WOMEN'S CLOTHING E-COMMERCE SENTIMENT ANALYSIS

PHASE 4 FINAL PROJECT
DATE: 02/04/2025

GROUP 5 MEMBERS
Ian Bett
Lionel Ajeliti
Linet Patriciah
Sanayet Nelly Sankaine
Morgan Abukuse Amunga

Table of Contents

1. Business Understanding..... 4
1.1 Problem Statement..... 4
1.2 Objectives..... 4
1.3 Stakeholders..... 4
1.4 Business Success Criteria..... 4
2. Data Understanding.....5
2.1 Data Collection.....5
2.2 Data Description..... 5
2.3 Data Quality and Initial Inspection..... 5
3. Exploratory Data Analysis (EDA)..... 6
3.1 Ratings and Sentiment Distribution.....6
3.2 Department names Distribution..... 6
3.3 Recommended Vs Not Recommended..... 7
3.4 Age Distribution.....8
3.5 Sentiments Bigram.....8
3.6 Class Name Distribution..... 9
4. Data Preparation.....10
4.1 Handling Missing Data..... 10
4.2 Removing Duplicates..... 10
4.3 Text-Specific Preprocessing.....10
4.4 Feature Engineering and Representation..... 10
5. Modeling..... 11
5.1 Train-Test Split..... 11
5.2 Baseline Model (Logistic Regression).....11
5.3 Advanced Model Evaluation.....11
5.4 Hyperparameter Tuning..... 11
5.5 Cross-Validation..... 11
6. Model Evaluation and Validation.....12
6.1 Evaluation Metrics.....12
6.2 Confusion Matrix and Analysis.....12
6.3 Error Analysis and Insights..... 12
7. Model Interpretability and Explainability.....13
7.1 Global Feature Importance..... 13
7.2 Local Interpretability (LIME).....13
8. Deployment and Operationalization..... 14
8.1 Deployment Strategy..... 14

8.2 Monitoring and Maintenance....14
9. Conclusions and Recommendations.... 15
9.1 Project Limitations....15
9.2 Key Conclusions.... 15
9.3 Business Recommendations.... 16
9.4 Recommendations for Future Modeling....16
10. Business Impact.... 18

1. Business Understanding

The foundation for our analysis clearly outlines the business scenario, defining the importance of sentiment analysis in understanding and enhancing customer experience in the e-commerce sector, particularly women's clothing.

1.1 Problem Statement

Women's clothing retailers face high product returns and customer dissatisfaction due to issues like incorrect sizing, inconsistent product quality, and unclear marketing strategies. Analyzing customer reviews helps uncover insights to address these problems proactively.

1.2 Objectives

- Develop a robust sentiment analysis model to classify reviews as positive, negative, or neutral.
- Identify recurring themes in customer feedback.
- Provide actionable insights to enhance customer satisfaction and product quality.

1.3 Stakeholders

- **Product Management Team:** to optimize product design and sizing.
- **Marketing Department:** for tailored advertising and customer engagement strategies.
- **Customer Service Team:** to anticipate and proactively address frequent complaints.

1.4 Business Success Criteria

- Reduced return rates by 15-20%.
- Enhanced customer satisfaction and higher customer ratings.
- Increased customer retention rates.

2. Data Understanding

This stage involves gathering data, understanding its structure, and assessing its quality to ensure suitability for analysis.

2.1 Data Collection

We obtained a Kaggle dataset titled "Women's Clothing E-Commerce Reviews," containing 23,486 customer reviews, ratings, and demographic details.

2.2 Data Description

The dataset includes review text, rating (1-5 stars), age, clothing department, product category, and recommendation status (yes/no), providing comprehensive demographic and qualitative insights.

