Project 2: Sentiment Analysis on Women's E-commerce Clothing Reviews

Course: MAI623 & DSC514: Natural Language Processing

Instructor: Prof. Demetris Paschalides
Department of Computer Science
University of Cyprus

Group Member 1
Name: Ahsan Fayyaz
Student Id: BB8119341
Email: fayyaz.ahsan@ucy.ac.cy

Group Member 2
Name: Hafiz Muhammad Soban Khan
Student Id: UZ4145952
Email: khan.hafiz-muhammad-soban@ucy.ac.cy

Group Member 3
Name: Mehak Mushtaq Malik
Student Id: FX0767161
Email: malik.mehak@ucy.ac.cy

Abstract

This project performs an extensive Exploratory Data Analysis (EDA) and develops a sentiment analysis model using a rich dataset of women's clothing reviews from an online retail platform. The principal objective is to get useful information that helps online businesses in better understanding customer sentiments, preferences, and trends. By achieving this, companies can more effectively modify their products and services to meet consumer demands. The dataset utilized in this study features multiple key attributes such as clothing ID, review texts, ratings, and customer recommendations. Additional data points include the age of the reviewer, the division, department, and class of the products reviewed, which allows for a multifaceted analysis of consumer behavior and trends across different demographic and product segments. Our methodology begins with an extensive data cleaning process. Following data preparation, the project analyzes the distribution of reviews across different age groups, product categories, and ratings, providing a detailed demographic analysis of customer feedback. We employ advanced Natural Language Processing (NLP) techniques to manipulate and process textual data from reviews. Furthermore, the use of word embeddings helps in capturing deeper semantic meanings of words, which enhances the quality of sentiment classification. The sentiment analysis model utilizes a variety of machine learning algorithms to evaluate performance using metrics such as accuracy, precision, recall, and F1-Score. This assessment allows us to determine the most effective approach for sentiment classification in the context of e-commerce reviews. Finally, the outcomes of this analysis not only highlight significant trends in customer feedback but also offer actionable insights that are critical for enhancing customer satisfaction and optimizing business strategies. The findings demonstrate the potential of machine learning and NLP in transforming how businesses interact with and respond to customer needs, ultimately contributing to a more personalized shopping experience.

Contents

1 Introduction

Outfits play a big role in defining a woman's character, highlighting her beauty, and shaping her unique identity. What she wears is more than just fabric; it's her way of expressing herself and showcasing her lifestyle. The main challenge addressed in this project is associated with the necessity for a systematic approach to analyze the sentiments expressed in reviews of women's clothing by customers. Through the utilization of data derived from an e-commerce platform, the project aims to reveal valuable understandings regarding customer sentiments, identify trends, and assess the influence of customer reviews on product recommendations and sales.
Sentiment analysis has emerged as a crucial tool in the e-commerce, enabling online businesses. It helps to understand and enhance the customer interactions through the evaluation of sentiments conveyed in reviews generated by users. Within the domain of the fashion industry, particularly in women's clothing, sentiment analysis offers unique insights into customer preferences, levels of satisfaction, and general market trends.
The objectives of this project are:

- To perform an extensive Exploratory Data Analysis (EDA) on a dataset of women's e-commerce clothing reviews by focusing on the distribution of reviews across different people, product categories, and ratings. This includes investigating the relationship between review sentiments, product recommendations, and the count of positive feedback.

- To develop and implement a sentiment analysis model that can classify reviews into positive, neutral, or negative categories and to assess the model's performance using metrics such as accuracy, precision, recall, and F1-Score.

This project utilizes a dataset from Kaggle that includes several key fields such as Clothing ID, Age, Title, Review Text, Rating, Recommended IND, Positive Feedback Count, and product categorization. The scope of the analysis is limited to this dataset, and the insights derived are specific to the customer base of the e-commerce platform from which the data was collected.

2 Data Description

The dataset used in this project consists of 22,641 rows and 10 columns, derived from customer reviews of women's clothing on an e-commerce platform. The dataset, obtained from Kaggle, represents real commercial data that has been anonymized; any references to the specific retailer have been replaced with ”retailer” in the review text and titles to maintain confidentiality.

2.1 Dataset Structure

Each row in the dataset corresponds to an individual customer review, containing both textual and categorical information about the review and the product. The key variables included in the dataset are as follows:

- **Clothing ID**: A unique identifier for each product.
- **Age**: The age of the reviewer.
- **Title**: The title of the review.
- **Review Text**: The detailed text of the review.
- **Rating**: The rating given to the product by the reviewer, on a scale from 1 to 5.
- **Recommended IND**: A binary indicator (1 for recommended, 0 for not recommended) showing whether the reviewer recommends the product.
- **Positive Feedback Count**: The number of other customers who found the review helpful.
- **Division Name**: The high-level division under which the product is categorized.
- **Department Name**: The department under which the product is categorized.
- **Class Name**: The class name under which the product is categorized.

