International Journal of Management Science Research, Vol.8, Issue 9, (Sep)
ISSN 2536-605X **2025**
DOI: 10.53469/ijomsr.2025.08(09).07

Sales Prediction Based on Textual Features of Online Product Reviews

Hongyan Yu[1], Qirong Yang[2], Jiacheng Yang[3], and Tianlan Zhou[4,*]

[1,4]School of Transportation and Communication, Shanghai Maritime University, Shanghai, China
[2]Shanghai Wind Information Co., Ltd.
[3]School of Engineering, University of Southampton, Southampton, UK
**Corresponding Author*

Abstract: ***Accurate sales forecasting is essential for effective enterprise decision-making. Reliable predictions not only optimize inventory management and reduce resource waste but also enhance customer experience. While prior research has primarily focused on improving traditional time series models, relatively limited attention has been given to forecasting sales based on textual features extracted from online product reviews. This study addresses this gap by incorporating consumer-generated reviews into a sales prediction framework. Using Amazon.com as a case study, we analyze the open-source Amazon food review dataset. Through feature engineering, three categories of textual variables are constructed: review topics (extracted via the Gensim library and Latent Dirichlet Allocation, LDA), star ratings, and review usefulness. These features, together with lagged sales, are used as inputs into a ridge regression model. Experimental results show that the proposed model achieves an R^2 of 0.88 on the training set and 0.78 on the test set, indicating both feasibility and strong predictive accuracy. Compared with traditional time series methods, the review-based approach leverages text mining to capture consumers' genuine perceptions and market responses. This innovation enhances forecasting accuracy and offers theoretical as well as practical implications for enterprises, including improved sales planning, more adaptive market strategies, and more efficient warehouse and inventory management.***

Keywords: Online product reviews, Sales prediction, Text mining, Latent Dirichlet Allocation (LDA), Ridge regression.

1. INTRODUCTION

In business operations, accurate sales forecasting is essential for efficient supply chain management and cost control. Reliable forecasts not only optimize inventory management but also enhance customer service, allowing firms to gain a competitive edge in dynamic markets [1]. However, studies show that many companies still rely heavily on traditional time series forecasting methods. While these approaches are effective in stable environments, they often fail to respond adequately to rapidly changing market demands [2].

Traditional qualitative forecasting methods, which depend on expert judgment, are limited in precision and provide insufficient strategic guidance for enterprises. In contrast, quantitative forecasting methods—such as time series analysis and machine learning models—offer a more scientific basis for predicting demand, thereby supporting production planning and resource optimization [3]. Although models like ARIMA (AutoRegressive Integrated Moving Average) are widely applied, they primarily rely on historical data and neglect influential non-time factors such as consumer reviews. Furthermore, their ability to capture complex nonlinear relationships is limited, which restricts their predictive performance and weakens their economic interpretability [4].

As traditional time series models can only represent growth trends, periodicity, and random fluctuations, their forecasting performance is often generic. Recent research suggests that incorporating feature engineering and machine learning into sales forecasting can significantly improve accuracy, especially in dynamic markets [5]. By extracting diverse features to explain variations in sales, feature learning methods not only improve prediction accuracy but also strengthen the economic interpretability of model parameters. Compared with traditional approaches, this framework provides feasible insights for advancing sales forecasting research and offers both theoretical and practical guidance for production decisions.

Building on prior studies, feature engineering has been shown to enhance the accuracy of time series forecasting by capturing implicit patterns in complex datasets, enabling firms to better detect and respond to market trends [6]. Moreover, integrating regression models and neural networks within machine learning frameworks has proven effective in addressing demand volatility and providing robust decision support [7]. Therefore, this paper proposes a novel approach: constructing multidimensional features from online reviews, embedding them into a time series forecasting model via feature engineering, and employing an econometric model for prediction. This method

represents a cutting-edge direction in sales forecasting research.

2. LITERATURE REVIEW

Latent Dirichlet Allocation (LDA) is one of the most influential models in text topic modeling, capable of automatically uncovering hidden thematic structures within large document collections [8]. The core assumption of LDA is that each document can be represented as a mixture of multiple topics, with each topic characterized by a probability distribution over words. By analyzing word co-occurrence patterns, LDA clusters terms into coherent themes and reveals the latent semantic structure of the text. Owing to its robust text analysis capabilities, LDA has been widely applied across domains such as natural language processing, social sciences, and digital humanities. For example, in social media analysis, LDA has been used to identify themes in user-generated content to understand public opinion dynamics [9], while in the digital humanities it has helped uncover the thematic evolution of historical texts [10].

Over time, LDA has been extended to address new analytical challenges. Blei and McAuliffe [11] introduced Supervised LDA (sLDA), which incorporates labeled data to improve document classification accuracy, while Blei and Lafferty [12] proposed Dynamic Topic Models (DTM) to capture temporal changes in topics, enhancing applicability to time-series data. More recently, text analytics has gained momentum in business applications. Makkar and Jaiswal [13] demonstrated the potential of LDA and feature engineering for predicting e-commerce sales, illustrating its value in real-world commercial scenarios. These developments highlight the versatility of LDA as both a methodological foundation and a practical tool.

