Randal Burks
April 16, 2025
CMSE 802

Sentiment Classification and Topic Modeling of Women's Clothing Reviews Using TF-IDF, Logistic Regression, and LDA

I completed a comprehensive project analyzing customer reviews from a women's clothing e-commerce dataset. The primary goal was to learn about hidden patterns in customer satisfaction using natural language processing and machine learning. Most online shopping platforms rely on star ratings to summarize customer experiences, but these ratings often lack nuance and may not accurately reflect the tone or sentiment expressed in the written review. For instance, a customer might give a product five stars while expressing frustration in the review text due to size or delivery issues. I wanted to go beyond the surface-level star ratings and explore how textual sentiment could be used to gain deeper insights into customer opinions. By classifying sentiment, comparing it to ratings, clustering customer behaviors, and identifying frequently discussed topics, I aimed to give a more complete picture of how customers felt about their purchases.

To begin, I prepared the dataset by performing data cleaning. I removed all reviews that had missing text because those couldn't be analyzed using text-based techniques. To ensure robustness, I implemented basic error handling throughout the data pipeline. I explicitly dropped reviews with missing or null values in the "Review Text" column to avoid issues during vectorization. I validated that all sentiment labels were correctly assigned after filtering. During TF-IDF vectorization, I handled potential issues with non-string input by checking that the text field was properly formatted. In modeling, I verified that all class labels were present in both the training and test sets to avoid label mismatch errors. These steps helped prevent runtime errors and ensured that the models trained and tested on clean, consistent input. Then, I created a new sentiment label using the star rating. Ratings of 4 and 5 were labeled as "Positive," a rating of 3 was labeled "Neutral," and ratings of 1 or 2 were labeled "Negative." This three-class system allowed me to train models to detect not just clear satisfaction or dissatisfaction, but also more ambiguous or balanced feedback. I split the data into training and testing sets using an 80/20 split. I then transformed the review text using TF-IDF vectorization. This method assigns weight to each word in a review based on how often it appears relative to the rest of the dataset. I included both unigrams and bigrams in the vectorization process to capture important short phrases like "too small" or "very comfortable," which added useful context to the model.

My first approach was to use a Naive Bayes classifier, which is often used for basic text classification tasks. While it initially performed well on the Positive class, it performed poorly on the Neutral and Negative classes, showing very low recall scores. I realized this was because the dataset was heavily imbalanced and most reviews were Positive, and the Neutral and Negative classes were underrepresented. To address this, I implemented Logistic Regression with class balancing using upsampling. This meant duplicating examples from the smaller classes so that all classes had an equal number of training samples. This change dramatically improved the model's performance, especially for the minority classes. The Neutral F1-score increased from 0.04 to 0.36, and the Negative class improved from 0.14 to 0.47. My overall weighted F1-score rose to

0.78, and accuracy remained strong at 77%, but now reflected more balanced predictive power across all three classes.

Beyond classification, I used Latent Dirichlet Allocation for topic modeling. This allowed me to identify common themes across the reviews without having to manually read thousands of entries. The five main topics that emerged involved product fit and sizing, fabric quality, overall comfort, color and design preferences, and emotional satisfaction such as customers saying they “loved” or “felt beautiful” in a product. These topics provided valuable context for what customers cared about most when reviewing a product. I also used clustering to segment reviewers into groups. I applied KMeans clustering on PCA-reduced TF-IDF vectors and metadata like age and sentiment score. This allowed me to group customers with similar feedback patterns, which could be useful for personalization or targeted product suggestions.

Throughout the project, I created a series of visualizations to better understand and communicate my findings. One of the most important visualizations was a strip plot showing the relationship between text sentiment scores which was calculated using VADER and the star ratings. This plot clearly showed that while higher ratings generally corresponded to more positive sentiment, there were many mismatches. Several 5-star reviews had neutral or even negative text sentiment scores. This confirmed that ratings alone often missed the full customer experience. Another visualization showed the distribution of review length by sentiment class. Interestingly, Neutral reviews tended to be longer, suggesting that customers with mixed opinions were more likely to explain themselves in detail. I also visualized average sentiment by product class and individual clothing ID. Chemises, Skirts, and Pants had the highest average sentiment, while Casual Bottoms, Sleepwear, and Trend categories had the lowest. These insights could help a business prioritize which categories to feature or improve. I also investigated the impact of reviewer influence. I grouped reviewers into Low and High influence based on the number of helpful votes their reviews received. I found that High-influence reviewers generally left more balanced reviews and were less likely to give overly positive or negative scores without explanation. These reviewers wrote longer reviews and provided more nuanced feedback, which could be especially useful for product improvement. This suggests that analyzing reviews based on reviewer credibility can enhance the reliability of sentiment analysis results.

One of the most important discoveries from this project was the frequent misalignment between the star ratings and the content of the review. Many customers gave high ratings but expressed dissatisfaction in the text, possibly due to brand loyalty or hesitation to leave low ratings. This validated the need for a more thorough sentiment analysis process in e-commerce platforms. If companies rely only on ratings, they could easily miss recurring problems that are clearly stated in the written feedback. There were limitations to my approach. The TF-IDF and Logistic Regression models, while interpretable and fast, do not capture the full context of language. For example, they can’t detect sarcasm or understand the relationship between distant words in a sentence. A model like BERT would be better suited for those tasks. Additionally, the dataset lacked time-based data, so I couldn’t track how customer sentiment changed over time. Including review dates would allow for trend analysis and seasonal behavior tracking, which would add another layer of value to the analysis.

Overall, I completed a full pipeline from data cleaning and sentiment labeling to model training, clustering, topic modeling, and detailed visualization. I improved the model’s performance through upsampling and parameter tuning, uncovered meaningful product and reviewer insights, and demonstrated that written reviews provide richer, more actionable feedback than star ratings alone. This project showed how natural language processing can transform qualitative customer reviews into structured data that supports better product development, marketing decisions, and customer experience strategies.