2.2 Data Modifications

During the preprocessing phase, a new column named **Merged Review** is created by merging the **Title** and **Review Text** columns. This modification is intended to provide a more comprehensive text input for sentiment analysis by combining the titles with the detailed review texts which potentially enriches the textual data available for analysis. This dataset serves as the foundation for both the EDA and the sentiment analysis model developed in this project.

3 Exploratory Data Analysis (EDA)

3.1 Data Cleaning and Preprocessing

The initial step in our exploratory data analysis involved a thorough data cleaning and preprocessing phase, which was essential for ensuring the quality and reliability of the dataset. This phase focused on dealing with common issues of duplicates and missing values that can impact analysis and modeling.

3.1.1 Handling Missing Values and Duplicates

Before preprocessing, our dataset contained several missing values and duplicates. We have detected that there were no duplicates in the dataset.
There were many missing values in several key columns with varying degrees of absence. The *Title* field had the most significant number of missing entries, totaling 3810. This was followed by the *Review Text* field which had 845 missing entries. Minor missing data were found in the *Division Name*, *Department Name*, and *Class Name*, each having 14 missing entries. To address these, we employed the following strategies:

- Titles with missing values were filled with the placeholder 'No Title' to maintain the integrity of the dataset and ensure that all reviews could be utilized in our analysis.
- Rows missing the *Review Text* were completely removed, considering the critical nature of this field for our sentiment analysis.
- For missing entries in *Division Name*, *Department Name*, and *Class Name*, we substituted 'Unknown' to indicate the absence of specific categorization data.

These steps were crucial for cleaning our data, and the impact was immediate and measurable. After the cleaning process, we reassessed the dataset to confirm the elimination of all duplicate entries and the appropriate handling of missing values. The results were satisfactory, showing a significant reduction in missing data, which allowed us to proceed with a more accurate and comprehensive analysis.

3.1.2 Data Merging

In addition to handling missing values and duplicates, we merged the *Title* and *Review Text* columns into a new column named **Merged Review**. This was done to enrich the text data available for sentiment analysis which provided a more comprehensive context for our NLP models to analyze. The merged text combines the concise descriptions in titles with the detailed content in review texts, offering a richer dataset for extracting sentiments. Following this, our dataset was well-prepared, with a structured format ready for in-depth exploratory analysis and sentiment modeling.

3.2 Distribution Analysis

Understanding the distribution of reviews across various demographic and product segments is crucial for identifying patterns in consumer behavior. We employed visual analysis techniques to explore the distribution of reviews across age groups, product divisions, departments, classes, and ratings.

3.2.1 Age Groups

The analysis of reviews across age groups revealed significant trends in the demographics of the reviewers. As illustrated in the accompanying bar chart, the majority of the reviews were contributed by customers in the age groups of ‘26-55‘ where ‘36-45‘ has highest number of reviews of more than 7000, which collectively accounted for the highest proportion of reviews. This indicates a strong engagement with the platform among middle-aged adults. Notably, the ‘18-25‘ age group showed the least engagement, suggesting either a lower interest to leave reviews or less frequent use of the platform by this demographic.