Despite these advantages, LDA still faces challenges. Topics generated by the model can be difficult to interpret, particularly when they overlap or when the number of topics must be specified in advance [14]. To address these issues, Teh et al. [15] proposed Hierarchical LDA (hLDA) and other non-parametric Bayesian models, which allow adaptive discovery of topic numbers. However, feature extraction in text mining often depends heavily on researchers' intuition rather than systematic methods, increasing the risk of omitting relevant variables or introducing noise. Moreover, extracting meaningful review-based features—such as topics, sentiment, and ratings—remains technically complex, as many tools in this area are still underdeveloped [16]. As the application of topic models expands, particularly in e-commerce platforms where user-generated reviews contain rich latent information, improving automated feature extraction methods will be crucial for enhancing both predictive performance and managerial decision-making.

3. FEATURE ENGINEERING BASED ON AMAZON REVIEW TEXTS

This study employs the Amazon open-source food review dataset, which is publicly available in CSV format from the official Amazon website. The dataset covers the period from 1999 to 2013 and contains 568,455 product reviews. To prepare the data for topic modeling and subsequent forecasting, we first provide a statistical overview of the dataset, followed by data preprocessing, including cleaning, transformation, and text-specific natural language processing.

3.1 Dataset Preparation and Description

Each review record contains the following fields: index number, product ID, consumer ID, username, product rating, timestamp, review text, number of opposing votes, and number of supportive votes. Table 1 presents the descriptive statistics of numeric variables.

Table 1: Descriptive statistics for numeric type variables

	Product Ratings	**Total number of commodity review comments**	**Number of product review support**
Reckoning	568455	568455	568455
Mean value	4.1823	2.2089	1.7276
Standard deviation	1.311	8.2592	7.6121
Minimum value	1	0	0
25 percent quartile	4	0	0
50 per cent quartile	5	1	0
75 per cent quartile	5	2	2
maximum values	5	923	866

The distribution of product ratings shows a left-skewed pattern, with higher-rated items generally corresponding to high-quality food products (see Figure 1). Accordingly, subsequent analyses use quality food as a baseline to examine the correlation between specific review indicators and product ratings. In contrast, the total number of reviews and the number of supportive votes is right-skewed, with extreme values exerting strong influence on the dataset. Moreover, review feedback and supportive votes are frequently missing.

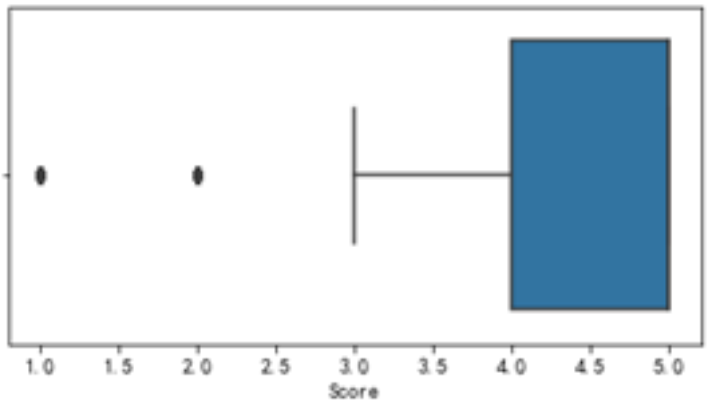

Figure 1: commodity star box diagram

To ensure data reliability, records containing NaN values or duplicate entries (caused by repeated reviews from the same user) were removed. Additionally, data from 1999–2005 were excluded, as sales during this period were sparse or even zero, which could distort time series analysis. Hence, all subsequent feature engineering and modeling used data from 2006 onward.

3.2 Data Type Conversion

Timestamps were originally recorded in Unix format and converted to Beijing time. Because daily intervals may introduce oscillation effects (e.g., higher weekend sales compared to weekdays) and overly coarse intervals would reduce sample size, weekly intervals were adopted. The conversion formula is:

$$timestamp' = \left[\frac{timestamp}{86400\times7}\right] \times 86400 \times 7 \quad (1)$$

where [ ] denotes the rounding function, and 86,400 is the number of seconds in a day. After conversion, 355 weeks were obtained, resulting in 564,901 valid review records, which are sufficient for subsequent model construction. Figure 2 illustrates the weekly sales trends, and Table 2 lists the data types of all dataset variables.

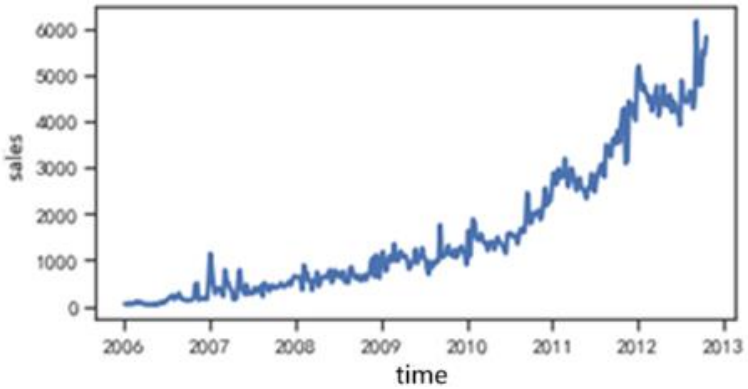

Figure 2: Line graph of weekly merchandise sales

Table 2: Data types and missing values after conversion of data set fields

field	reckoning	Whether there are deficiencies	data type
Product ID	564901	clogged	string
user ID	564901	clogged	string
Total number of commodity review comments	564901	clogged	integer
Number of product review support	564901	clogged	integer
Product Ratings	564901	clogged	integer

