction Performance; Figure 6(a) Error Analysis; Figure 6(b) Direction Accuracy

signals provided by GPT-3: its precise recognition of neutral and complex semantic reviews effectively reduces sentiment feature noise, enabling the model to more accurately capture emotion-driven consumer behavior changes. Furthermore, the TFT architecture, through a variable selection network and interpretable attention mechanism, adaptively weights GPT-3 sentiment features, strengthening the modeling of the correlation between sentiment dynamics and market share evolution. In contrast, LSTM architectures generally perform worse than TFT due to their lack of explicit modeling of static features and future variables. VADER, however, suffers from its crude sentiment discrimination, limiting its performance even when embedded in TFT. The results demonstrate that the coordinated optimization of high-quality sentiment features and advanced time series architectures is key to improving market forecast accuracy. GPT-3, as a sentiment extraction engine, provides leading indicators with significant marginal value for business forecasting.

Using TFT's interpretable attention mechanism, it visualizes the contribution of each feature at each historical time step to the market share forecast for July 2023, revealing the dominant role of sentiment signals in key forecast windows. The results are shown in Figure 7.

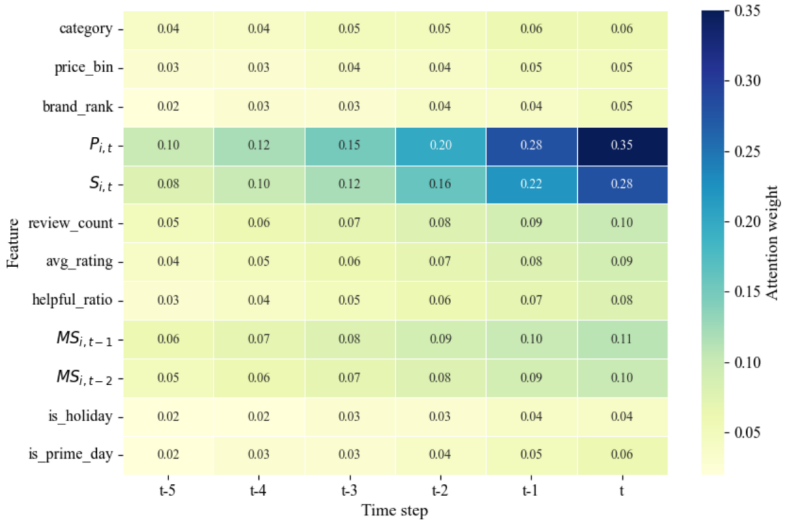

FIGURE 7. TFT attention weight heat map

The attention heatmap in Figure 7 reveals the dynamic dependency patterns of various features used by the TFT model in predicting market share for July 2023. The GPT-3-generated sentiment features Pi,t and Si,t receive the highest attention weights near time steps (t-1, t), indicating that recent consumer sentiment is the core signal driving predictions. The static feature weights are stable and low, indicating that the model primarily relies on dynamic behavior rather than inherent attributes. This result not only validates the effectiveness of GPT-3 sentiment features but also demonstrates the high interpretability achieved by TFT through the attention mechanism, providing intuitive evidence for the business side that sentiment leads sales.

D. ABLATION EXPERIMENT

An ablation experiment is conducted to analyze the impact of removing sentiment features on market share prediction performance. The results are shown in Table 3.

TABLE 3. Ablation experiment results

Model	MAE	RMSE	MAPE (%)	DA (%)
Full	0.024	0.032	4.29	83.7
w/o $P_{i,t}$	0.028	0.039	5.23	81.2
w/o $S_{i,t}$	0.029	0.037	5.36	80.8
w/o $P_{i,t}$ and $S_{i,t}$	0.035	0.052	6.32	75.3

Note: $P_{i,t}$ is the confidence-weighted positive rate of product i in month t; $S_{i,t}$ is the confidence-weighted sentiment score of product i in month t; w/o means without.