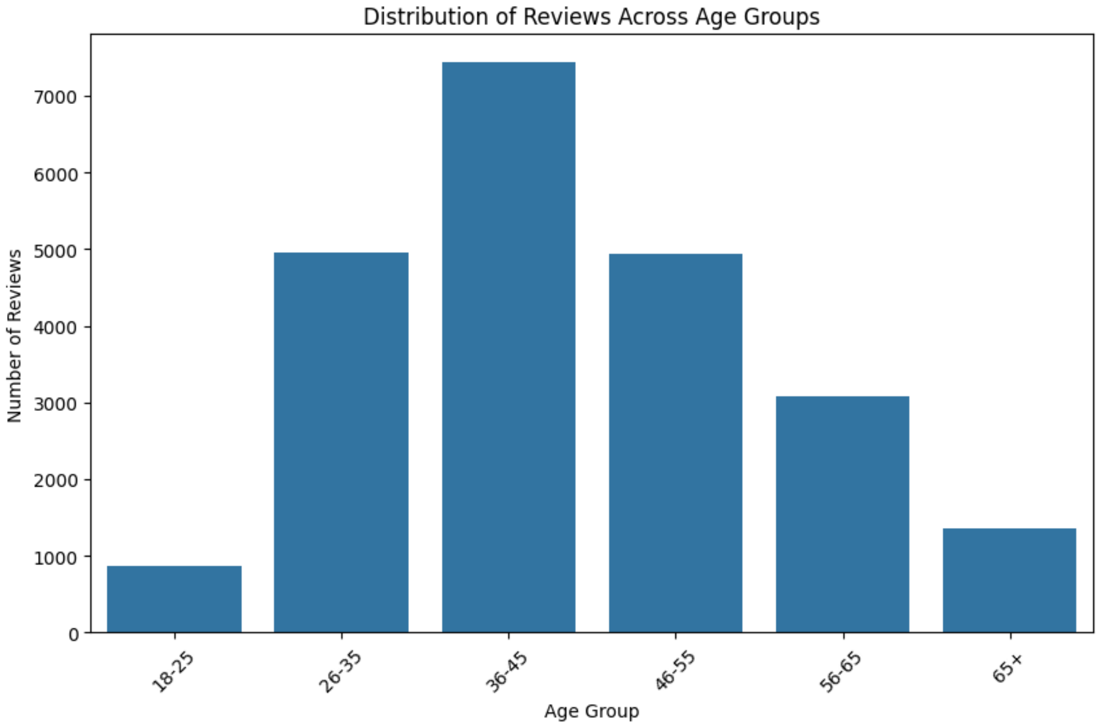

Figure 1: Distribution of Reviews Across Age Groups

3.2.2 Divisions

In terms of product divisions, the ‘General‘ division got the most reviews, followed by the ‘General Petite‘ division. The ‘Intimates‘ division received comparatively fewer reviews. This distribution suggests that the General division’s offerings are most popular among the users of the women’s e-commerce clothing platform which reflects broader consumer preferences within this online community.

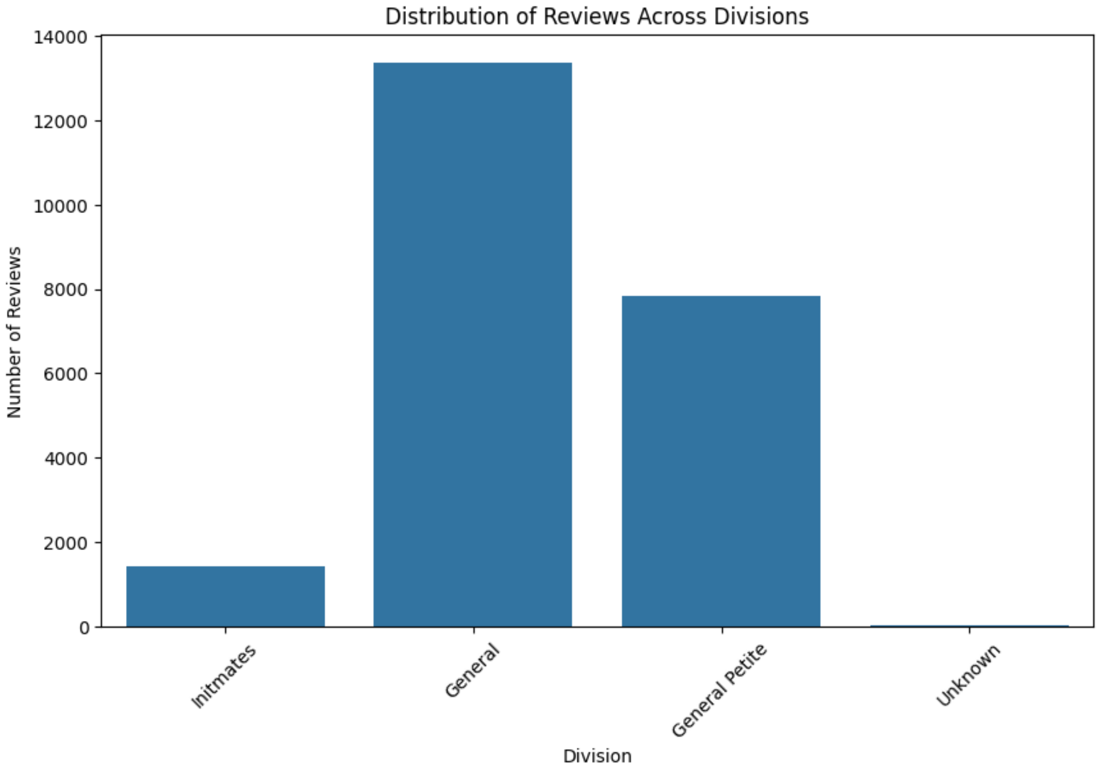

Figure 2: Distribution of Reviews Across Divisions

3.2.3 Departments

The review distribution across departments highlighted that the ‘Tops‘ department received the highest number of reviews, followed by ‘Dresses‘ and ‘Bottoms‘. This trend could indicate a higher consumer interest or satisfaction in these categories, possibly due to a wider variety of offerings or higher purchase rates in these departments.