times	564901	clogged	Data Frame Timestamp
Comment text	564901	clogged	string

3.3 Natural Language Processing of Text

The raw dataset contained various issues, including irrelevant information, punctuation marks, inconsistent capitalization, and redundant words. To transform review texts into a bag-of-words representation suitable for Latent Dirichlet Allocation (LDA) topic modeling, it was necessary to perform Natural Language Processing (NLP). The main steps are as follows:

1) Symbol and Affix Processing

Using Python's re library, each review was processed line by line and sentence by sentence. Punctuation, emoticons, and tags—elements irrelevant to computational understanding of written language—were removed. Informal English suffix contractions (e.g., 's, 've, 'll, 're, 't, 'd, 'm) were also stripped to ensure lexical consistency.

2) Text Segmentation and Case Normalization

Since the dataset is in English, tokenization was performed by splitting on spaces. To reduce inconsistencies caused by capitalization, special cases, or misspellings, all words were converted to lowercase except for the initial character, which was standardized to uppercase. This normalization facilitates accurate word recognition by the algorithm.

3) Word Frequency Weighting

The Term Frequency–Inverse Document Frequency (TF-IDF) method was applied to assess word importance within the corpus. TF-IDF balances word frequency within a document against its distribution across the entire dataset, highlighting terms most representative of individual reviews.

4) Stopword Filtering and Vocabulary Refinement

After segmentation, non-informative words were removed in several stages. First, Python's stopword list was adopted to exclude prepositions, pronouns, articles, conjunctions, exclamations, and numbers. Second, words in the bottom 10% of TF-IDF scores were discarded, as they lacked topical significance. Third, words with extremely high or low document frequencies were excluded to reduce noise, while medium-frequency terms were retained to better capture thematic patterns. Fourth, tokens shorter than four characters (excluding standard stopwords) were removed, as these were often typographical errors or meaningless abbreviations. Finally, numerous food-category terms (e.g., product names) were eliminated, since their overwhelming presence risked biasing the analysis. This step allowed the extraction of cross-category features that reflect general patterns in consumer reviews rather than product-specific terminology.

As shown in Figure 3, the processed text differs substantially from the raw input. Reviews are segmented into token sequences forming a bag-of-words representation. After stopword filtering, nearly all irrelevant content is removed, leaving primarily descriptive terms related to food quality and consumer perception. These processed tokens provide a robust foundation for unsupervised learning with the LDA model.

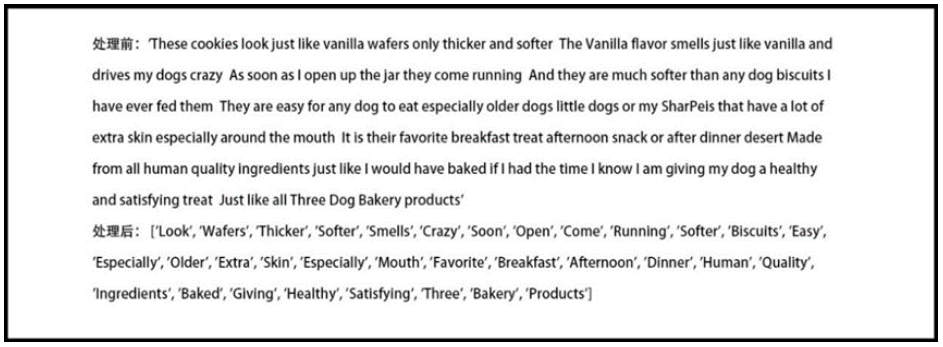

Figure 3: Schematic before and after preprocessing of product review text

Feature engineering—defined as the systematic extraction and transformation of raw data into informative variables—is a critical component of machine learning. Effective feature engineering not only improves model performance but also enhances interpretability. Its core techniques include data preprocessing, feature selection, and dimensionality reduction. In this chapter, the dataset is analyzed using the LDA topic model in conjunction with descriptive statistics to examine review features from three perspectives: (i) thematic content, (ii) star ratings, and (iii) perceived usefulness.

3.4 Review Topic Features Based on LDA Model Mining

This study applies the Latent Dirichlet Allocation (LDA) model to extract latent themes from the preprocessed Amazon food review dataset. The implementation uses the Gensim library in Python. Following common practice, the number of topics (K) is set to 5, the total vocabulary size (N) is 4,166, and the total number of documents (M) is 564,901. The Dirichlet hyperparameters are set as $\alpha = 10$ and $\beta = 0.01$, which reports that these values accelerate model convergence [17].

The processed dataset, number of keywords, and specified topics are entered into the LDA training function. The results of the topic modeling experiment are summarized in Table 3.