Ablation experiment results clearly reveal the critical contribution of sentiment features to market share prediction. The full model (Full) achieves optimal performance (MAE = 0.024, DA = 83.7%) after applying the confidence-weighted positivity Pi,t and sentiment score Si,t. Removing either sentiment metric individually degraded performance: removing Pi,t increases the MAE to 0.028, while removing Si,t increases it to 0.029, demonstrating that they provide complementary information—Pi,t captures the intensity of positive sentiment, while Si,t reflects net sentiment polarity. Removing both simultaneously increases the MAE to 0.035 and decreases the DA to 75.3%, degrading performance to levels approaching those of LSTM + RoBERTa, confirming that sentiment signals are the core driver of improved prediction accuracy. Removing sentiment features results in a drop in directional accuracy of over 8 percentage points (from 83.7% to 75.3%), demonstrating that consumer sentiment not only influences absolute market share but also dominates its short-term fluctuations.

E. CASE ANALYSIS

A typical case study demonstrates GPT-3's emotional signal's ability to provide early warning of market share fluctuations, visually verifying the predictive value of emotion as a leading indicator in real business scenarios. The results are shown in Figure 8.

Figure 8, using a Bluetooth headset as an example, vividly demonstrates GPT-3's sentiment signals' ability to proactively predict market fluctuations. In March 2023, GPT-3 detects a sharp drop in the positive rating from 0.84 to 0.70 in April, issuing a warning approximately two weeks in advance of the April market share decline (from 0.15 to 0.13, a 13.3% drop). This lag effect aligns with consumer behavior: negative reviews first spread on social platforms, then influence potential purchase decisions, ultimately reflecting sales share. GPT-3, with its sensitivity to semantic details, successfully captured this sentiment inflection point. This case study not only validates the effectiveness of sentiment as a leading indicator but also highlights the strength of large language models in identifying subtle negative signals, providing companies with a valuable window for intervention—for example, initiating customer service

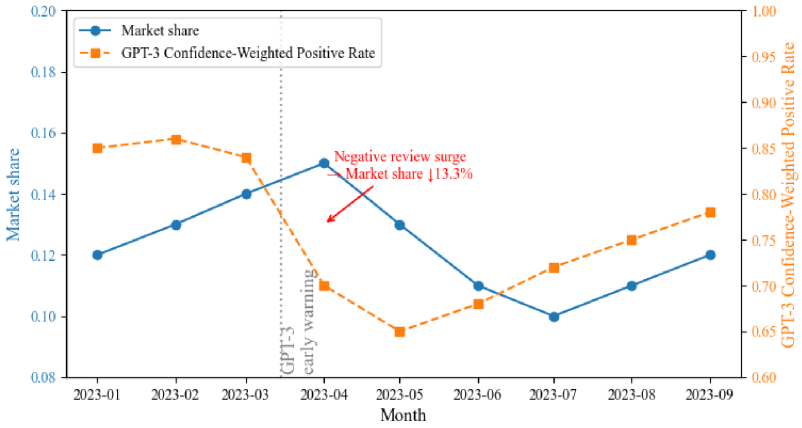

FIGURE 8. Bluetooth headset market share and GPT-3 positive rate time series

responses or product improvements before a substantial decline in market share occurs, thereby mitigating market losses.

V. CONCLUSION

This study establishes a novel and effective framework that synergistically integrates GPT-3–derived high-fidelity sentiment time series with the Temporal Fusion Transformer (TFT) to enable accurate, interpretable, and forward-looking market share prediction in the textile and apparel e-commerce domain. By leveraging GPT-3's superior contextual understanding—particularly its state-of-the-art performance in neutral sentiment recognition (F1 = 0.789) and overall Macro-F1 of 0.842—the model generates confidence-weighted sentiment indicators that significantly enhance predictive power. The TFT+GPT-3 architecture achieves a MAE of 0.024, RMSE of 0.032, MAPE of 4.29%, and directional accuracy of 83.7%, with ablation experiments confirming that sentiment features alone account for a >30% reduction in prediction error. This work not only quantifies the leading predictive value of nuanced consumer sentiment in real-world demand forecasting but also provides a reproducible, large language model–driven paradigm that bridges descriptive review analysis and actionable business intelligence, empowering textile manufacturers to dynamically align production, inventory, and design strategies with evolving consumer preferences. In this study, the full use of gpt-3 aims to verify the marginal value of its emotional signals. The actual industrial deployment can adopt the sampling set analysis strategy. Future research can explore low-cost alternatives such as knowledge distillation and model fine-tuning to reduce the application threshold while maintaining performance.