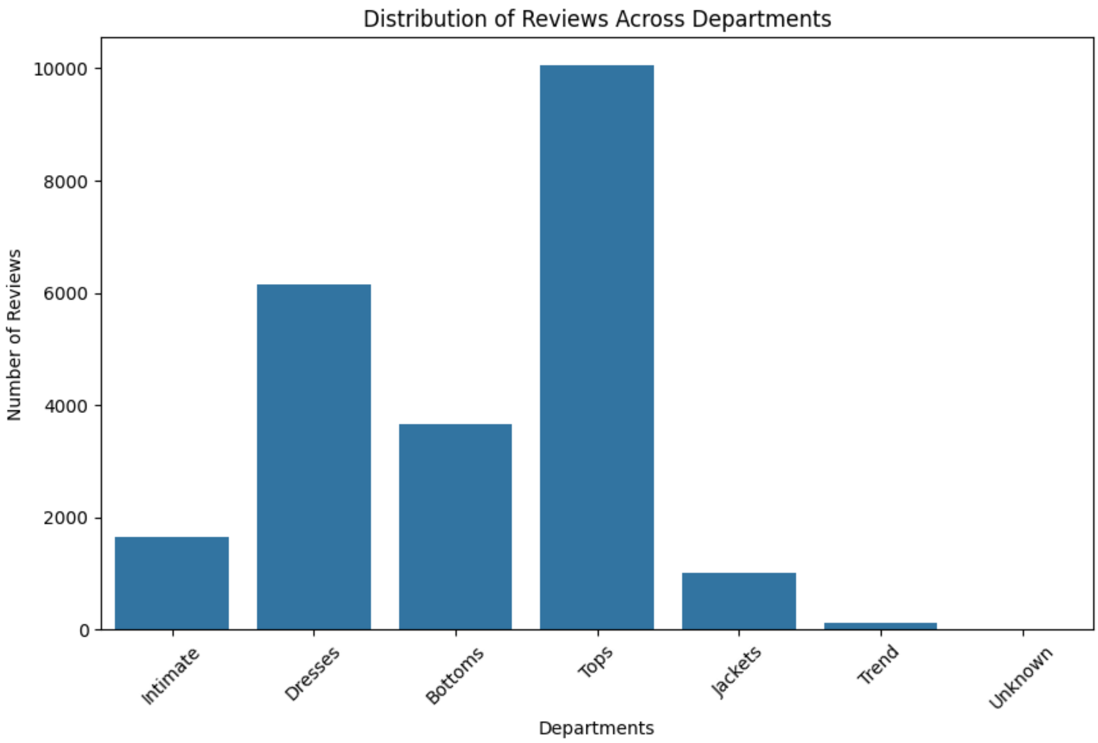

Figure 3: Distribution of Reviews Across Departments

3.2.4 Classes

The analysis extended into product classes where ‘Dresses‘, ‘Blouses‘, and ‘Knits‘ received the most attention in terms of reviews. The detailed exploration into classes helps in understanding specific product types that are most popular among consumers, guiding potential inventory and marketing strategies for the retailer.

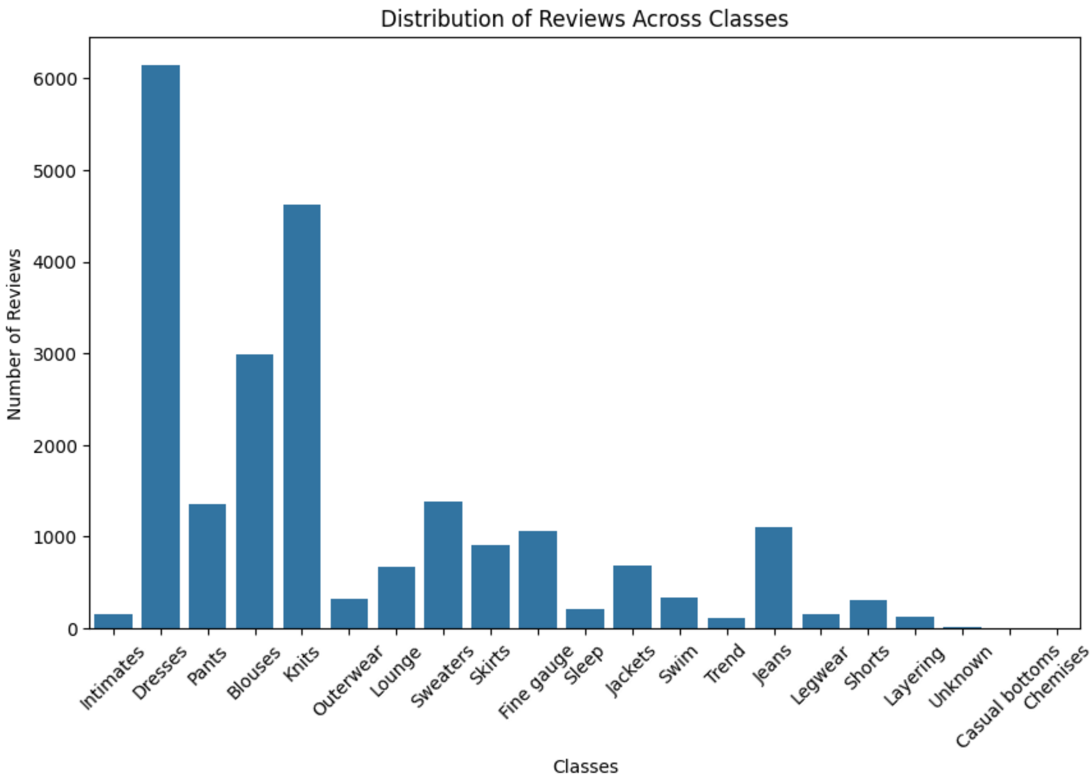

Figure 4: Distribution of Reviews Across Classes

3.2.5 Ratings

The review distribution skewed towards higher ratings, mostly rated as '5'. This positive skew suggests overall customer satisfaction with the products offered. However, the presence of reviews across the spectrum indicates diverse consumer experiences and expectations.