Table 3: Thematic content extracted from reviews

Theme 1	Theme 2	Theme 3	Theme 4	Theme 5
Strong	Price	Sweet	Order	Ingredients
Green	Store	Delicious	Price	Diet
Blend	Loves	Texture	Received	Natural
Nice	Local	Tasty	Bought	Protein
Smooth	Bought	Well	Package	Calories
Roast	Well	Nice	Arrived	Energy
Tastes	Brand	Favorite	Shipping	High
Bitter	Stores	Tastes	Bags	Tastes
Favorite	Grocery	Perfect	Purchase	Health
Flavored	Buying	Eating	Free	Bottle

Each review is assigned to the topic with the highest posterior probability. To construct topic-level features, the topic assignments are encoded using One-Hot Encoding, where the topic of the current review is marked as 1 and all others as 0. Based on this representation, the relative proportions of different topics across time periods are calculated. These proportions serve as the first set of input features for the predictive model, enabling analysis of whether shifts in review content (e.g., taste vs. nutrition) influence sales dynamics.

From Figure 4, it can be observed that only a small fraction of reviews focusses on food composition and nutritional health, while the majority are distributed more evenly across texture, taste, price, and logistics. These machine-learning-derived topics and their relative frequencies are largely consistent with the practical content of consumer reviews in contemporary e-commerce platforms.

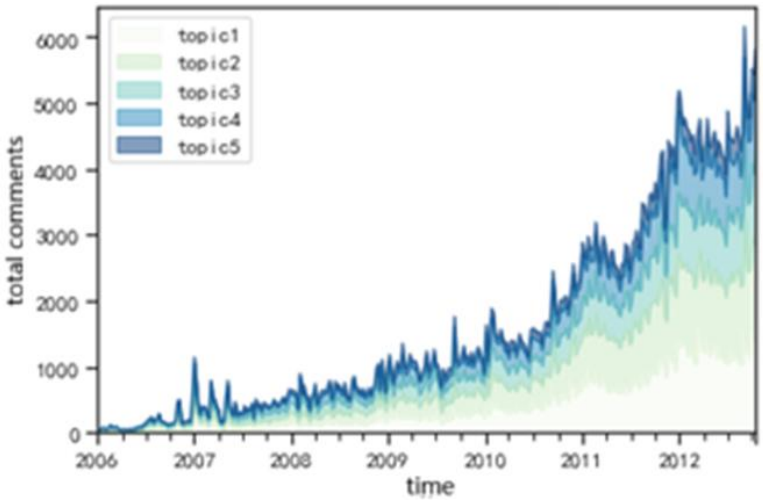

Figure 4: Line chart of the number of reviews by topic over time

3.5 Review Star Rating Feature Mining

Star ratings are the most direct reflection of consumer evaluations of product quality, and they can significantly influence subsequent purchase intentions. In this study, review data are aggregated on a weekly basis. Three descriptive statistical measures are extracted as features: (i) the mean star rating, (ii) the standard deviation of star ratings, and (iii) the proportion of reviews at each star level. Together, these indicators capture both the overall evaluation level and the degree of consensus among consumers.

3.6 Review Usefulness Feature Mining

To construct review usefulness features, the average number of helpfulness votes per review within each time period is first calculated. Since consumers typically view only a limited number of reviews when browsing products, a decline in the average number of votes indicates that the ratio of product sales to review votes has increased. This trend can serve as a proxy signal of rising sales, and thus a specialized metric is created and incorporated into the predictive model.

In e-commerce platforms, consumers can rate the helpfulness (or validity) of a review, akin to "likes" or "dislikes." Reviews with low helpfulness scores often deviate from the actual product experience, whereas highly rated reviews provide stronger assurances of authenticity and reliability. Such evaluations can influence potential buyers: high usefulness reinforces trust, while low usefulness may raise suspicions of review manipulation (e.g., astroturfing) or reflect divergent consumer opinions about the same product.

Accordingly, a usefulness indicator is introduced in feature engineering. The Helpfulness Rate (HR) is defined as:

$$Helpful_{Rate} = \frac{\sum HelpfulnessNumerator}{\sum HelpfulnessDenominator} \quad (2)$$

where the numerator is the number of users who rated a review as helpful, and the denominator is the total number of users who evaluated its helpfulness. Changes in this metric over time are illustrated in Figure 5(a) and Figure 5(b).

