Sentiment Analysis of Women's Clothing E-Commerce Reviews

Jiadan Zhao **Miya Huang** **Qinyi Hu** **Wenxin Xu**

Abstract

Sentiment analysis plays a critical role in enhancing e-commerce platforms by improving product recommendations and understanding customer feedback. This study explores sentiment classification in women's clothing reviews using three vectorization techniques—Count Vectorizer, TF-IDF, and Word2Vec—paired with multiple classification models, including Logistic Regression, Naïve Bayes, SVM, Random Forest, AdaBoost, GRU, and LSTM. Our findings indicate that among traditional machine learning models, SVM with TF-IDF provides the most effective sentiment classification, particularly for shorter reviews. Meanwhile, LSTM demonstrates superior performance for longer, context-rich reviews due to its ability to capture sequential dependencies. These insights can aid e-commerce businesses in selecting the most suitable sentiment analysis approach based on review characteristics.

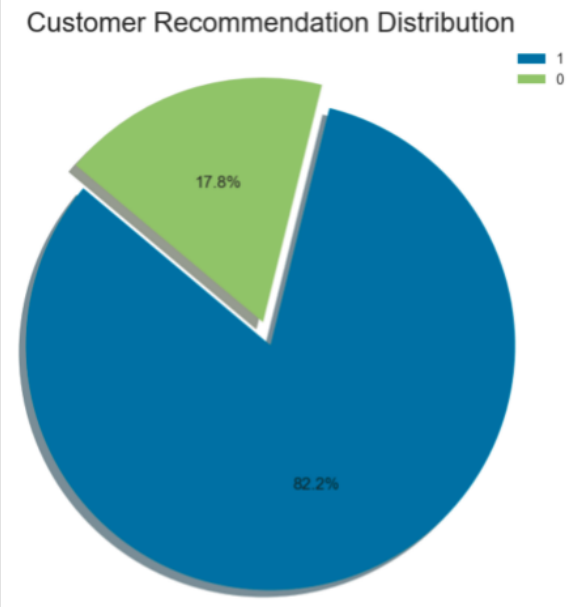

Figure 1: The percentage of customers recommendation

1 Dataset

1.1 Data Source

The dataset utilized in this study is the Women's Clothing E-Commerce Reviews Dataset, comprising 22,641 customer-generated textual reviews alongside several metadata attributes. Each data entry represents an individual customer's review and includes fields such as the product rating, reviewer's age, textual content of the review ("Review Text"), and an indicator of whether the customer recommended the product ("Recommended_IND"). This binary variable, indicating customer satisfaction, serves as our target for sentiment analysis. Given the commercial sensitivity, all identifying references to brands or retailers have been anonymized.

1.2 Exploratory Data Analysis (EDA)

To systematically understand the data and ensure informed modeling decisions, a comprehensive exploratory data analysis (EDA) was conducted. We explored various aspects of the dataset, including missing values, class distribution, review rating distribution, review text length, and frequently mentioned words, to uncover underlying patterns relevant to customer sentiment.

1.2.1 Missing Values

The primary feature for sentiment classification is "Review Text." Instances with missing review text were removed to ensure data quality. Other columns with minor missing values were ignored as they were not utilized in modeling. The final dataset retained 22,641 valid reviews.

1.2.2 Class Distribution

The dataset is highly imbalanced, with 82% positive (recommended) and 18% negative (not recommended) reviews. This imbalance may cause models to favor positive predictions. Therefore, macro F1-score and recall were prioritized in model evaluation.

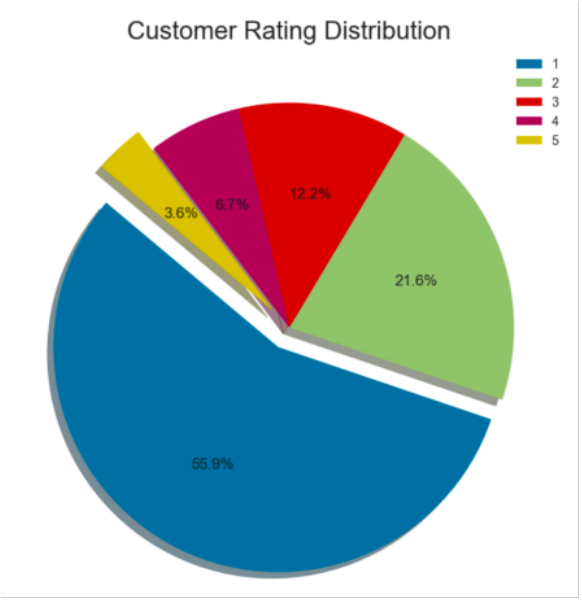

Figure 2: The percentage of customers ratings

1.2.3 Rating Analysis

Product ratings in the dataset ranged from 1 (worst) to 5 (best). Ratings distribution heavily skewed towards positive ratings, predominantly ratings of 4 or 5, further highlighting a positive feedback bias inherent in the dataset. Such skewness underscores the importance of using nuanced text-based models rather than relying solely on numerical ratings to predict customer recommendations.