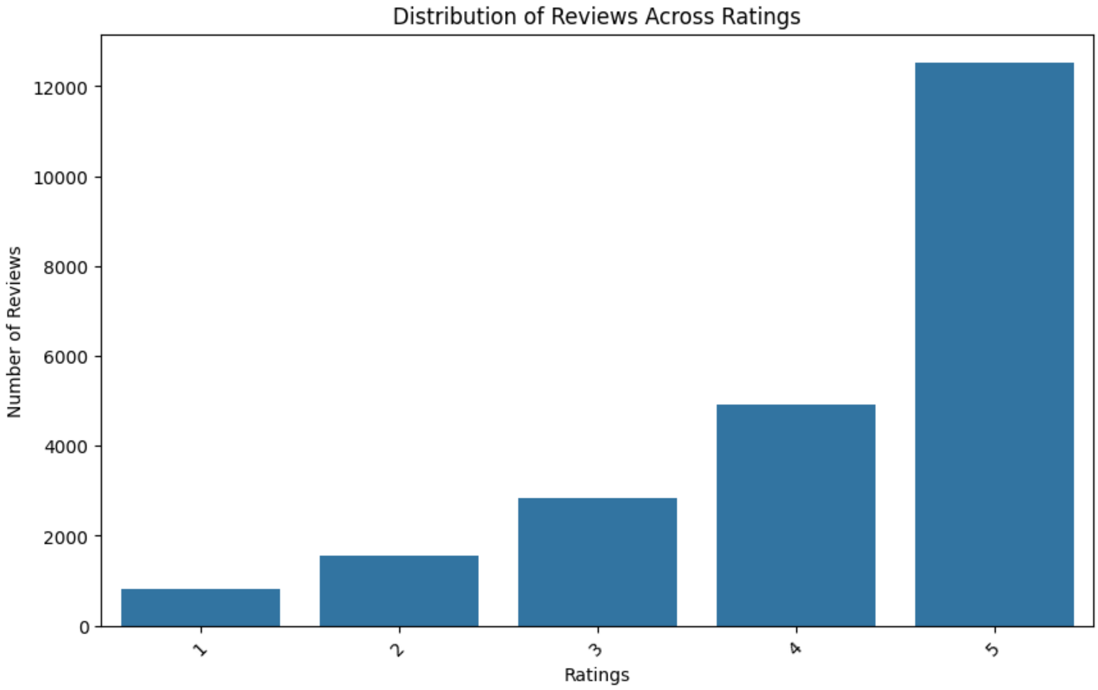

Figure 5: Distribution of Reviews Across Ratings

Each of these findings provides valuable insights into the consumer base of the women's e-commerce clothing platform. By understanding these distributions, the retailer can tailor marketing and product strategies to better align with customer preferences and behaviors. The visual representations of these distributions were crucial in illustrating these trends.

3.3 Relationship Investigation

To further understand the dynamics between customer sentiments, product recommendations, and the influence of positive feedback on reviews, we performed a correlation analysis. This analysis aimed to uncover any significant relationships that could provide insights into how sentiments and recommendations affect customer engagement and feedback mechanisms.

3.3.1 Correlation Analysis

Using Python's `TextBlob` library, we computed sentiment scores for each merged review. These sentiment scores were then categorized as positive, neutral, or negative based on predefined thresholds. The analysis primarily focused on exploring how these sentiment categories correlate with whether the product was recommended (*Recommended IND*) and the count of positive feedback received by each review.