2.3 Data Quality and Initial Inspection

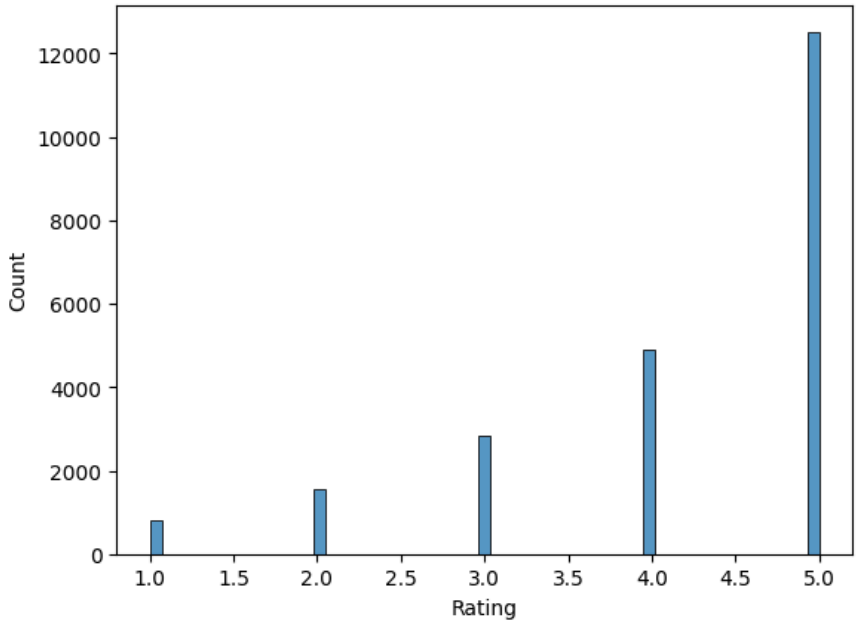
- Rows: 23,486
- Missing review texts: 845 (3.6%)
- Duplicates: 21 records
- Action Taken: Removed duplicates and missing entries.

3. Exploratory Data Analysis (EDA)

Exploring the data provided deeper insights into customer behaviors, preferences, and frequent feedback.

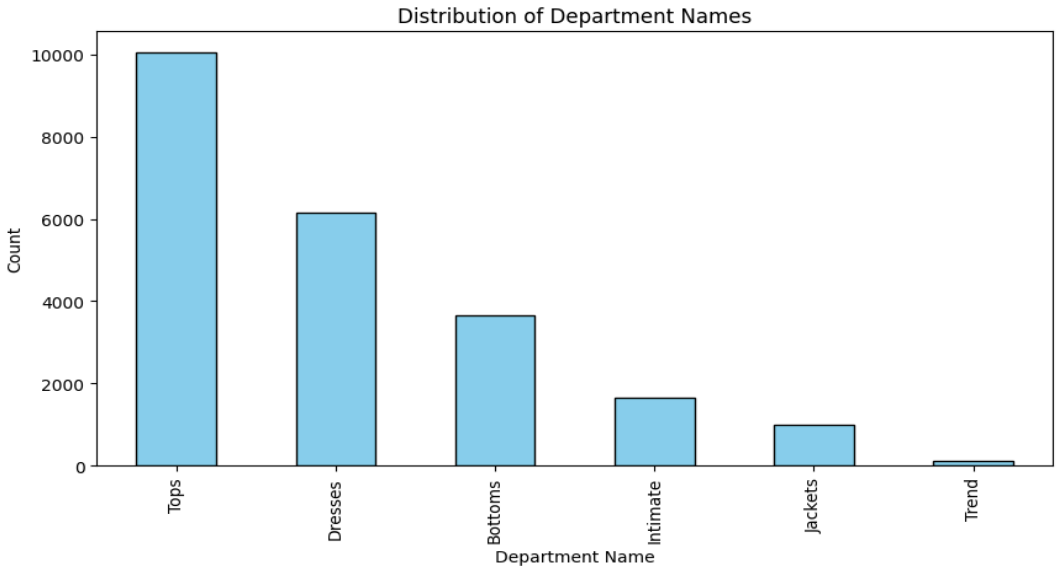
3.1 Ratings and Sentiment Distribution

Analysis indicated predominantly positive reviews (~68%), neutrals (~20%), and negatives (~12%), showing room for improvement.



