# Module 4: Technique Practice

ALY6040 – Data Mining, Northeastern University

Professor Justin Grosz

12/03/23

Submitted by: Sudhamshu Vidyananda

# Introduction

In the dynamic and fiercely competitive e-commerce landscape, understanding customer preferences, expectations, and experiences is paramount for businesses seeking to maintain relevance and achieve sustainable growth. The vast array of customer-generated content, particularly product reviews, harbors rich insights into consumer behavior and market trends. However, the sheer volume and unstructured nature of this data pose a significant analytical challenge. This paper presents a novel approach to distilling actionable business intelligence from online customer reviews using advanced natural language processing and machine learning techniques.

## ***Business Problem***

The primary business problem addressed in this study is the *optimization of product offerings and enhancement of the customer purchase journey in Amazon online shopping.* Given the crucial role of customer reviews in shaping perceptions and influencing buying decisions, our focus is on extracting meaningful patterns and sentiments from textual reviews on a popular e-commerce platform, Amazon. The goal is to identify key factors that contribute to customer satisfaction and purchasing decisions, which, in turn, could inform product development, marketing strategies, and customer service enhancements.

## ***Approach and Methodology***

The methodology employed in this investigation comprises several steps, starting with the collection of a large dataset of Amazon reviews. Preliminary data processing includes cleaning and normalizing the text data, followed by the removal of stop words and extraction of significant nouns to reduce noise and focus on the most relevant terms. The core of our analytical framework involves two main components: topic modeling and word cloud visualization. Topic modeling is utilized to uncover latent topics within the review corpus. This unsupervised learning technique provides a probabilistic model that identifies words commonly occurring together, which we interpret as indicators of overarching themes in the data. Complementing the topic modeling, word cloud visualizations offer a straightforward and impactful representation of the frequency and relevance of terms within the reviews. These visualizations serve to highlight the most prominent and recurrent themes, providing an at-a-glance understanding of customer sentiments and priorities. The analytical process is iterative, with initial findings leading to refined queries and deeper analysis.

## Data Cleanup

Data cleaning for the Amazon review dataset is an indispensable step that enhances the accuracy and relevance of the subsequent analysis. By preparing the dataset through a meticulous cleaning process, researchers ensure that the foundation upon which they build their insights is robust and reliable, leading to more credible and actionable outcomes. The first step was to find out the basic statistical features of the dataset and through this process, we could find out that Amazon Review dataset contained 5 numerical and 5 categorical values. **ProfileName and Summary** had null/missing value in them. These 2 values will not be used for text analysis and hence the null values will not be removed as the concentration will be on the **Text** column only. Removing the null values from these columns will eventually remove the data from Text column and we will end up losing the data to be analyzed as Text column is not having any null values in them. Dataset was also checked for **duplicated records** and found none. Duplicated records will disrupt the data modelling process and will not produce accurate results and needs to be removed from the dataset.

## Exploratory Data Analysis

EDA is done to answer the hypothesis questions and have a visual interpretation of it.

***Hypothesis: 5-star reviews will predominantly contain language that expresses satisfaction with the product's quality, value for money, and user experience.***

To test this hypothesis, text mining techniques such as sentiment analysis and keyword extraction could be applied to the text of the 5-star reviews to identify common themes and sentiments. For example, if the hypothesis is correct, we would expect to find a significant occurrence of positive sentiment words such as happy, satisfied, excellent, worth and recommend. Additionally, we will also check whether lower-star reviews (1-star and 2-star) share common characteristics in their text that indicate specific aspects of dissatisfaction. This could include words like disappointed, poor, return, or bad. **Figure 1 (see Appendix)** shows a bar chart giving the count of reviews by star ratings, which appears to be from the Amazon reviews dataset mentioned earlier. It presents a distribution of reviews across different star ratings on a scale of 1 to 5, with each bar representing the number of reviews for each rating. The **purple** bar is the tallest, representing 5-star reviews with a count of 363,122. This suggests that many customers were very satisfied with their purchases. The **red** bar represents 1-star reviews with a count of 52,268. This indicates the number of users who were highly dissatisfied with their purchase. The **green** bar represents 3-star reviews with a count of 42,640. These are neutral reviews, indicating a middling customer experience. The bar graph in **Figure 2(see Appendix)**, supports this hypothesis by showing that the most common words in 5-star reviews are overwhelmingly positive. Words like **great, good, love and best** suggest strong satisfaction with the products, while **recommend, delicious, perfect, free, healthy, easy, and fresh** could be related to the quality, value, and user experience of the products. The presence of these words at high frequencies confirms the hypothesis to a significant degree. The customers' satisfaction is clearly reflected in the positive language they use. The word **amazon** being prominent suggests a positive association with the purchasing platform itself, which might also influence the overall customer experience and satisfaction.