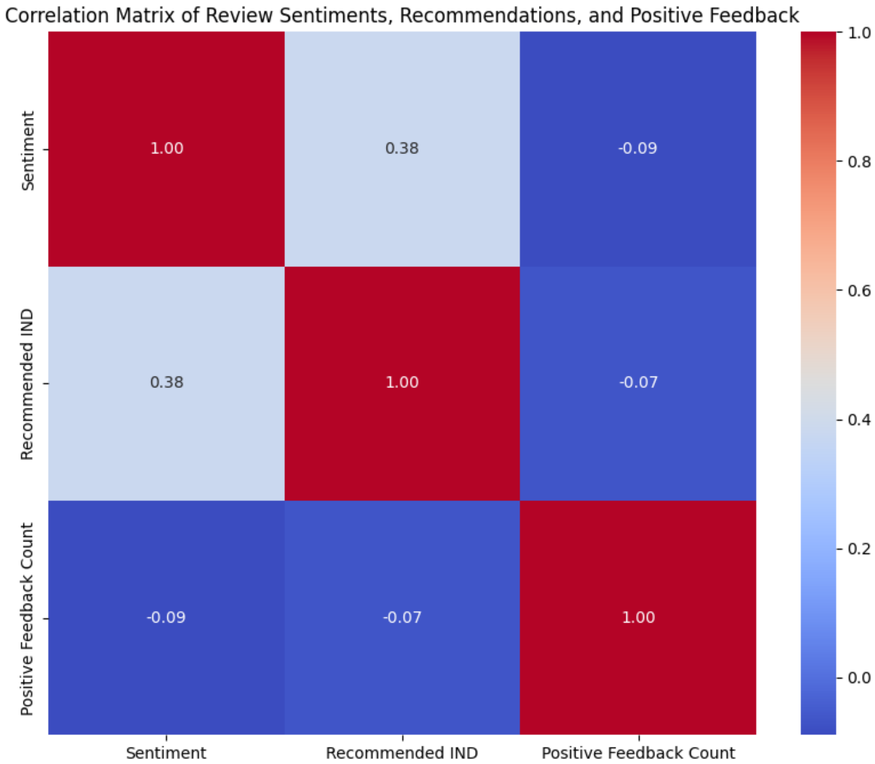

Figure 6: Correlation Analysis

The correlation matrix, visualized through a heatmap, revealed the following key relationships:

- **Sentiment and Recommendations**: There is a moderate positive correlation (`0.38`) between sentiment scores and product recommendations. This indicates that reviews with more positive sentiments are likely to recommend the product, suggesting that emotional tone in reviews significantly impacts the recommendation behavior of customers.

- **Sentiment and Positive Feedback Count**: The correlation between sentiment and positive feedback count was slightly negative (`-0.09`), although this relationship is very weak. This suggests that more positive sentiments in reviews do not necessarily correlate with higher numbers of positive feedback, indicating that other factors may influence the reception of feedback.

- **Recommendations and Positive Feedback Count**: Similarly, the relationship between product recommendations and positive feedback count was also weakly negative (`-0.07`). This unexpected finding could imply that recommendations are not the primary driver for other users finding a review helpful.

These insights help in understanding the outcomes relationships between the emotional content of reviews, customer endorsements of products, and the community response in terms of positive feedback. The outcomes of these findings are significant for e-commerce platforms as they suggest areas for enhancing customer interaction strategies and review management practices.

3.4 Visualization of Common Words

To gain deeper insights into the sentiments expressed in customer reviews, we utilized WordCloud visualizations, which provide a graphical representation of word frequency. Words that appear more frequently in the review text are given greater prominence in the WordCloud, thus allowing us to visually perceive the most common themes and terms used in positive and negative reviews.

3.4.1 WordCloud Analysis

Two separate WordClouds were generated for reviews categorized as positive and negative based on their sentiment scores. This categorization was accomplished using the sentiment analysis functionality of Python's `TextBlob` library.

Figure 7: WordCloud Analysis

Positive Reviews: The WordCloud for positive reviews prominently features words such as "perfect," "love," "great," and "comfortable," suggesting that these are key factors contributing to positive customer experiences. The frequent appearance of words like "fabric," "color," and "fit" indicates that these aspects are particularly appreciated in the products reviewed.

Negative Reviews: On the other hand, the WordCloud for negative reviews frequently includes words such as "disappointed," "small," "return," and "poor." These highlight common issues faced by customers, primarily related to sizing, expectations not being met, and poor quality of the products. The occurrence of words like "tight" and "color" also suggests dissatisfaction with how the products fit and their visual appeal.

These visualizations are crucial in identifying the most significant factors influencing customer satisfaction and dissatisfaction. By understanding these key themes, retailers can better address the concerns highlighted in negative reviews and reinforce the positive attributes that establish a strong connection with customers.

4 Sentiment Analysis Model

4.1 Preprocessing for Model Ingestion

Effective preprocessing is essential for refining raw text data into a structured format suitable for sentiment analysis. Our preprocessing pipeline involves several critical steps designed to normalize and enhance the text data, preparing it for accurate sentiment classification.

4.1.1 Text Normalization

We started by expanding contractions (e.g., changing ”can’t” to ”cannot”) to standardize the text and make it more analyzable. This was achieved using regular expressions that methodically replaced contracted forms with their full forms across the dataset.

4.1.2 Tokenization and Cleaning

After normalization, the text was tokenized, breaking it down into individual tokens. This process allows for the filtering of punctuation and numbers, which are removed to reduce noise in the data.

4.1.3 Stop Word Removal and Lemmatization

We also eliminated commonly used stopwords, which do not contribute significantly to sentiment analysis. The tokens were then lemmatized and converted to their base or dictionary form. This step utilizes part-of-speech tagging to accurately identify and transform words and ensuring that the text data retains its contextual meaning.