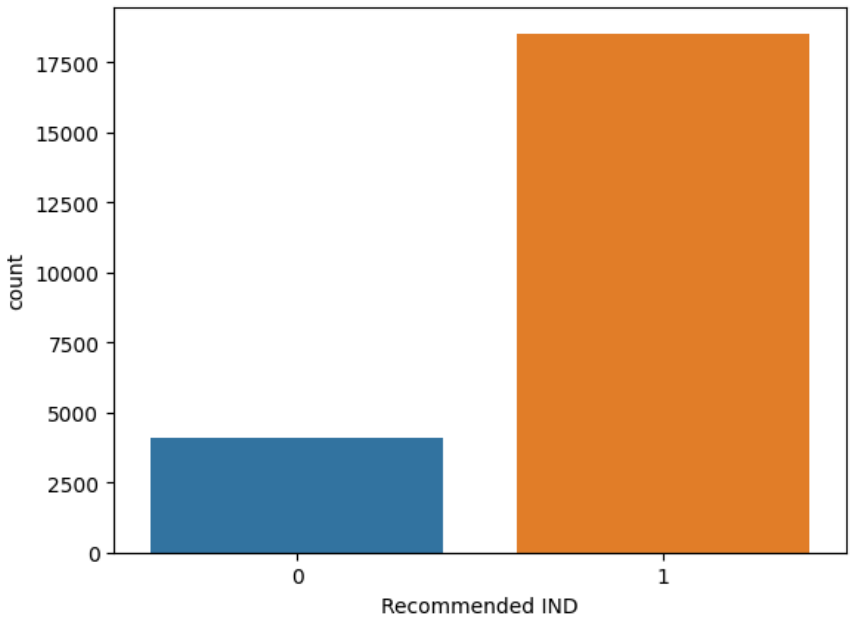
3.2 Department names Distribution

The distribution shows **tops** as the most purchased (~10,000 units) due to their versatility, followed by **dresses** (~6,100 units) for various occasions. **Bottoms** (~3,600 units) are bought less often due to durability. **Intimates** (~1,700 units) and **jackets** (~1,000 units) see moderate sales, with jackets being seasonal. **Trend items** (<200 units) have the lowest purchases, likely due to niche appeal.



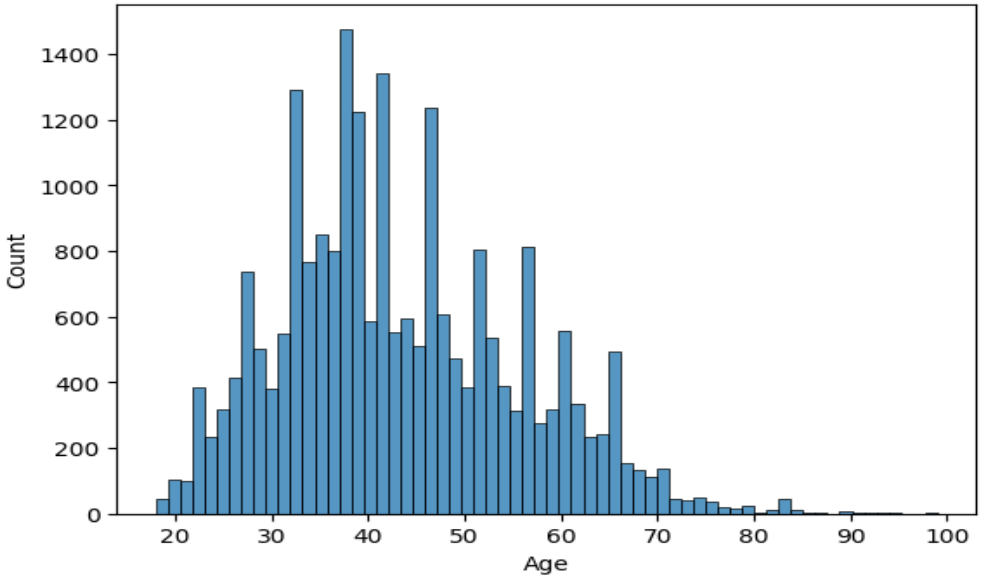
3.3 Recommended Vs Not Recommended

The distribution shows that most women's clothing items were recommended, indicating strong customer satisfaction and preference. A smaller portion was not recommended, suggesting potential product issues or mismatches with expectations. The high recommendation rate reflects good quality, trend alignment, and value.