AUTHOR CONTRIBUTIONS

Ting Lu designed, collected and analyzed the data, and drafted the manuscript. Ting Lu conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Ting Lu participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This research was supported by, 2023 Jiangsu Province higher vocational college teacher professional leader research funding project,ID: 2023GRFX009.

Jiangsu Province Higher Education Teaching Reform Research Project in 2025: Innovation and Practice of Industrial Colleges in Digital Talent Cultivation under the Background of "New Double High" (Project No. 2025JGYB049).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] Z. Pu, C. Zhang, and Xu Zi, "A review of consumer preference mining based on online reviews," Chinese Journal of Management Science, vol. 33, no. 1, pp. 209-220, 2025, doi: 10.16381/j.cnki.issn1003-207x.2024.0483.

[2] C. Shen, S. Liu, and T. Xu, "Research on consumer online shopping review behavior under merchant-induced reviews," Journal of Information Resources Management, vol. 10, no. 3, pp. 92-101, 2020, doi: CNKI:SUN:XNZY.0.2020-03-010.

[3] N. M. Alharbi, N. S. Alghamdi, E. H. Alkhammash, and J. F. Al Amri, "Evaluation of sentiment analysis via word embedding and RNN variants for Amazon online reviews," Mathematical Problems in Engineering, vol. 2021, no. 1, Art. no. 5536560, 2021, doi: 10.1155/2021/5536560.

[4] A. Vollero, D. Sardanelli, and A. Siano, "Exploring the role of the Amazon effect on customer expectations: An analysis of user-generated content in consumer electronics retailing," Journal of Consumer Behaviour, vol. 22, no. 5, pp. 1062-1073, 2023, doi: 10.1002/cb.1969.

[5] P. M. Chu, Y. S. Mao, S. J. Lee, and C. L. Hou, "Leveraging user comments for recommendation in E-commerce," Applied Sciences, vol. 10, no. 7, pp. 2540, 2020, doi: 10.3390/app10072540.

[6] T. Chen, C. Tong, Y. Bai, J. Yang, G. Cong, T. Cong, et al., "Analysis of the public opinion evolution on the normative policies for the live streaming e-commerce industry based on online comment mining under COVID-19 epidemic in China," Mathematics, vol. 10, no. 18, pp. 3387, 2022, doi: 10.1145/2808797.2809334.

[7] C. Fang, Y. Jin, G. Chen, Y. Zhang, S. Li, Y. Ma, et al., "Multimodal Speech Emotion Recognition Based on Large Language Model," IEICE TRANSACTIONS on Information and Systems, vol. 107, no. 11, pp. 1463-1467, 2024, doi: 10.1587/transinf.2024EDL8034.

[8] E. H. Nfaoui and H. Elfaik, "Evaluating Arabic emotion recognition task using ChatGpt models: a comparative analysis between emotional stimuli prompt, fine-tuning, and in-context learning," Journal of Theoretical and Applied Electronic Commerce Research, vol. 19, no. 2, pp. 1118-1141, 2024, doi: 10.3390/jtaer19020058.

[9] M. E. Alzahrani, T. H. H. Aldhyani, S. N. Alsubari, M. M. Althobaiti, and A. Fahad, "Developing an Intelligent System with Deep Learning Algorithms for Sentiment Analysis of E-Commerce Product Reviews," Computational Intelligence and Neuroscience, vol. 2022, no. 1, Art. no. 3840071, 2022, doi: 10.1155/2022/3840071.