4.2 Text Representation Techniques

To prepare the text data for sentiment analysis, three different text representation techniques were employed: Count Vectorization, TF-IDF Vectorization, and Word Embeddings. Each technique transforms raw text into a numerical format that machine learning algorithms can interpret, but they do so in uniquely beneficial ways. Beyond the basic vectorization methods, BERT (Bidirectional Encoder Representations from Transformers) was also utilized.

4.2.1 Count Vectorization

Count Vectorization converts text documents into a matrix of token counts, effectively transforming the dataset into a 'bag of words'. This method counts the frequency of each word or phrase from a predefined dictionary of words (n-grams) that occur in the dataset. It is a basic form of text representation where the context and order of words are ignored, but it provides a robust initial approach for analyzing text data.

4.2.2 TF-IDF Vectorization

Term Frequency-Inverse Document Frequency (TF-IDF) builds on Count Vectorization by reducing the weight of tokens that occur very frequently, thus diminishing their significance in the model. It increases the importance of words that occur in fewer documents, which can be more informative. TF-IDF helps to adjust for the fact that some words appear more frequently in general.

4.2.3 Word Embeddings

Word Embeddings provide a dense representation of words and their relative meanings. They involve transforming textual information into vectors where each word is represented by a point in the embedding space. This technique captures more context and relationships between words, making it powerful for many NLP tasks. For our analysis, embeddings were generated using the ‘Sentence-Transformers‘ library which uses pre-trained models based on GloVe (Global Vectors for Word Representation).

4.2.4 BERT Embeddings

BERT models are pre-trained on a large corpus of text and then fine-tuned for specific tasks, which in our case, involves generating embeddings for sentiment analysis. BERT considers the entire context of a word by looking at the words that come before and after it, which is particularly powerful for understanding complex language in customer reviews. These embeddings were generated using the SentenceTransformer implementation of BERT, specifically the ”all-MiniLM-L6-v2” model which is optimized for greater speed and efficiency.

Each of these techniques was applied to our preprocessed, lemmatized sentences to convert them into formats suitable for training our sentiment analysis models. The effectiveness of each method in capturing the instances of sentiment in text will be evaluated and compared in next sections of this analysis.

4.3 Model Training and Evaluation

The Logistic Regression classifier was trained using various text representations to evaluate their effectiveness in sentiment analysis. The performance of each model was measured using precision, recall, f1-score, and accuracy, which are critical metrics in classification tasks.

4.3.1 Model Performance

- **CountVectorizer**: Achieved an accuracy of 90% with weighted averages for precision, recall, and f1-score at 0.89, 0.90, and 0.89 respectively. The model showed a balanced performance across different classes, though it struggled slightly with the 'Negative' class.

	precision	recall	f1-score	support
Negative	0.42	0.36	0.38	70
Neutral	0.64	0.52	0.58	591
Positive	0.93	0.96	0.95	3868
accuracy			0.90	4529
macro avg	0.66	0.61	0.64	4529
weighted avg	0.89	0.90	0.89	4529

Figure 8: Classification Report of Count Vectorization

- **TF-IDF**: Similar to CountVectorizer in accuracy at 90%, but showed varied performance in recall, especially lower in the 'Negative' class, which could indicate overfitting on the dominant 'Positive' class.

	precision	recall	f1-score	support
Negative	0.57	0.06	0.10	70
Neutral	0.68	0.42	0.52	591
Positive	0.92	0.98	0.95	3868
accuracy			0.90	4529
macro avg	0.72	0.49	0.52	4529
weighted avg	0.88	0.90	0.88	4529

Figure 9: Classification Report of TF-IDF Vectorization

- **Word Embeddings**: Posted an accuracy of 88% with the highest recall for the 'Positive' class but lower performance metrics for 'Negative' and 'Neutral' classes, suggesting a potential bias toward more frequent labels.

	precision	recall	f1-score	support
Negative	0.83	0.21	0.34	70
Neutral	0.56	0.34	0.42	591
Positive	0.90	0.97	0.94	3868
accuracy			0.88	4529
macro avg	0.77	0.51	0.57	4529
weighted avg	0.86	0.88	0.86	4529

Figure 10: Classification Report of Word Embeddings

- **BERT Embeddings**: Also recorded an 88% accuracy with similar challenges as word embeddings, highlighting the complexity of capturing sentiment in less frequent classes.

	precision	recall	f1-score	support
Negative	0.65	0.16	0.25	70
Neutral	0.57	0.34	0.42	591
Positive	0.90	0.97	0.94	3868
accuracy			0.88	4529
macro avg	0.71	0.49	0.54	4529
weighted avg	0.86	0.88	0.86	4529

Figure 11: Classification Report of BERT Embeddings

These results suggest that while advanced embeddings capture richer linguistic complexities, traditional methods like CountVectorizer and TF-IDF remain competitive, particularly in datasets with imbalanced classes.