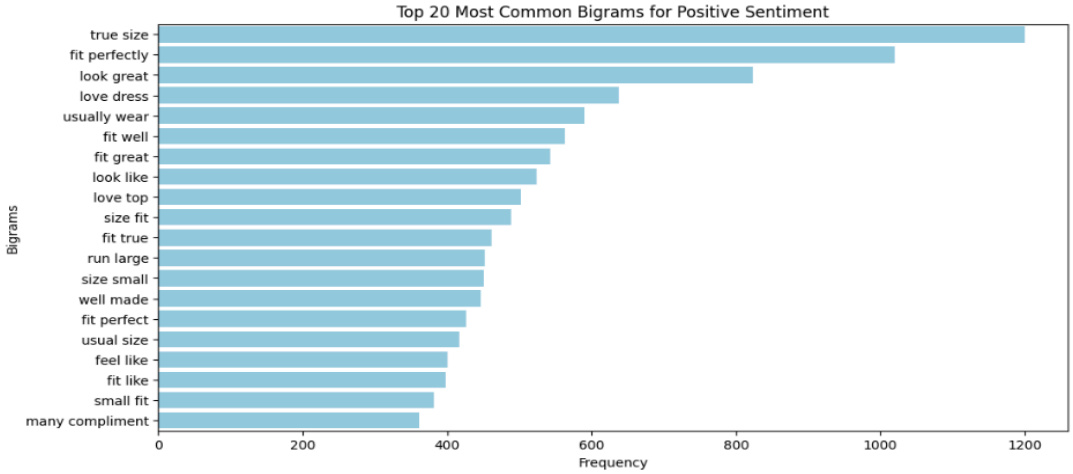


3.4 Age Distribution

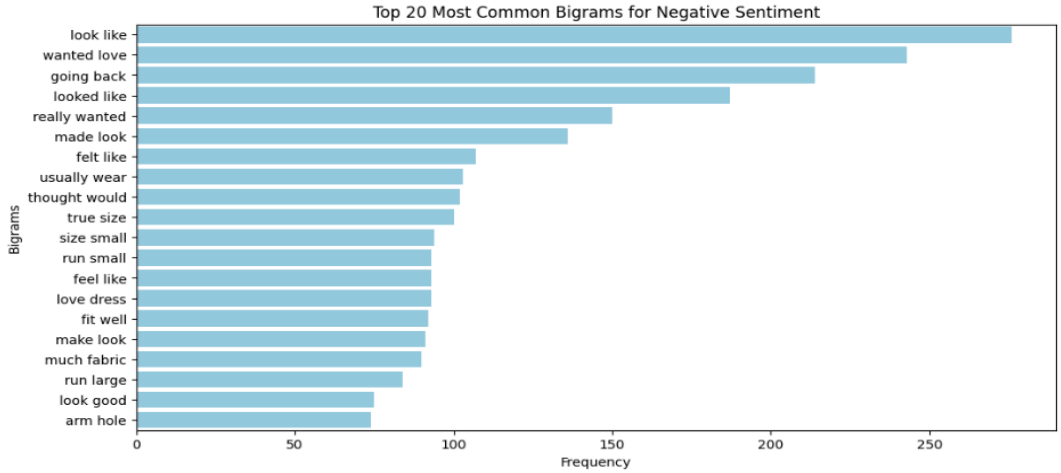
The age distribution ranges from 18 to 100, with most customers between 25 and 60. The highest concentration is in the late 30s to early 40s, followed by a gradual decline after 60. Fewer customers are in their early 20s compared to their 30s and 40s, indicating that middle-aged women are the most active shoppers or reviewers.