1.2.4 Review Length Analysis

The lengths of customer reviews exhibited considerable variability. While some reviews were concise, comprising only a few words, others exceeded 100 words, with an average length around 60 words. Reviews typically clustered between 50 to 100 words, indicating sufficient textual content for meaningful sentiment analysis. Review length may significantly influence the effectiveness of different models, particularly favoring sequential models for longer reviews. Notably, reviews with fewer words may provide limited contextual clues, thus benefiting more from simpler frequency-based models, whereas longer reviews might contain richer context, better captured by recurrent neural network models.

1.2.5 Word Cloud Analysis

The overall word cloud highlighted frequently mentioned words such as “dress,” “top,” “love,” “fabric,” and “sweater,” reflecting general customer concerns and interests. Positive reviews predominantly

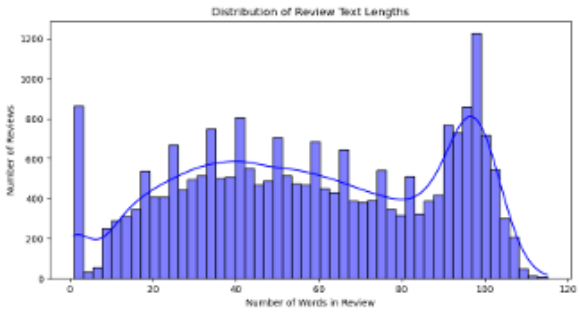

Figure 3: The distribution of review text lengths

Figure 4: Word cloud of the most used words in customer reviews

featured terms like “perfect,” “comfortable,” and “great,” indicating high customer satisfaction regarding fit, comfort, and overall product quality. Conversely, negative reviews frequently mentioned terms such as “small,” “tight,” “disappointed,” and “returned,” emphasizing problems related to sizing and fit. These findings clearly illustrate that accurate product descriptions, especially regarding sizing and fabric quality, are critical for improving customer satisfaction on e-commerce platforms. Thus, enhancing sizing accuracy and providing detailed product information can significantly reduce dissatisfaction rates.

Figure 5: Word cloud of the most used words in positive customer reviews

Figure 6: Word cloud of the most used words in negative customer reviews

2 Literature Study

Sentiment analysis of online customer reviews has become essential for e-commerce businesses seeking to improve product offerings and enhance customer experience. Recent research has explored various methodologies, offering valuable insights into model selection, preprocessing techniques, and sentiment classification accuracy.

Mahmud et al. (2023) applied advanced NLP methods, including topic modeling and sentiment visualization, to women's clothing e-commerce reviews. Their study found that Random Forest outperformed other models, achieving an accuracy of approximately 96.51%, demonstrating the strength of ensemble models in capturing complex, nonlinear sentiment patterns. This insight influenced our decision to evaluate Random Forest and AdaBoost in our experiments.

Karthik and Ganapathy (2021) proposed a fuzzy logic-based recommendation system that integrates sentiment analysis and ontology to predict consumer preferences dynamically. Their study illustrates how incorporating sentiment-based insights can enhance personalization and improve recommendation accuracy. Hassan and Ghonem (2022) highlighted that textual reviews provide more reliable sentiment indicators than numerical ratings, as the latter often exhibit biases. They evaluated different vectorization methods (frequent sentence vectors and TF-IDF) alongside oversampling techniques to address class imbalance. Their findings suggest that oversampling significantly improves minority class representation but comes at the cost of increased computational complexity. Inspired by their approach, we adopted class-weight adjustments and SMOTE oversampling to mitigate class imbalance while maintaining computational efficiency.

Prince et al. (2024) found that simpler classifiers, such as KNN, outperformed deep learning models (LSTM, CNN) in clothing review sentiment classification. This finding underscores the impact of dataset characteristics on model performance, revealing that traditional models, including Naïve Bayes and SVM, can sometimes match or exceed deep learning methods depending on text representation techniques.

Kamal (2024) further reinforced this notion by demonstrating that simple classifiers consistently achieved competitive performance when coupled with effective data preprocessing and vectorization methods. Motivated by Kamal's findings, we systematically compared machine learning and deep learning classifiers across different text vectorization techniques to examine how feature representation influences classification accuracy.