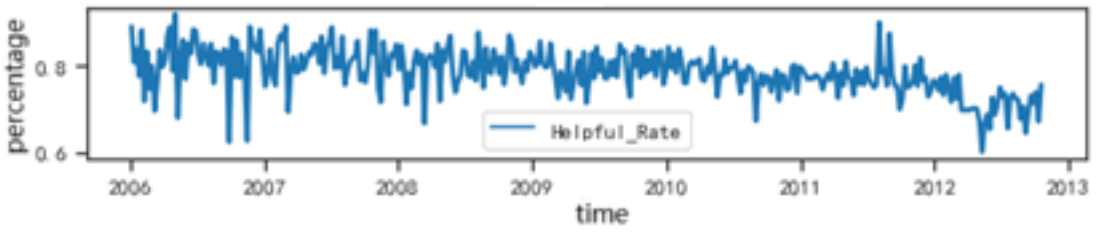

Figure 5 (a): Line graph of indicators of usefulness of comments

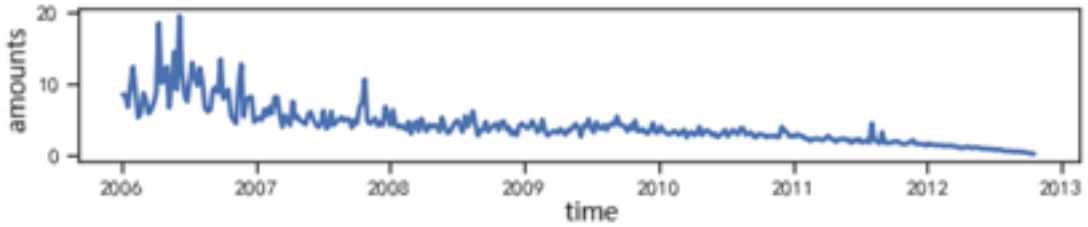

Figure 5 (b): Line graph of indicators of usefulness of comments

4. SALES PREDICTION MODEL BASED ON AMAZON TEXT FEATURES

Building on the results of the previous chapter, three categories of text-based review features were constructed: topic features, star rating features, and usefulness features, yielding a total of 15 variables for model input. Assuming that sales dynamics follow a Markov process, sales in the next period depend only on current features. Therefore, in addition to the constructed review features, the model also includes the lagged sales variable period's salesst−1 as an input.

4.1 Ridge Regression Model Construction and Fitting

To examine correlations among variables, a heatmap was generated using the Pearson Correlation Coefficient (PCC) (see Figure 6).

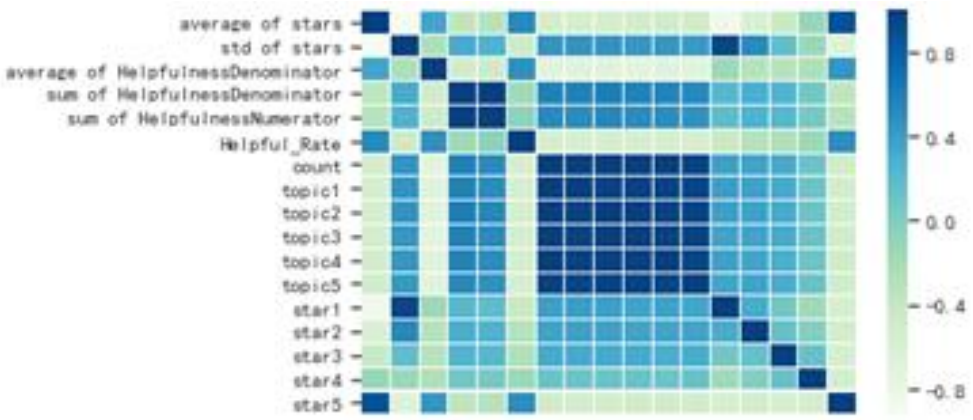

Figure 6: Heatmap of Pearson correlation coefficients among feature variables

The results show strong covariance among the constructed features. This is expected because many of the features are proportion-based, and their values sum to one, naturally introducing multicollinearity. Consequently, conventional multivariate regression models are unsuitable. To address this issue, this study employs the Ridge Regression model, which incorporates L2 regularization to mitigate multicollinearity. Although ridge regression sacrifices the strict unbiasedness of the Ordinary Least Squares (OLS) method and may reduce the goodness of fit, it improves model interpretability and produces more stable coefficient estimates, better reflecting economic reality.

When the ridge penalty parameter $\alpha = 0$, ridge regression degenerates to multiple linear regression. However, in high-dimensional settings with many variables, OLS often suffers from large coefficient variances, poor interpretability, and unrealistic estimates. As α increases, coefficients deviate from the OLS solution, but the L2 penalty term effectively shrinks variances, improving robustness. The key lies in choosing an appropriate value of α to balance variance reduction and bias. Following [18], a ridge trace plot was generated to determine the optimal α (Figure 7).

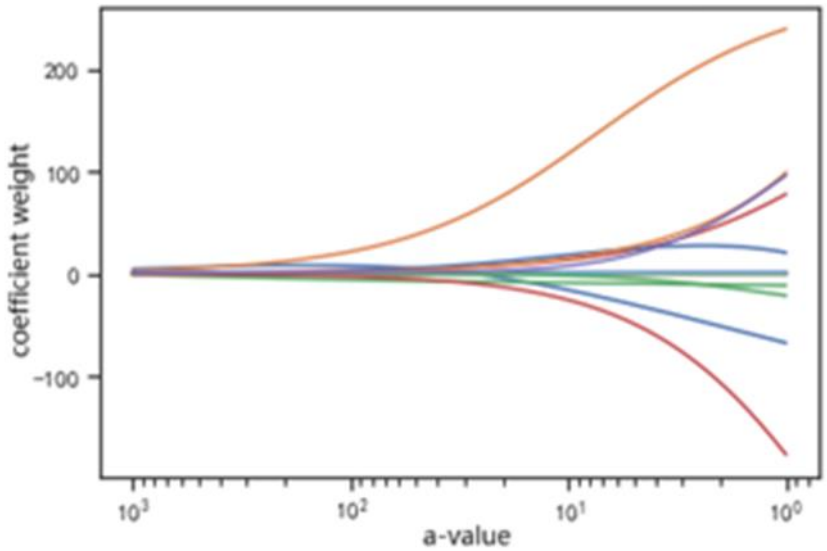

Figure 7: Ridge regression coefficients fitted along the ridge trace

The ridge trace indicates that at $\alpha = 20$, the regression coefficients stabilize, achieving an effective trade-off between variance and bias. Therefore, $\alpha = 20$ was selected for model fitting.