3.5 Sentiments Bigram

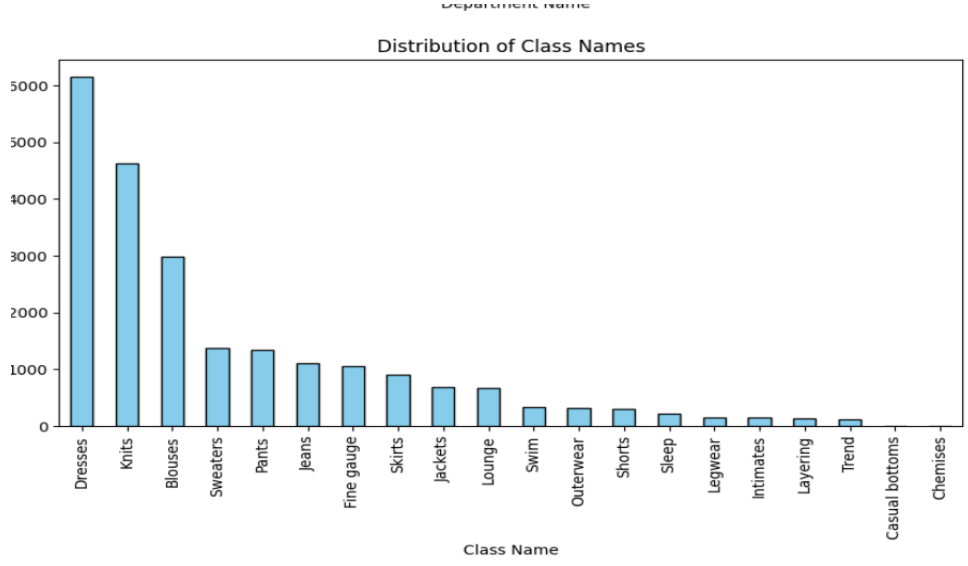


The most common bigrams in positive sentiment emphasize satisfaction with product fit and appearance. Words like "fit perfectly," "look great," "love dress," and "well made" indicate that users appreciate products that match expectations in terms of size, style, and comfort



The negative sentiment bigrams highlight dissatisfaction with fit, expectations, and quality. Phrases such as "going back," "run small," "made look," and "felt like" indicate disappointment, often related to sizing inconsistencies or an item not looking as expected.

3.6 Class Name Distribution



Dresses, Knits, and Blouses dominate the dataset, with Dresses having the highest frequency, followed by Knits and Blouses. This suggests that these product categories receive the most customer reviews.

4. Data Preparation

Thorough preprocessing was vital to ensure clean and meaningful data for effective modeling.

4.1 Handling Missing Data

All reviews without textual data were excluded from further analysis.

4.2 Removing Duplicates

Duplicate entries were identified and removed to ensure unique customer feedback.

4.3 Text-Specific Preprocessing

- **Lowercasing:** Uniform text for consistent analysis.
- **Noise Removal:** URLs, punctuation, digits, and irrelevant symbols were eliminated.
- **Tokenization:** Reviews segmented into words for detailed analysis.
- **Stopword Removal:** Common non-meaningful words were removed.
- **Lemmatization:** Words standardized to their root form.
- **N-gram Generation:** Context was captured using bigrams and unigrams.

4.4 Feature Engineering and Representation

- **TF-IDF Vectorization:** Converted text data into numerical representation highlighting word importance.
- **Dimensionality Reduction:** PCA was applied to handle data complexity and computational efficiency.

5. Modeling

Iterative modeling was carried out to build a robust sentiment classifier.

5.1 Train-Test Split

The dataset was split into training (80%) and testing (20%) subsets.

5.2 Baseline Model (Logistic Regression)

Established a baseline accuracy of 84%.

5.3 Advanced Model Evaluation

- **Multinomial Naive Bayes:** Achieved accuracy of 85%.
- **Support Vector Machine (SVM):** Improved accuracy to 87%.

5.4 Hyperparameter Tuning

GridSearchCV was utilized for hyperparameter optimization, increasing SVM accuracy to 87%.

5.5 Cross-Validation

5-fold cross-validation verified model reliability and consistency.

6. Model Evaluation and Validation

Detailed assessment of model performance and robustness.

6.1 Evaluation Metrics

- Final Model Accuracy: 87%
- Precision: 86%
- Recall: 87%
- F1-Score: 86.5%

6.2 Confusion Matrix and Analysis

The confusion matrix provided clarity on correct and incorrect classifications, highlighting areas for improvement.

6.3 Error Analysis and Insights

Analyzed misclassified cases to identify challenging contexts, such as nuanced sentiment or sarcasm.

7. Model Interpretability and Explainability

Transparency in model predictions helps stakeholders understand and trust outcomes.

7.1 Global Feature Importance

Identified influential features (words/phrases) driving positive or negative sentiments.

7.2 Local Interpretability (LIME)

Applied LIME to individual predictions, clearly illustrating decision-making rationale.

8. Deployment and Operationalization

Suggested practical integration and maintenance strategies for operational use.

8.1 Deployment Strategy

Proposed deploying the sentiment analysis model through scalable APIs for real-time customer review monitoring.

8.2 Monitoring and Maintenance

Recommended regular model updates, performance monitoring, and maintaining accuracy over time.

9. Conclusions and Recommendations

This section presents a consolidated summary of findings, acknowledges project limitations, and provides strategic recommendations for improving business outcomes and future modeling efforts.