[10] U. Singh, A. Saraswat, H. K. Azad, K. Abhishek, and S. Shitharth, "Towards improving e-commerce customer review analysis for sentiment detection," Scientific Reports, vol. 12, no. 1, Art. no. 21983, 2022, doi: 10.1038/s41598-022-26432-3.

[11] R. Hernández-Pérez, P. Lara-Martínez, B. Obregón-Quintana, L. S. Liebovitch, and L. Guzmán-Vargas, "Correlations and fractality in sentence-level sentiment analysis based on Vader for literary texts," Information, vol. 15, no. 11, pp. 698, 2024, doi: 10.3390/info15110698.

[12] F. Huang, X. Li, C. Yuan, S. Zhang, J. Zhang, S. Qiao, et al., "Attention-emotion-enhanced convolutional LSTM for sentiment analysis," IEEE transactions on neural networks and learning systems, vol. 33, no. 9, pp. -4345, 2021, doi: 10.1109/TNNLS.2021.3056664.

[13] F. Ullah, X. Chen, S. B. H. Shah, S. Mahfoudh, M. A. Hassan, N. Saeed, et al., "A novel approach for emotion detection and sentiment analysis for low resource Urdu language based on CNN-LSTM," Electronics, vol. 11, no. 24, pp. 4096, 2022, doi: 10.3390/electronics11244096.

[14] M. K. Shaik Vadla, M. A. Suresh, and V. K. Viswanathan, "Enhancing product design through AI-driven sentiment analysis of Amazon reviews using BERT," Algorithms, vol. 17, no. 2, pp. 59, 2024, doi: 10.3390/a17020059.

[15] M. Mostafa and A. AlSaeed, "Sentiment analysis based on BERT for Amazon reviewer," Journal of the ACS Advances in Computer Science, vol. 13, no. 1, pp. 1-10, 2022, doi: 10.21608/asc.2022.283539.

[16] J. Kim, H. S. Kim, and S. Y. Choi, "Forecasting the S&P 500 index using mathematical-based sentiment analysis and deep learning models: A FinBERT transformer model and LSTM," Axioms, vol. 12, no. 9, pp. 835, 2023, doi: 10.3390/axioms12090835.

[17] C. Suhaeni and H. S. Yong, "Mitigating class imbalance in sentiment analysis through GPT-3-generated synthetic sentences," Applied Sciences, vol. 13, no. 17, pp. 9766, 2023, doi: 10.3390/app13179766.

[18] B. Dhyani, M. Kumar, P. Verma, and A. Jain, "Stock market forecasting technique using arima model," International Journal of Recent Technology and Engineering, vol. 8, no. 6, pp. 2694-2697, 2020, doi: 10.35940/ijrte.F8405.038620.

[19] A. M. Wahid, "Time series analysis of bitcoin prices using ARIMA and LSTM for trend prediction," Journal of Digital Market and Digital Currency, vol. 1, no. 1, pp. 84-102, 2024, doi: 10.47738/jdmdc.v1i1.1.

[20] L. Liu and B. Ma, "CA-VAR-Markov model of user needs prediction based on user generated content," Scientific Reports, vol. 15, no. 1, pp. 7716, 2025, doi: 10.1038/s41598-025-92173-8.

[21] T. Alsaedi, M. R. R. Rana, A. Nawaz, A. Raza, and A. Alahmadi, "Sentiment Mining in E-Commerce: The Transformer-based Deep Learning Model," International Journal of Electrical and Computer Engineering Systems, vol. 15, no. 8, pp. 641-650, 2024, doi: 10.32985/ijeces.15.8.2.

[22] K. Kadarsih and D. Pujianto, "Application of Transformer Model and Word Embedding in Sentiment Analisys of Indonesian E-Commerce Application Review," Journal of Computer Networks, Architecture and High Performance Computing, vol. 7, no. 3, pp. 720-732, 2025, doi: 10.3389/fcomp.2021.649508.