### *Correlation Matrix*

The correlation matrix heatmap provided in **Figure 3 (see Appendix),** shows the relationship between various variables in the Amazon review dataset. **HelpfulnessNumerator** and **HelpfulnessDenominator** have a very high positive correlation **(0.99)**, which makes sense because the numerator (number of helpful votes) cannot exceed the denominator (total votes). As one increases, so often does the other. **Time** and **ProductId** have a moderate positive correlation **(0.42)**, suggesting that certain products may receive more reviews at specific times. This might be due to a new product release, a sale, or seasonal purchasing patterns. The variable **Score** shows little to no linear correlation with most other variables. This could mean that the star rating does not strongly depend on the product ID, user ID, or the helpfulness of the review.

## Data/Text mining

Text mining is crucial because it allows businesses and researchers to analyze vast amounts of unstructured text data efficiently, which would be impractical to process manually. By employing text mining techniques, valuable insights can be extracted from data such as customer reviews, social media posts, emails, and documents. These insights can reveal patterns, trends, and sentiments that inform decision-making in various domains like marketing, customer service, and product development. The analysis will be focused on processing and examining **Text** data from Amazon reviews. Since the dataset is large, it is filtered based on certain criteria, specifically focusing on reviews with a score greater than **2** and limiting the analysis to the **first**

**50,000** entries. All the Stopwords and nouns are stripped out from the Text column indicating a focus on preprocessing the textual data for more meaningful analysis. Few Stopwords are additionally added since we are creating document-term matrix. The data is converted into a matrix format, for subsequent analysis. A corpus is created to convert a sparse matrix of term-document frequency into a Gensim corpus object. The sparse matrix is transposed because Gensim expects terms as rows, not columns. **id2wordn** is a dictionary that maps each term's integer ID to its string form. This is generated by reversing the 'cv.vocabulary_' mapping which originally maps terms to their IDs. An **LDA (Latent Dirichlet Allocation)** model is instantiated and trained on the corpus with two topics (num_topics=2), mapping the term IDs to the actual words (id2word=id2wordn), and iterating over the corpus ten times (passes=10). **Figure 4(see Appendix)** shows the topic modeling results from the above modeling steps taken. For each topic identified by the LDA model, a **word cloud (Figure 5(a) & 5(b), see Appendix)** is created to show the words that make up each topic. *The text size depends on the weightage of the text.* Larger the text, more weightage and vice versa. From our topic modeling results, for **topic 0**, the key terms such as **food, bars, snack, dog, butter, protein, and treats** suggest a focus on edible products, possibly including health bars, pet food, and snacks. A company selling these products can deduce that customers are interested in the nutritional content (protein) and possibly the convenience (snack, bars) of their products. If customers frequently mention **protein** in snack bars, it suggests a market trend toward health-conscious eating. A food manufacturer can use this information to create new products or improve existing ones with higher protein content. **Topic 1**, it is centered around beverages, with words like **tea, coffee, flavor, chocolate, and cup** being prominent. A beverage company could use this information to improve or introduce new flavors that resonate with consumer preferences. The emphasis on **flavor and taste** indicates that these are significant factors in customer satisfaction. If **flavor** is a significant factor in beverage reviews, beverage companies can innovate with new flavors or highlight the taste profiles of their drinks. The **word cloud** for the entire dataset **Figure 6 (see Appendix)**, displays terms such as **love, taste, make, Amazon, buy, good, need, easy, and used**. Mentions of **Amazon** and **buy** indicate the importance of the purchasing platform and the transaction experience. Ensuring that product listings are clear, accurate, and that customer service is responsive on this platform can directly affect customer satisfaction and loyalty. The prevalence of words like **love and good** indicates a positive reception of products. However, the presence of **need and easy** suggests that customers are also voicing desires or expectations for practicality and convenience. Positive words like "love" and "good" in the reviews can be used in marketing materials to create a positive brand image. For instance, testimonials or phrases from positive reviews can be quoted in advertisements to attract potential customers. Retailers can capitalize on this by highlighting how their products meet these needs in their marketing strategies.