The dataset was partitioned into training and test sets: 250 weeks of data were used for training and the remaining 100+ weeks for testing. This split allows evaluation of model performance, particularly with respect to overfitting or underfitting. The model was implemented in Python using the scikit-learn (sklearn) library, and the results are

shown in Figure 8.

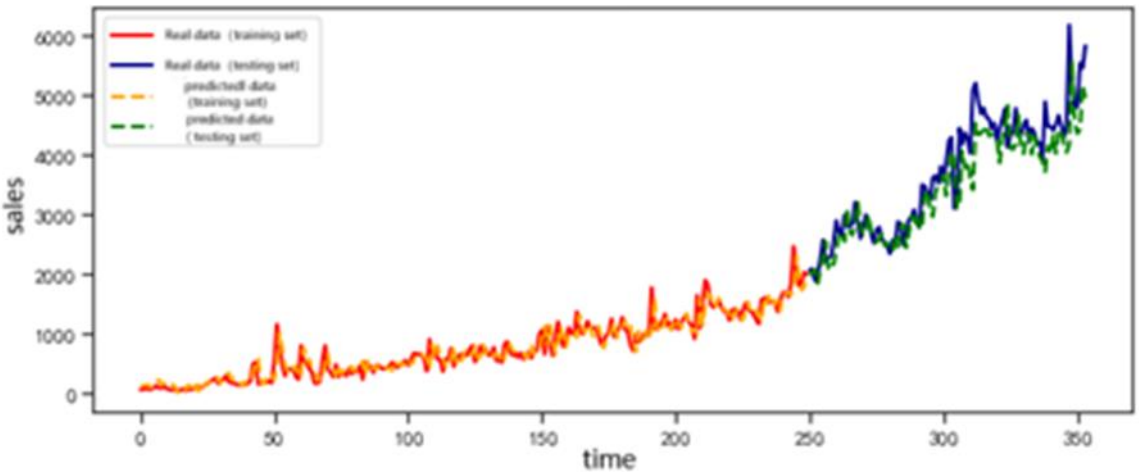

Figure 8: Ridge regression prediction results: training and test sets

4.2 Analysis of Model Residuals

From the fitted results, the regression model demonstrates strong predictive accuracy. The coefficient of determination (R^2) is 0.88 for the training set and 0.78 for the test set, indicating a good overall fit.

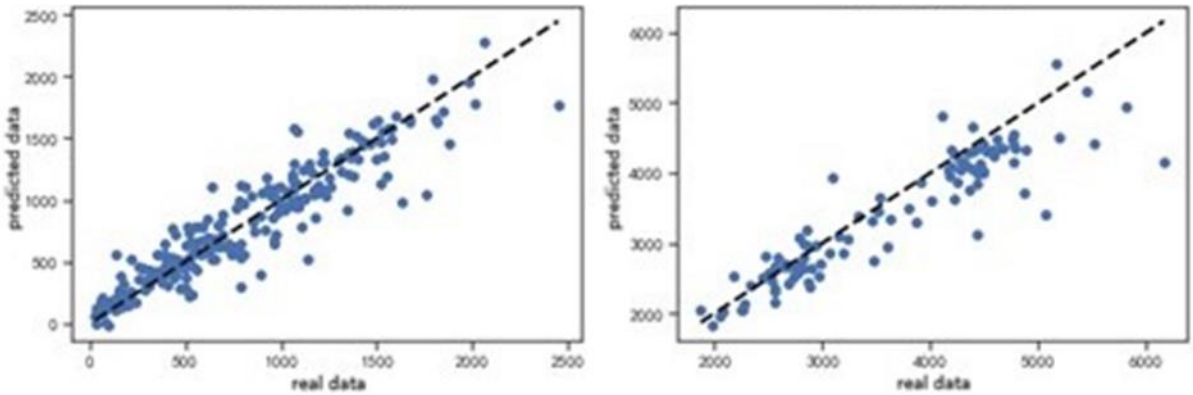

Figure 9: Scatterplot of regression residuals for training and test sets

As shown in Figure 9, the predicted values in the training set align closely with the observed data, with few large residuals. In the test set, only a small number of extreme cases show larger residuals, while most predictions remain close to the actual values, further confirming the good fit.

To assess residual normality, Quantile–Quantile (QQ) plots were generated. In statistics, residuals represent the difference between observed and predicted values, capturing the effect of factors not explained by the model.

