Course Project Part 2

Hannah Canela, Berkeley Rebman, Jayne VanKirk
Department of Business Information & Analytics, University of Denver

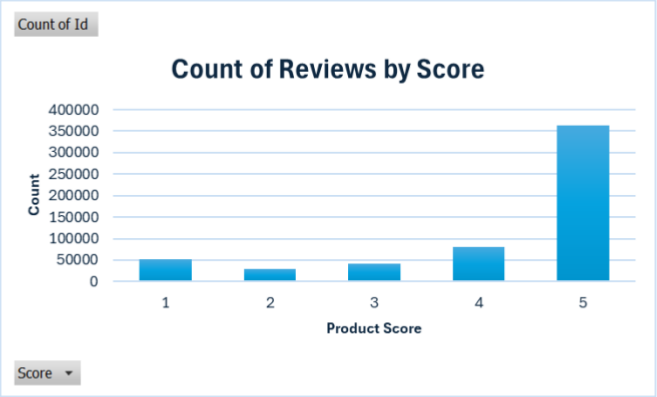
INFO-3400-1 (Complex Data Analytics)
Professor Tianjie Deng
March 12, 2025

Part 1: Data Introduction and Descriptive Analysis

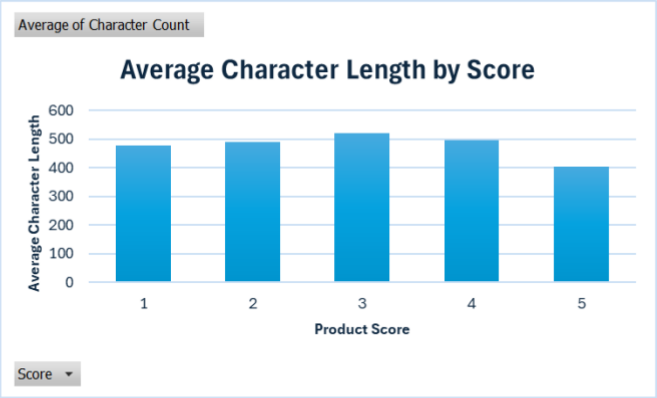
The Amazon Reviews dataset explores reviews for a variety of products. The original dataset contained 568,454 rows. The features in the dataset include:

Variable Name	**Variable Description**
Id	Unique identifier of the review
ProductId	Unique identifier of the product
UserId	Unique identifier of the user
ProfileName	User's online name
HelpfulnessNumerator	Number of users who found a review helpful
HelpfulnessDenominator	Number of users who indicated that a review was helpful or not
Score	Item's star rating between 1 (lowest) and 5 (highest)
Time	Timestamp of the review
Summary	Summary of the review
Text	Review

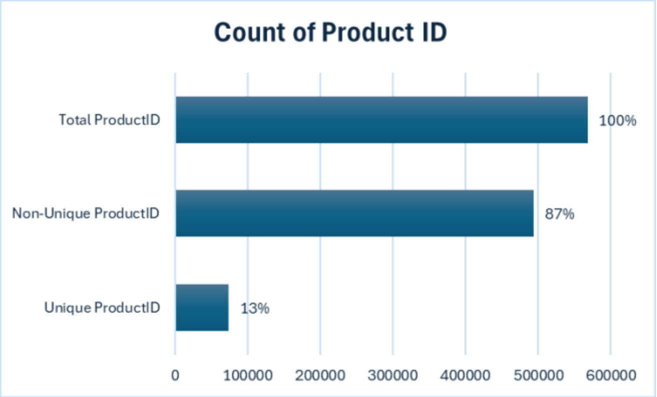
To understand the data before running text analytics, the group created graphs to identify some general metrics.



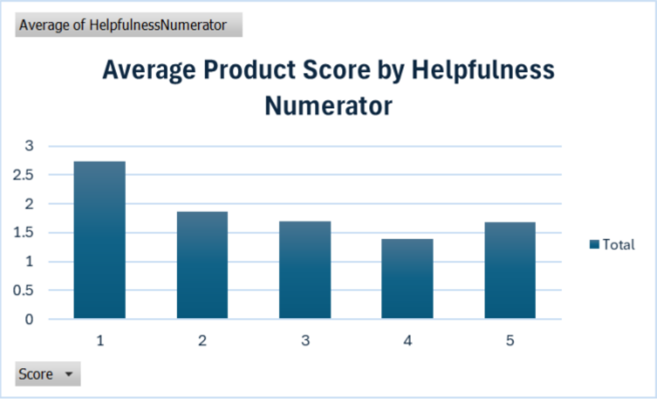
The "Count of Reviews by Score" graph shows the distribution of reviews by score. Most reviews are very positive with a score of 5 followed by a score of 4. There are only about 75,000 negative reviews (1s and 2s).



In the "Average Character Length by Score" graph, there is a relatively even distribution. The scores for the 5s reviews have the least amount of characters. This is potentially due to people having more to say when they are complaining and less to say when all is well.



Presented by the "Count of Product ID" bar chart, 87% of products are mentioned in more than one review. Only 13% of the 568,454 reviews (about 73,900 reviews) were only mentioned once in the dataset.