[23] A. R. Chowdhury, R. Paul, and F. Z. Rozony, "A systematic review of demand forecasting models for retail e-commerce enhancing accuracy in inventory and delivery planning," International Journal of Scientific Interdisciplinary Research, vol. 6, no. 1, pp. 1-27, 2025, doi: 10.63125/mbbfw637.

[24] Q. Zhang, X. Li, and P. Gao, "Forecasting Sales in Live-Streaming Cross-Border E-Commerce in the UK Using the Temporal Fusion Transformer Model," Journal of Theoretical and Applied Electronic Commerce Research, vol. 20, no. 2, pp. 92, 2025, doi: 10.3390/jtaer20020092.

[25] M. Sukel, S. Rudinac, and M. Worring, "Multimodal Temporal Fusion Transformers Are Good Product Demand Forecasters," IEEE MultiMedia, vol. 31, no. 2, pp. 48-60, 2024, doi: 10.1109/MMUL.2024.3373827.

[26] S. McCarthy and G. Alaghband, "Enhancing financial market analysis and prediction with emotion corpora and news co-occurrence network," Journal of Risk and Financial Management, vol. 16, no. 4, pp. 226, 2023, doi: 10.3390/jrfm16040226.

[27] Y. Hou, J. Li, Z. He, A. Yan, X. Chen, J. McAuley, et al., "Bridging language and items for retrieval and recommendation," arXiv preprint arXiv:2403.03952, 2024, doi: 10.48550/arXiv.2403.0395.

[28] R. Mousavi, B. Hazarika, K. Chen, and M. Razi, "The effect of online Q&As and product reviews on product performance metrics: Amazon," com as a case study. Journal of Information & Knowledge Management, vol. 20, no. 01, Art. no. 2150005, 2021, doi: 10.1142/S0219649221500052.

[29] U. Norinder and P. Norinder, "Predicting Amazon customer reviews with deep confidence using deep learning and conformal prediction," Journal of Management Analytics, vol. 9, no. 1, pp. 1-16, 2022, doi: 10.1080/23270012.2022.2031324.

[30] J. W. Kang and S. Y. Choi, "Comparative investigation of gpt and finbert's sentiment analysis performance in news across different sectors," Electronics, vol. 14, no. 6, pp. 1090, 2025, doi: 10.3390/electronics14061090.

[31] O. Shobayo, S. Adeyemi-Longe, O. Popoola, and B. Ogunleye, "Innovative sentiment analysis and prediction of stock price using FinBERT, GPT-4 and logistic regression: A data-driven approach," Big Data and Cognitive Computing, vol. 8, no. 11, pp. 143, 2024, doi: 10.3390/bdcc8110143.

[32] M. C. Lee, "Temporal Fusion Transformer-Based Trading Strategy for Multi-Crypto Assets Using On-Chain and Technical Indicators," Systems, vol. 13, no. 6, pp. 474, 2025, doi: 10.3390/systems13060474.

[33] Y. Zhou, L. Liao, Y. Gao, R. Wang, and H. Huang, "TopicBERT: A topic-enhanced neural language model fine-tuned for sentiment classification," IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 1, pp. 380-393, 2021, doi: 10.1109/TNNLS.2021.3094987.

[34] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," Electronics, vol. 9, no. 3, pp. 483, 2020, doi: 10.3390/electronics9030483.

[35] J. Hartmann, M. Heitmann, C. Siebert, and C. Schamp, "More than a feeling: Accuracy and application of sentiment analysis," International Journal of Research in Marketing, vol. 40, no. 1, pp. 75-87, 2023, doi: 10.1016/j.ijresmar.2022.05.005.

[36] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artificial Intelligence Review, vol. 55, no. 7, pp. 5731-5780, 2022, doi: 10.1007/s10462-022-10144-1.