Collectively, these studies emphasize the importance of aligning feature representation with model complexity, forming the basis for our experimental approach. Our study builds upon these findings by evaluating how vectorization methods impact classifier performance, addressing class imbalance issues, and investigating the trade-offs between traditional and deep learning models in e-commerce sentiment analysis.

3 Methodology

This study employs a structured approach to sentiment classification by processing review text, converting it into numerical representations, and applying various machine learning and deep learning models. The dataset was first cleaned to ensure the removal of missing values in the review text, as textual content is the primary feature for classification. Standard preprocessing techniques were applied, including lowercasing, punctuation removal, stopword filtering, and lemmatization, to standardize the text and enhance model efficiency. Additionally, rare words appearing fewer than three times were removed to balance vocabulary size while retaining meaningful linguistic patterns.

To convert text into numerical representations, three vectorization methods were used: Count Vectorization, TF-IDF, and Word2Vec. Count Vectorization provided a straightforward frequency-based representation, while TF-IDF refined this approach by weighting words based on their importance across documents. Word2Vec, in contrast, generated dense vector embeddings to capture semantic relationships between words. These vec-

Model	Vector	Accuracy	F1-score
LR	count	0.87	0.81
LR	TF-IDF	0.85	0.79
LR	w2v	0.78	0.72
Naive Bayes	count	0.89	0.82
Naive Bayes	TF-IDF	0.84	0.60
Naive Bayes	w2v	0.72	0.66
SVM	count	0.87	0.81
SVM	TF-IDF	0.85	0.79
SVM	w2v	0.78	0.72
RF	count	0.85	0.78
RF	TF-IDF	0.83	0.76
RF	w2v	0.83	0.73
Ada Boost	count	0.88	0.79
Ada Boost	TF-IDF	0.88	0.78
Ada Boost	w2v	0.85	0.72
DL-GRU	-	0.87	0.81
DL-LSTM	-	0.87	0.80

Table 1: Summary of different models performances on the training set.

torized representations were then used as input for various classification models.

Both traditional machine learning and deep learning models were explored. Logistic Regression, Naïve Bayes, SVM, Random Forest, and AdaBoost were trained using Count Vectorized and TF-IDF features, leveraging different learning approaches such as probabilistic classification, support vector methods, and ensemble learning. Additionally, deep learning models, specifically GRU and LSTM, were employed to capture sequential dependencies in text data. The choice of these models was based on their established effectiveness in text classification tasks.

To ensure robust evaluation, an 80/20 train-test split was applied, and models were fine-tuned through cross-validation. Given the dataset's class imbalance, macro F1-score and recall were prioritized to provide a fair evaluation of performance across both sentiment classes. The results and performance comparisons of these models are detailed in the following section.

4 Experiments

In this section, we describe the experimental setup used for sentiment classification of women's clothing reviews. Our objective was to determine whether a customer recommends a product based on their review text. To achieve this, we implemented three different text vectorization techniques: Count Vectorization, TF-IDF, and Word2Vec. These representations were then used as input for various classification models, including machine learning algorithms such as Logistic Regression, Naïve Bayes, Support Vector Machine (SVM), Random Forest, and AdaBoost, as well as deep learning models like GRU and LSTM.

The dataset was split into training and test sets using an 80/20 ratio. Given the class imbalance, with 82% of the reviews being positive and 18% negative, we prioritized macro-averaged F1-score and recall over simple accuracy for model evaluation. Hyperparameter tuning was conducted using a 5-fold cross-validation approach to prevent overfitting and ensure robust performance across different models.

We systematically evaluated multiple classifiers, including traditional machine learning models (Logistic Regression, Naïve Bayes, SVM, Random Forest, AdaBoost) and deep learning methods (GRU and LSTM). Class imbalance issues were addressed using class-weight adjustments, and model performance was assessed primarily using macro F1-score and recall metrics.

4.1 Count Vectorizer

TF-IDF (Term Frequency-Inverse Document Frequency) improves upon Count Vectorization by adjusting the weights of words based on their importance across the corpus. The resulting feature space also had 5,794 dimensions, maintaining a similar computational complexity as the Count Vectorizer.

SVM with TF-IDF provided the best results among traditional machine learning models, achieving 85% accuracy and 79% macro F1-score. Logistic Regression performed similarly, while Naïve Bayes struggled with TF-IDF due to its reliance on raw frequency distributions, leading to a significantly lower macro F1-score of 60%. Random Forest and AdaBoost performed slightly worse than they did with Count Vectorization.