4.3.2 Classifier Selection Using Lazypredict

To determine the most suitable classifier for our sentiment analysis model using CountVectorizer, we employed lazypredict, a library that quickly fits and compares a wide range of classification models. Logistic Regression was chosen based on its performance and characteristics ideal for our application.

Why Logistic Regression?

Logistic Regression was selected for several reasons:

- **Performance:** It achieved high and balanced accuracy, making it effective for our imbalanced dataset.
- **Interpretability:** Logistic Regression offers clear interpretability, which is crucial for understanding feature influence on the prediction outcomes—essential in sentiment analysis to trace back what drives positive, neutral, or negative sentiments.
- **Efficiency:** It is computationally efficient compared to more complex models, ensuring quick training and prediction phases, suitable for real-time analysis applications.

Below is a detailed table showcasing the performance of various classifiers evaluated through lazypredict, highlighting why Logistic Regression stood out.

Table 1: Performance Comparison of Various Classifiers using Lazypredict

Model	**Accuracy**	**Balanced Accuracy**	**F1 Score**	**Time Taken (s)**
BernoulliNB	0.83	0.64	0.84	1.22
NearestCentroid	0.76	0.63	0.80	1.07
Logistic Regression	0.89	0.62	0.89	12.44
LinearSVC	0.88	0.59	0.88	102.19
GaussianNB	0.61	0.59	0.69	1.33
SGDClassifier	0.89	0.57	0.88	34.94
Perceptron	0.88	0.56	0.87	3.17
LinearDiscriminantAnalysis	0.87	0.55	0.87	9.12
PassiveAggressiveClassifier	0.88	0.55	0.88	5.43
LGBMClassifier	0.89	0.54	0.88	5.97
BaggingClassifier	0.86	0.49	0.85	44.74
DecisionTreeClassifier	0.82	0.48	0.82	11.80
AdaBoostClassifier	0.88	0.48	0.86	32.83
ExtraTreeClassifier	0.80	0.46	0.80	1.58
CalibratedClassifierCV	0.89	0.46	0.87	382.84
RidgeClassifier	0.88	0.43	0.86	2.10
RidgeClassifierCV	0.88	0.43	0.86	8.48
QuadraticDiscriminantAnalysis	0.76	0.41	0.78	10.21
SVC	0.87	0.40	0.84	391.14
KNeighborsClassifier	0.85	0.38	0.81	6.02
RandomForestClassifier	0.86	0.36	0.81	19.31
ExtraTreesClassifier	0.86	0.36	0.81	40.63
LabelPropagation	0.02	0.33	0.01	32.19
LabelSpreading	0.02	0.33	0.01	43.83
DummyClassifier	0.85	0.33	0.79	0.95

5 Conclusions and Future Work

5.1 Summary of Findings

This project conducted an in-depth exploratory data analysis (EDA) and developed a sentiment analysis model that utilized various advanced text representation techniques to understand consumer sentiments on women's clothing from an e-commerce platform. Key insights revealed:

- Dresses, blouses, and knits are the most reviewed classes, indicating a high consumer interest and satisfaction. This is likely due to the broad variety and the essential nature of these items in women's fashion.
- The most active reviewers are middle-aged adults (36-45), which suggests that marketing efforts targeted at this demographic could be particularly effective. The data also showed that these age groups likely find these products highly relevant to their lifestyle needs.
- Intimates and other specific classes like trousers received fewer reviews, which might indicate lesser popularity or issues with these products that deter high review volumes.
- Negative reviews often mentioned issues like poor fit, disappointment with product quality, and mismatch with expectations (color, size). These areas are critical for improvement.
- Logistic Regression used in sentiment analysis, showed a robust performance especially with features extracted via CountVectorizer, balancing efficiency with interpretability.

5.2 Limitations

Several limitations were encountered during this analysis:

- The dataset was limited to one e-commerce platform, which may not generalize across the broader retail industry.

- Missing data in critical fields like review text could have biased the sentiment analysis towards users who are more engaged or have more pronounced opinions.
- The classifiers struggled with the imbalanced dataset, particularly in accurately classifying less frequent negative and neutral sentiments.

5.3 Recommendations for Future Research

To build on the findings of this project, future research could explore:

- Integrating additional data sources such as social media, to enrich the dataset and possibly enhance the model's generalizability across different platforms.
- Employing more sophisticated ML models such as deep learning techniques which might improve the accuracy, especially in under-represented classes.
- Conducting extended observations to observe how sentiment trends evolve over time in response to changing fashion trends and market dynamics.

Further, exploring sentiment analysis with real-time data could provide dynamic insights into customer behavior, enabling businesses to react more promptly to emerging trends. This approach would help in changing marketing strategies more effectively and potentially increasing customer satisfaction.

The continuous advancement in NLP and ML techniques presents an exciting opportunity to enhance sentiment analysis tools and making them more adaptable to the complex and ever-evolving language of customer reviews. By addressing the limitations and implementing the suggested future work the effectiveness and applicability of sentiment analysis in e-commerce can be significantly improved.