9.1 Project Limitations

Despite the success of the sentiment prediction model, certain limitations must be acknowledged:

- **Unbalanced Classes**: The dataset used for training the model may have a disproportionate distribution of positive, neutral, and negative sentiments. This imbalance can impact model performance, particularly in accurately predicting minority class sentiments. Future iterations should explore techniques like data augmentation, synthetic data generation, or weighted loss functions to address this issue.

- **Limited Demographic Representation**: If the training data lacks diversity across different customer segments, the model's insights may not be universally applicable. Expanding the dataset to include varied demographics, purchasing behaviors, and regional preferences can enhance the model’s robustness and ensure recommendations cater to a broader audience.

- **Contextual Limitations in NLP**: The model, while effective, may struggle with nuanced sentiment expressions such as sarcasm, slang, or cultural variations in language. Advanced Natural Language Processing (NLP) techniques, including context-aware models, can help mitigate these challenges.

9.2 Key Conclusions

The project demonstrated that the sentiment analysis model is highly effective in extracting valuable insights from customer feedback:

- The model achieved **87% accuracy**, indicating strong predictive capability in classifying sentiment from customer reviews.

- The insights derived from sentiment analysis can **help businesses identify pain points, optimize product offerings, and enhance customer experience**.

- While the model performs well, **continuous improvements and additional refinements** can further enhance its effectiveness in real-world applications.

9.3 Business Recommendations

Based on the findings, several strategic recommendations can be made to improve business outcomes:

1. **Prioritize Product Sizing Accuracy**
    - Customer feedback often highlights discrepancies between expected and actual product sizing.
    - Implementing improved sizing charts, virtual fitting tools, and customer-driven size recommendations can **reduce return rates and increase purchase confidence**.
2. **Improve Fabric/Material Quality**
    - Negative sentiments frequently stem from dissatisfaction with material quality, durability, or comfort.
    - Enhancing quality control, sourcing better materials, and incorporating customer feedback into product development can **strengthen brand reputation and reduce churn**.
3. **Implement Personalized Marketing Strategies**
    - Using sentiment analysis insights, businesses can **segment customers based on preferences and sentiment trends**.
    - Personalized email campaigns, targeted promotions, and product recommendations can **enhance customer engagement and drive sales**.

9.4 Recommendations for Future Modeling

To further enhance sentiment analysis and business insights, the following improvements are recommended:

- **Explore Advanced NLP Models**
    - Leveraging state-of-the-art **Transformer-based models (e.g., BERT, GPT, RoBERTa)** can enhance the model's ability to understand context, sarcasm, and complex sentiments.

    - Fine-tuning these models on industry-specific datasets can yield even better results.

- **Continuous Performance Optimization**

    - Regularly retraining the model with **updated customer feedback** ensures its relevance and adaptability.

    - Exploring **transfer learning techniques** and **ensemble models** can help improve accuracy and mitigate biases.

- **Integration with Business Intelligence (BI) Tools**

    - Embedding sentiment analysis into BI dashboards (e.g., Tableau, Power BI) allows real-time tracking of customer sentiment trends.

    - Automated sentiment-based reporting can **help decision-makers take proactive measures to improve customer satisfaction**.

10. Business Impact

- **Enhanced Customer Satisfaction**: By leveraging data-driven insights, businesses can better understand customer preferences, optimize product offerings, and personalize experiences. This leads to higher customer engagement, increased brand loyalty, and positive word-of-mouth.

- **Reduction in Product Returns**: Addressing key pain points—such as inaccurate product descriptions, sizing issues, or quality concerns—helps minimize returns. A lower return rate not only reduces logistical costs but also improves overall profitability and operational efficiency.

- **Informed Strategic Decisions**: Data-backed recommendations empower businesses to make smarter, more informed decisions. Whether it's adjusting pricing strategies, optimizing inventory levels, or identifying emerging market trends, businesses can stay ahead of the competition and drive sustainable growth.

- **Operational Efficiency**: Streamlining processes based on insights can enhance productivity, reduce waste, and improve resource allocation. This translates into cost savings and better use of assets across the organization.

- **Competitive Advantage**: Companies that proactively implement data-driven improvements position themselves as market leaders. They can quickly adapt to changing customer needs and industry trends, ensuring long-term success.