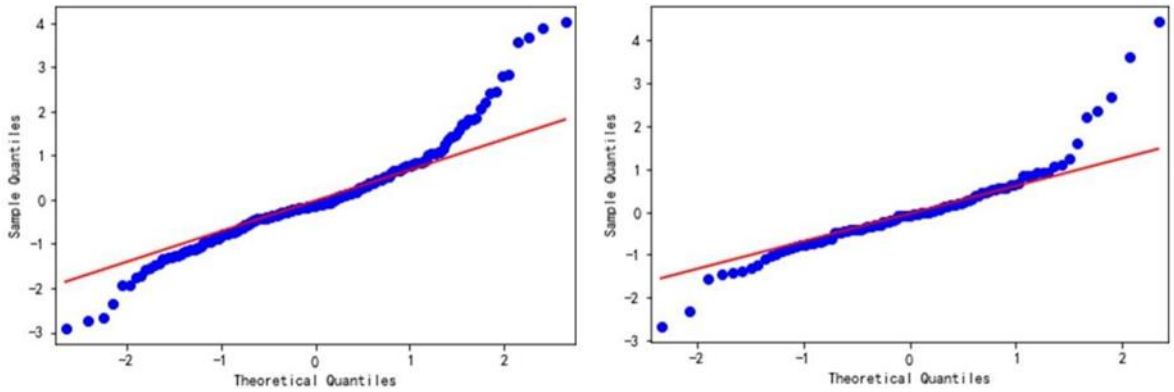

Figure 10: QQ plot of regression residuals

As shown in Figure 10, because ridge regression is a biased estimation method, its residuals do not strictly satisfy the normality assumption of classical multiple linear regression. Nevertheless, after standardization, the residuals closely approximate a normal distribution, with only a few deviations due to extreme values. Overall, the model provides a robust fit.

4.3 Model Comparison

To benchmark the proposed model, the Holt linear trend method was selected as a control, representing a traditional time-series forecasting approach. The Mean Absolute Percentage Error (MAPE) was employed to evaluate predictive accuracy. In this comparison, Holt's method was implemented with smoothing parameters $\alpha = 0.3$ and $\beta = 0.7$.

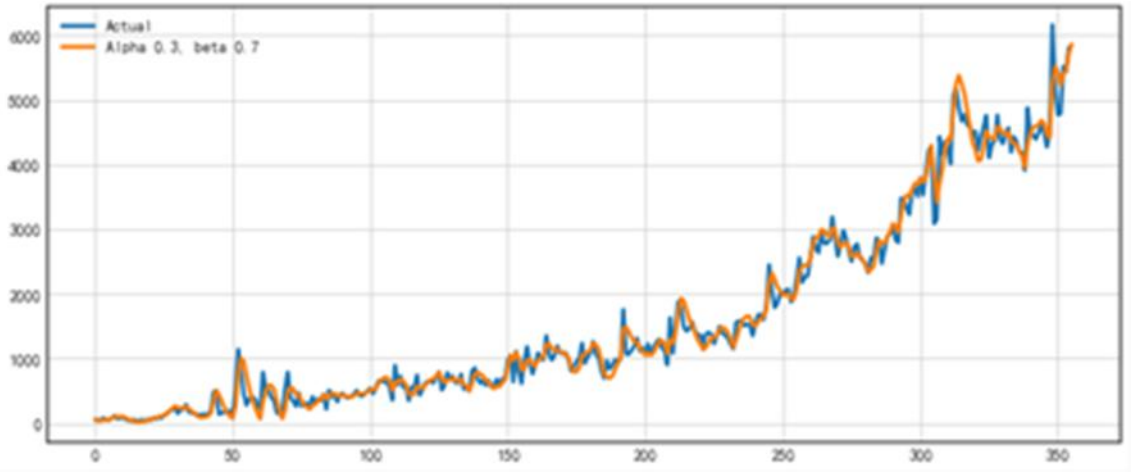

Figure 11: Fitting results of Holt's linear trend method

The results show that the MAPE of the feature-engineering-based model is 0.0840, compared with 0.1442 for Holt's method. This demonstrates that the proposed model achieves higher predictive accuracy, verifying its reliability and efficiency.

4.4 Model Discussion and Interpretation

Table 4: Parameters estimated by the ridge regression model

variant	Star rating average	Standard deviation of star rating	Average number of votes	Comments on usefulness	s_{t-1}
parameters	-35.96	42.16	-10.05	1.01	0.86
Thematic variables	Theme 1	Theme 2	Theme 3	Theme 4	Theme 5
parametric	0.26	0.46	-0.53	-0.23	0.90
starred variable	Star 1	Star 2	Star 3	Star 4	Star 5
parameters	7.71	5.48	-0.89	-9.54	-2.76

Table 4 reports the coefficients estimated by the ridge regression model. A positive coefficient indicates that the corresponding variable contributes to sales growth, whereas a negative coefficient suggests an inhibitory effect. The magnitude of the coefficient reflects the relative importance of the variable in influencing sales changes.

From a practical perspective, one might expect higher ratings and favorable reviews to consistently drive higher sales. However, the experimental results reveal more nuanced consumer behavior. Once products are generally rated above four stars and considered "high-quality," consumers appear more interested in consulting negative reviews to validate potential risks. The results show that:

- A lower average star rating combined with a higher standard deviation indicates divergent consumer opinions, which can amplify consumer interest.