## Interpretation and Recommendation

The prominence of food, bars, snack, and protein in the word cloud reinforces the findings from topic 0, showing a clear interest in convenient and healthy snack options. Words like tea, flavor, coffee, and chocolate suggest a varied interest in hot beverages. This points to a diverse customer palate and the importance of offering a range of flavors to meet different tastes. The broader terms in this cloud, including Amazon, buy, love, and easy suggest that customers are not only interested in the products themselves but also in the purchasing experience. Positive sentiments like love and good imply customer satisfaction with the products or the shopping experience.

For the data owner, which appears to be a retailer or a product manufacturer selling on Amazon, the analysis indicates that:

**Product Development:** There's a clear market interest in nutritious, flavorful, and convenient products. It is recommended to focus on these aspects in product development.

For beverages, variety in flavor is important. Consider introducing new flavors or highlighting existing ones in the product line.

**Marketing Strategy:** Use the positive sentiment words from the word clouds in marketing campaigns to reinforce the emotional connection customers have with the products.

Emphasize the convenience and nutritional aspects of food products, and the variety and quality of flavors in beverages, as these are key drivers of customer satisfaction.

**Customer Experience and Pricing:** Since the term **price** is visible in the word clouds, ensure that the products are competitively priced, and consider running promotions or discounts to attract price-sensitive customers. Highlight the ease of purchase and fast delivery options on Amazon to improve and streamline the customer experience.

**Inventory and Sales:** Stock levels should be managed to ensure availability of the high-interest and positively reviewed items, particularly those associated with strong positive sentiments like love and good.

### ***Additional Variables to Incorporate***

To deepen the analysis, we can consider integrating customer demographic data such as age, location, and purchasing power, which can offer insights into which segments are most engaged with these products. Track changes in review sentiments over time to understand how product adjustments or market trends affect customer opinions. By aligning product attributes with customer preferences and ensuring that the buying experience is seamless, the data owner can enhance product offerings and potentially increase customer satisfaction and sales.

## Conclusion

The analysis of Amazon reviews using topic modeling and word cloud visualizations has underscored the importance of customer preferences for product attributes, such as the nutritional value of food and the flavor diversity in beverages. Positive customer sentiments, highlighted by words like "love" and "good," along with considerations around purchasing experiences and cost, signal satisfaction and influence buying behavior. These insights offer actionable intelligence for businesses, enabling them to refine products, tailor marketing, and adjust pricing strategies to align with consumer expectations. From this analysis, we have learned the importance of actively listening to customer feedback through data. It's not just about collecting data but interpreting it in a way that leads to better products, improved customer experiences, and smarter business strategies. This approach exemplifies how big data and analytics can be transformative tools for businesses in the digital age.