TF-IDF successfully enhanced classification by highlighting distinguishing terms while reducing the influence of common words. However, in some cases, it overly emphasized rare words, introducing noise into the classification process.

4.2 TF-IDF Vectorizer

Word2Vec generates dense vector representations of words by capturing their semantic relationships. We trained a 100-dimensional Word2Vec model

on the full dataset and represented each review by averaging the embeddings of its constituent words. This significantly reduced the feature space compared to the sparse matrices of Count and TF-IDF vectorization.

Traditional classifiers applied to Word2Vec embeddings yielded mixed results. SVM and Logistic Regression achieved 78% accuracy and 72% macro F1-score, but Naïve Bayes was not well-suited for this approach due to its assumption of independent features. Random Forest and AdaBoost also showed inconsistencies, as the semantic relationships captured by Word2Vec were not always effectively leveraged by these models.

The main advantage of Word2Vec was its ability to retain meaning beyond individual word frequencies, making it particularly useful for identifying sentiment cues in longer reviews. However, the simple averaging method used to generate sentence vectors lost important contextual dependencies, limiting its performance.

4.3 Word2Vec

To capture sequential dependencies in review text, we implemented recurrent neural networks using Gated Recurrent Units (GRU) and Long Short-Term Memory (LSTM) networks. These architectures are well-suited for sentiment analysis as they process text sequentially, preserving context over longer passages.

The models were trained using an embedding layer initialized with 100-dimensional random vectors, followed by three recurrent layers. Dropout regularization was applied to prevent overfitting. The output layer used a sigmoid activation function, and the Adam optimizer was employed with a learning rate of 0.001.

LSTM slightly outperformed GRU, achieving 87% accuracy and 81% macro F1-score, compared to 87% accuracy and 80% macro F1-score for GRU. Both models performed particularly well on longer reviews, where contextual relationships played a more significant role. However, training deep learning models required significantly more computational resources, and without sufficient regularization, there was a risk of overfitting to majority-class data.

4.4 GRU and LSTM

Recurrent neural networks (GRU and LSTM) explicitly modeled sequential context in longer, detailed reviews, addressing the shortcomings of basic embedding aggregation methods. These models excelled in capturing complex contextual patterns, particularly beneficial for sentiment analysis of lengthy and nuanced reviews. Utilizing embedding layers and dropout regularization techniques effectively mitigated overfitting risks due to dataset imbalance and complexity. LSTM exhibited slightly superior performance compared to GRU (Accuracy: 87%, Macro F1-score: 81%), highlighting its strength in processing complex sequential textual information.

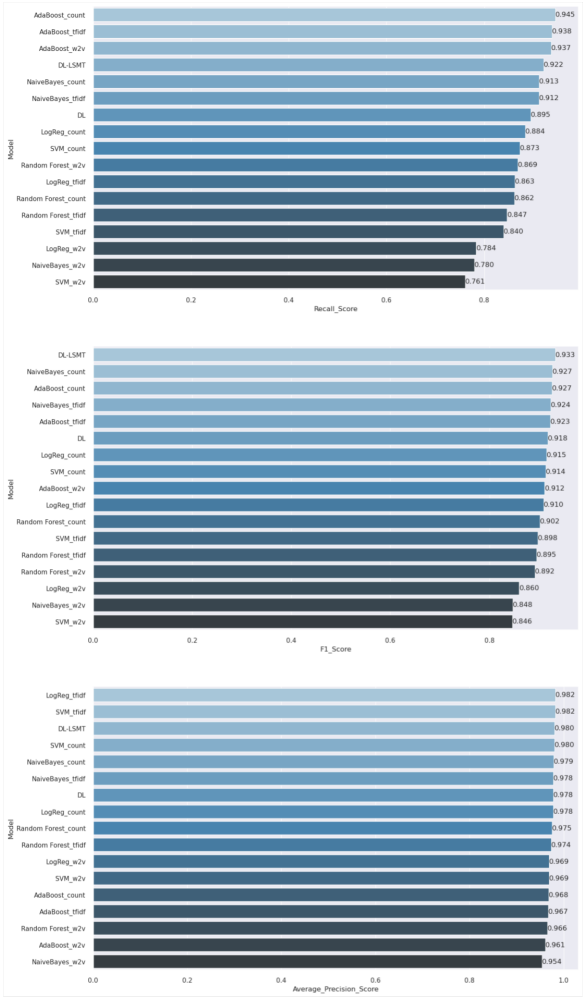

Figure 7: The Recall score, F1 score, and average precision score of different models

In summary, the experiments demonstrated that matching text vectorization and classification methods to specific data characteristics significantly enhances sentiment analysis performance, providing valuable guidance for model selection in practical e-commerce contexts.