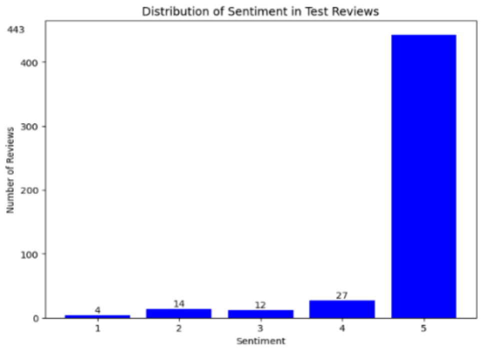


The "Average Product Score by Helpfulness Numerator" graph displays which numerators are the most helpful by scores. The reviews that were more helpful were negative ones, likely because they convinced other shoppers not to buy a certain product.

Part II: Text Mining Analysis

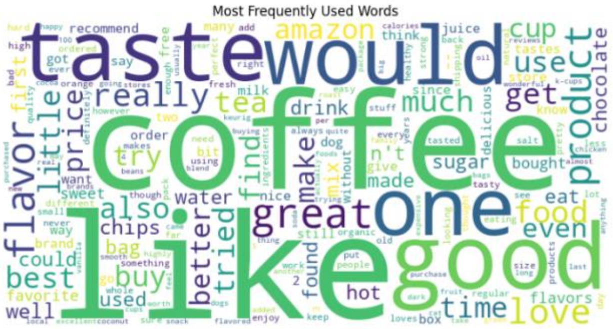
For ease of analysis and running complex models, the dataset was shortened to the first 10,000 rows for the final two sections. To begin an effective text analysis, the data needed to be cleaned first. HTML tags, punctuation, English stop words, and extra spaces were removed with contractions expanded into two words. The text was then preprocessed and tokenized for efficient analysis. The group defined three questions to explore in the analysis:

1. Sentiment Analysis: What is the distribution of the sentiment (positive, negative, or neutral) of all the reviews?



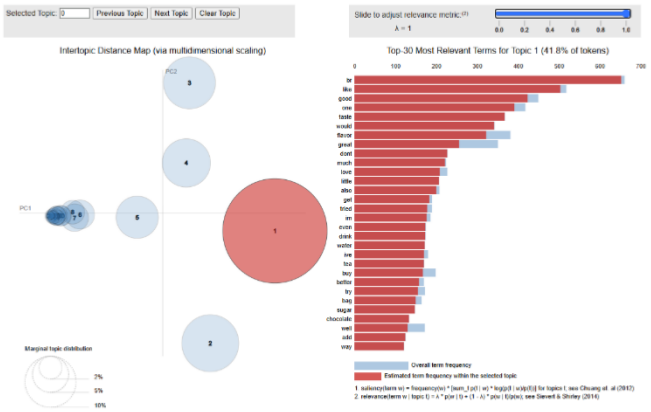
From the testing data (similar to the overall number of reviews per score), a strong majority of reviews are positive. The sentiment scores from the text align with the review (star) scores that customers gave on Amazon.

2. Keyword Extraction: What are the most common words in product reviews?



The most popular words in the Amazon reviews were coffee, like, taste, and good. All these words have a positive or neutral sentiment which aligns with almost 85% of the reviews being neutral or positive.

3. Topic Modeling: What are the most relevant terms for prevalent topics?



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Looking at the top topic with 41.8% of tokens, the words "like", "good", and "taste" are in the top five tokens which are also in the word cloud above. These words are generic and since reviews are about food, "taste" makes sense as a top word.

Completing these text analyses allows for recognition of patterns in the unstructured data which will help inform the predictive analysis.

Part 3: Predictive Analysis

To best predict the importance of factors the question "Can the textual contents of a product's review to predict the product's rating (score)?" was asked.

A robust analysis was completed with some preprocessing steps. First, any reviews that had a score of 3, 4, or 5 were listed as positive and reviews of 1 or 2 stars were defined as negative. 80% of the data (8,000 rows) were used in the training set and the remaining 20% (2,000 rows) were used in the testing dataset. The text features were converted into TF-IDF features which measured how important the words were by calculating the number of times a word was in a document (review) and the number of documents (reviews) the token was present in.

Once the preparation was completed, a logistic regression was created. With an accuracy of 89.225%, the liblinear (optimization algorithm used for smaller datasets) model performed the best. In 89.225% of cases, the model correctly predicted whether a review was positive (score of 3, 4, or 5) or negative (1 or 2). It is important to note that 85% of the reviews fit into the positive category and to improve model performance with other datasets, it would be wise to have a more balanced dataset since the model could only train on 15% negative reviews. This model performed well because it had a binary decision (positive or negative). More complex models with multiple categories to choose from (ex: selecting the correct score) would have most likely resulted in lower accuracy and predictions. In this scenario, sentiment helped predict whether a review was positive or negative. This can be evidenced from the sentiment analysis chart in Part II above.

Conclusion

In conclusion, the analysis and modeling provided useful information that demonstrated how stars/review scores are given by consumers. The models and analyses completed above were beneficial in identifying key tokens that were mentioned in positive, negative, or both reviews. The naming of these keywords could help suppliers identify pain points and areas of opportunity in their businesses. Moving forward, there are more analyses and modeling that can be completed. For example, creating a more complex model would allow for better sarcasm detection. Additionally, with more time, it would be interesting to see the prediction accuracy if the model were trained and tested with all 568,454 reviews. Looking at time stamps could also be interesting to see if people were more likely to give a positive review at a certain time of day. Overall, this process allowed for a robust text analysis of Amazon reviews.

Appendix

Arham Rumi. (2021). *Amazon Product Reviews* [CSV Data File]. Retrieved from https://www.kaggle.com/datasets/arhamrumi/amazon-product-reviews