## Appendix

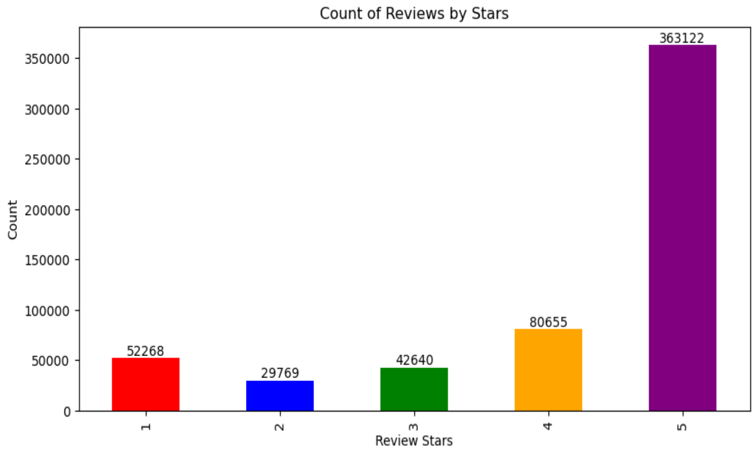


*Figure 1: Bar graph for Scores*

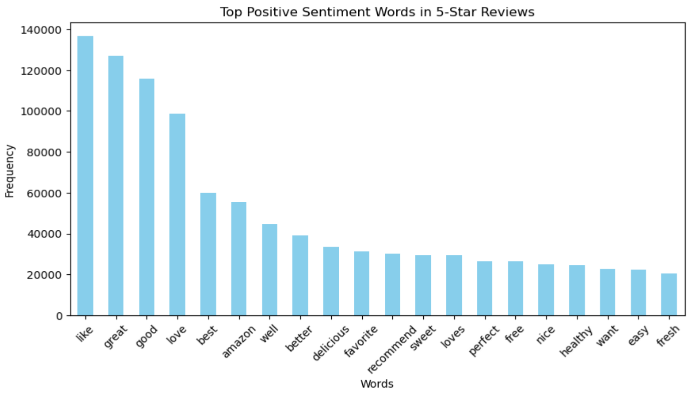


*Figure 2: Sentiment Analysis for 5-star Reviews*

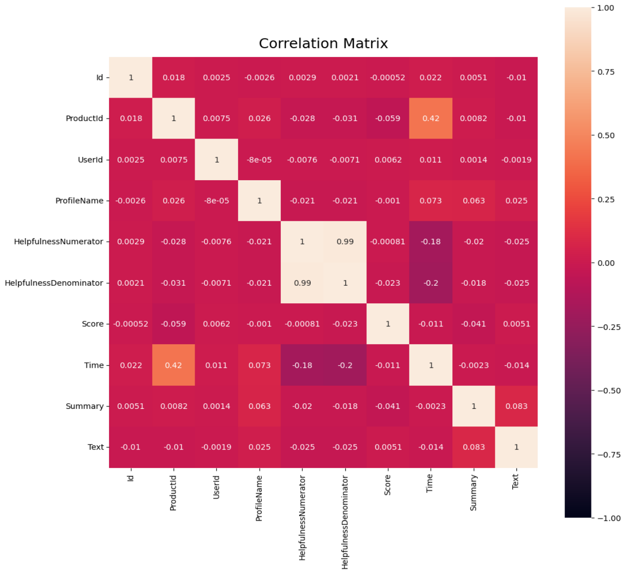


*Figure 3: Correlation Matrix*

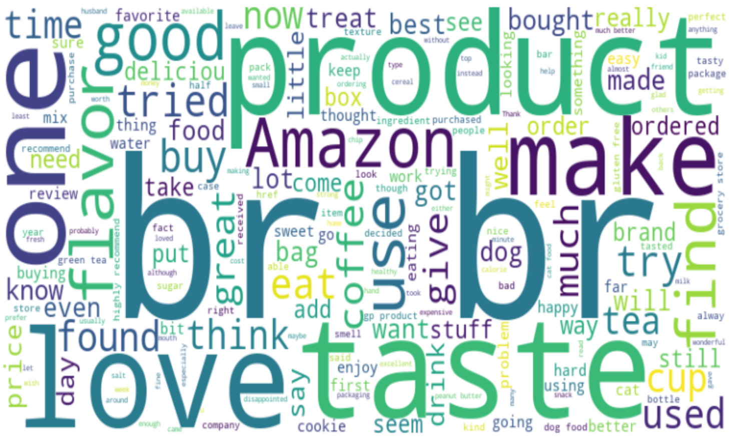
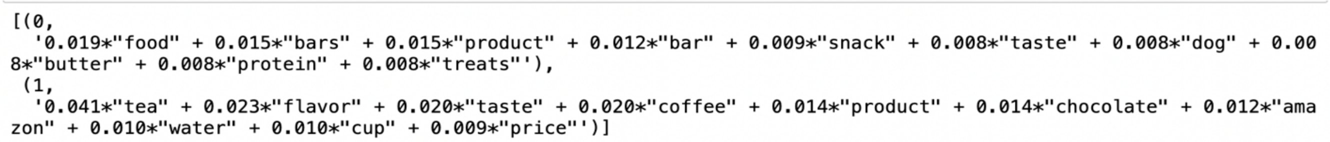