4.5 Summary and Discussion

Our experiments demonstrated that different text representation techniques and classification models impact sentiment classification performance in distinct ways. SVM with TF-IDF consistently outperformed other traditional models, achieving 85% accuracy and a macro F1-score of 79%. Naïve Bayes performed best with Count Vectorization but struggled with TF-IDF due to its probabilistic assumptions. Word2Vec embeddings provided meaningful semantic information but resulted in slightly lower classification performance compared to TF-IDF.

Among deep learning models, LSTM achieved the highest performance with an accuracy of 87% and a macro F1-score of 81%, particularly excelling in analyzing longer reviews. GRU showed comparable results but had slightly lower recall, making LSTM the preferred choice for handling sequential text data. However, these deep learning models required significantly more computational resources, and their effectiveness varied depending on the length and complexity of the reviews.

The results indicate that traditional machine learning models, particularly SVM with TF-IDF, remain highly competitive for sentiment classification when computational efficiency is a priority. Deep learning approaches provide more nuanced sentiment capture but come with increased complexity and computational demands. The trade-offs between model interpretability, resource efficiency, and performance must be carefully considered depending on the application.

5 Conclusion

This study explored sentiment classification in women's clothing e-commerce reviews using various text vectorization techniques and classification models. Our findings suggest that SVM with TF-IDF provides a strong baseline for sentiment classification, offering a balance between accuracy and computational efficiency. Among deep learning models, LSTM demonstrated superior performance in capturing sentiment nuances in longer reviews but required more resources and fine-tuning.

The choice of model should be driven by the nature of the reviews and the available computational resources. For short and straightforward reviews, SVM with TF-IDF is a highly effective and interpretable solution, whereas for longer and more complex reviews, LSTM offers better contextual understanding. These findings emphasize the need for context-aware model selection in sentiment analysis applications.

While our models performed well, handling class imbalance and incorporating contextual embeddings remain areas for improvement. More sophisticated deep learning architectures or domain-specific enhancements could further enhance performance. Future research should explore these directions to refine sentiment classification techniques for real-world applications.

6 Future Work

Although our models achieved strong classification performance, several areas warrant further exploration. First, addressing class imbalance more effectively could improve the recall of the minority class (negative reviews). Techniques such as data augmentation or synthetic review generation could be investigated.

Second, the use of transformer-based models such as BERT could provide richer contextual representations compared to traditional word embeddings like Word2Vec. Fine-tuning pre-trained models on e-commerce-specific data may improve accuracy, particularly for longer reviews where sentiment is more nuanced.

Third, integrating multi-modal data sources, such as product images, structured metadata, and customer demographics, could enhance predictive capabilities. Sentiment analysis models could benefit from combining textual data with additional context to provide more personalized recommendations.

Finally, optimizing computational efficiency for deep learning models remains a challenge. Exploring techniques such as knowledge distillation, pruning, or low-rank factorization could make LSTM and transformer models more viable for real-time sentiment analysis in e-commerce platforms.

References

M. Mahmud, R. A. Mullick, and M. C. Anas. 2023. "Sentiment Analysis of Women's Clothing Reviews on E-commerce Platforms: A Machine Learning Approach." *Online*, June 2023.

R. V. Karthik and S. Ganapathy. 2021. "A fuzzy recommendation system for predicting the customers' interests using sentiment analysis and ontology in e-commerce." *Applied Soft Computing, vol. 108, p. 107396*. doi: 10.1016/j.asoc.2021.107396.

M. Hassan and H. Ghonem. 2022. "Sentimental Analysis on Women Clothes E-commerce Reviews." *Online*, April 2022. doi: 10.13140/RG.2.2.12974.18248.

N. Uddin Prince, M. Nazmul, H. Shawon, and Z. Kamal. 2024. "E-commerce Clothing Review Analysis by Advanced ML Algorithms." *World Journal of Advanced Research and Reviews, vol. 24, no. 1*. doi: 10.30574/wjarr.2024.24.1.3044.

Z. Kamal. 2024. "E-commerce-Clothing-Review-Analysis-by-Advanced-ML-Algorithms." Presented at the *Conference: E-commerce Clothing Review Analysis by Advanced ML Algorithms*, Dhaka, Bangladesh, November 2024.

A Appendix

Github repository link: https://github.com/rsm-wex015/415-final