- The coefficients for one-star and two-star reviews are positive and relatively large, suggesting that negative reviews may paradoxically stimulate sales by increasing consumer trust in review authenticity.

- Conversely, the coefficients for four-star and five-star reviews are negative and smaller, implying a weaker influence on incremental sales.
- Three-star (neutral) reviews exert little impact, indicating that consumers pay less attention to neutral opinions and are more influenced by polarized sentiments (positive or negative).

These findings suggest that consumers rely not only on favorable ratings but also on a balance of critical feedback when forming purchase decisions.

5. SUMMARY AND OUTLOOK

5.1 Contributions

Drawing on real business data and prior research, this study conducts an in-depth analysis of the Amazon food review dataset. Review topic classifications, derived through text mining, are incorporated as key input variables into the proposed sales prediction model. The main contributions are as follows:

1) Improved enterprise inventory management: By introducing a sales forecasting model based on consumer online reviews, firms can more accurately predict product demand. This enables optimization of safety stock levels, reduces the risk of overstocking or stockouts, and enhances inventory turnover. It also lowers warehousing and logistics costs while supporting production scheduling, logistics execution, and departmental coordination, thereby significantly improving operational efficiency and service quality.

2) Enhanced market strategy development: The model reveals potential market responses to consumer comments, providing firms with deeper insights into market trends and consumer demand. This helps adjust marketing and pricing strategies, guide new product development, and ensure that strategic decisions are more closely aligned with consumer expectations.

3) Improved customer experience: By analyzing online reviews, firms can quickly detect issues in products or services and implement timely improvements. Such responsiveness enhances customer satisfaction and loyalty, strengthens brand reputation, and ultimately increases market competitiveness.

5.2 Research Directions

1) Improved data representation: In this study, the number of reviews was used as a proxy for sales volume. However, not all consumers leave reviews after purchase, which may bias prediction results. Future research should employ datasets that separately provide sales figures, review counts, and review content for specific products and periods. This would allow more accurate validation of the model's practical applicability.

2) Interaction between review sentiment and sales: Future work could examine the relationship between review sentiment (positive, negative, neutral) and sales performance, analyzing how variations in sentiment intensity influence demand. Such findings would provide firms with more refined decision support for marketing and product management.

3) Multimodal data fusion: Integrating multimodal data—such as images and videos—with textual reviews represents a promising avenue for improving sales prediction accuracy. This approach would require more sophisticated feature extraction and model fusion techniques, potentially enhancing the robustness of forecasting systems.

REFERENCES

[1] Chopra, S., & Meindl, P. Supply Chain Management: Strategy, Planning, and Operation. Pearson, 2016.
[2] Waller, M. A., & Fawcett, S. E. Click Here for a Data Scientist: Big Data, Predictive Analytics, and Theory Development in the Era of a Maker Movement Supply Chain. Journal of Business Logistics, 2013, 34(4): 249-252.
[3] Makridakis, S., & Hibon, M. The M3-Competition: Results, Conclusions, and Implications. International Journal of Forecasting, 2000, 16(4): 451-476.
[4] Makkar, Sandhya, and Sneha Jaiswal. "Predictive analytics on e-commerce annual sales." Proceedings of Data Analytics and Management: ICDAM 2021, Volume 1. Springer Singapore, 2022.

[5] Hyndman, R. J., & Athanasopoulos, G. Forecasting: principles and practice. OTexts, 2018.
[6] Fulcher, B. D., & Jones, N. S. hctsa: A Computational Framework for Automated Time-Series Phenotyping Using Massive Feature Extraction. Cell Systems, 2017, 5(5): 527-531.
[7] Bontempi, G., Taieb, S. B., & Le Borgne, Y. A. Machine Learning Strategies for Time Series Forecasting. European Business Intelligence Summer School, 2012, 62-77.
[8] Blei, D. M., Ng, A. Y., & Jordan, M. I. *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 2003.
[9] Hong, L., & Davison, B. D. *Empirical study of topic modeling in Twitter*. Proceedings of the First Workshop on Social Media Analytics, 2010.
[10] Meeks, E., & Weingart, S. *The Digital Humanities Contribution to Topic Modeling*. Journal of Digital Humanities, 2012.
[11] Blei, D. M., & McAuliffe, J. D. *Supervised topic models*. Advances in Neural Information Processing Systems, 2007.
[12] Blei, D. M., & Lafferty, J. D. *Dynamic topic models*. Proceedings of the 23rd International Conference on Machine Learning, 2006.
[13] Makkar, S., & Jaiswal, S. *Predictive analytics on e-commerce annual sales*. Proceedings of Data Analytics and Management: ICDAM 2021, Volume 1. Springer Singapore, 2022.
[14] Griffiths, T. L., & Steyvers, M. Finding scientific topics. Proceedings of the National Academy of Sciences, 2004.
[15] Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 2006.
[16] Makkar, Sandhya, and Sneha Jaiswal. "Predictive analytics on e-commerce annual sales." Proceedings of Data Analytics and Management: ICDAM 2021, Volume 1. Springer Singapore, 2022.
[17] Wei X, Croft W B. LDA-based document models for ad-hoc retrieval[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. 2006: 178-185.
[18] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to linear regression analysis. Wiley.