*Figure 6: Word Cloud for Entire Dataset*



```
[(0,
  '0.019*"food" + 0.015*"bars" + 0.015*"product" + 0.012*"bar" + 0.009*"snack" + 0.008*"taste" + 0.008*"dog" + 0.00
8*"butter" + 0.008*"protein" + 0.008*"treats"'),
 (1,
  '0.041*"tea" + 0.023*"flavor" + 0.020*"taste" + 0.020*"coffee" + 0.014*"product" + 0.014*"chocolate" + 0.012*"ama
zon" + 0.010*"water" + 0.010*"cup" + 0.009*"price"')]
```

*Figure 4: Topic Modeling*

*Figure 5(a): Word Cloud for Topic 0*

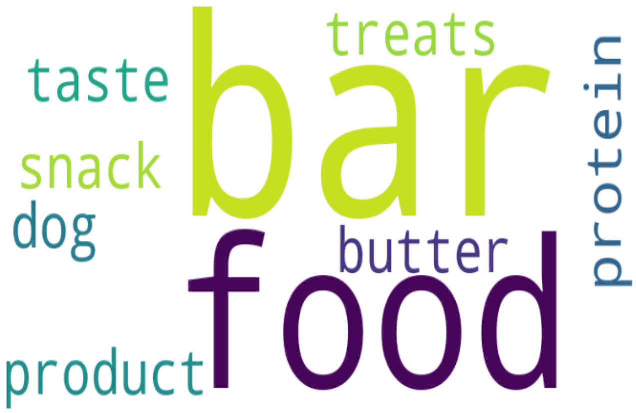
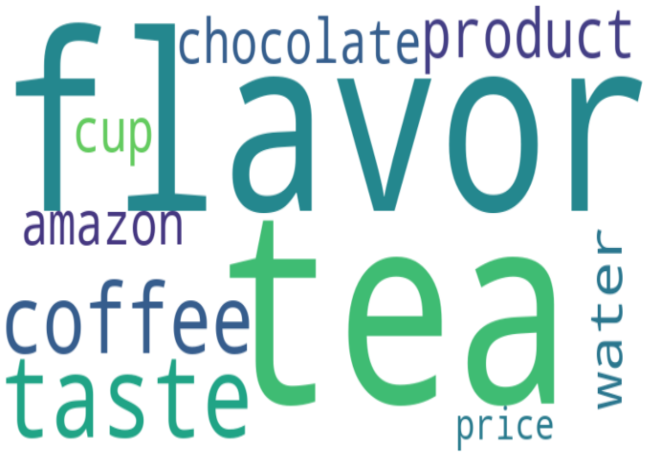


*Figure 5(b): Word Cloud for Topic 1*



## References

Barde, B. V., & Bainwad, A. M. (2017, June). An overview of topic modeling methods and tools. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 745-750). IEEE.

Bose, R., Dey, R. K., Roy, S., & Sarddar, D. (2020). Sentiment analysis on online product reviews. In *Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2018* (pp. 559-569). Springer Singapore.

Cambria, E., Das, D., Bandyopadhyay, S., & Feraco, A. (Eds.). (2017). *A practical guide to sentiment analysis* (Vol. 5). Cham: Springer International Publishing.

Cohen, K. B., & Hunter, L. (2008). Getting started in text mining. *PLoS computational biology*, *4*(1), e20.

Dasu, T., & Johnson, T. (2003). *Exploratory data mining and data cleaning*. John Wiley & Sons.

Heimerl, F., Lohmann, S., Lange, S., & Ertl, T. (2014, January). Word cloud explorer: Text analytics based on word clouds. In *2014 47th Hawaii international conference on system sciences* (pp. 1833-1842). IEEE.