BA820 Group-A2 BA820_Amazon.ipynb Github

Chih-Hsin Peng, Gaurangi Agrawal, Jiacheng Li, Ya Chu Hsu, Yi-Cheng Chung

**Trend Tracking: Leverage Review Data for Industry Insight on Amazon**

Exploratory Data Analysis: The Amazon Fashion data from UCSD consists of 883,636 reviews covering diverse fashion products from 1998 to 2018. It includes 12 columns: The original dataset is too large. Therefore, we applied the sampling method to retrieve 5% data. Now we have 44,182 reviews in the dataset.

Pre-processing:

- **Handling Missing Values:**
    1. Null values in columns *reviewerName*, *reviewTex*t, and *summary* were dropped as they accounted for a minimal percentage of the data.
    2. *style* and *image* columns, with over 90% null values were dropped entirely.
    3. Null values in the *vote* column, representing the number of helpful votes, were replaced with 0.
- **Handling Duplicates:** Duplicate rows were identified in these columns, *asin* (item ID), *reviewerID*, *unixReviewTime* (review time), and *reviewText*. Duplicates were then removed, ensuring each review was unique.
- **Text Cleaning:** Lower-casing, removing all non-alphanumeric characters and stop words, stemming, and lemmatization for text.

**Analysis Plan**:

- **Tokenization:** We proceeded with word tokenization on the review text and opted for a TF-IDF technique. This choice allowed us to convert textual data into numerical representations, facilitating quantitative analysis of customer reviews. By employing this approach, we aimed to extract meaningful features and identify prevalent sentiments and themes within the dataset. Ultimately, our goal was to gain insights into customer preferences and behaviors within the fashion sector through the quantification of text.
- **NLP:** We summarized the *reviewText* column and compared the generated summaries with the *summary* column. By condensing detailed reviews into concise summaries, we aim to extract valuable insights into consumer behavior within the fashion sector. This approach will enable us to understand the key factors influencing purchase decisions and identify prevalent sentiments expressed by customers regarding fashion products.

**Preliminary Results**:

- **NLP:** Summarizing the *reviewText* and comparing the result with *summary*. We created a new column called *summary(after pre-process)*. This summary is derived from *reviewText(after pre-process)*, where we extract verbs and adjectives. Later, we compare the *summary* provided by the original dataset with the *summary(after pre-process)* to assess their similarity by using cosine similarity. We retain only those with cosine similarity above 0.5, defining them as similar. The analysis indicates that only 5.11% of reviews have similar summaries to their provided ones, suggesting that most extracted summaries differ from the originals.

- **Word cloud:** The word cloud generated from the *reviewText* column presents a striking portrayal of consumer sentiment and product characteristics. Central to the customer feedback are prominent terms such as 'great,' 'comfort,' 'material,' and 'color,' which are notably large, signifying their frequent mention and presumed significance to the reviewers. The words 'use' and 'work' stand out, indicating the practicality of the products. Positive words like 'perfect,' 'beautiful,' 'nice,' and 'good' point to generally favorable feedback. Conversely, terms such as 'disappoint' and 'return' signal some negative experiences. Descriptive words like 'short,' 'small,' 'big,' 'long,' and 'tight' likely pertain to the fit or style of the items purchased. Verbs such as 'wear,' 'order,' 'get,' and 'need' highlight the transactional aspects of the purchasing process. The sizable appearance of words like 'dress,' 'shirt,' 'shoe,' and 'wallet' suggests these are items frequently discussed, thereby garnering more reviews. Lastly, the prominent size of 'nice' and 'love 'suggests a tendency toward *positive* reviews in the data set. The word cloud shown in Figure 3, is derived from preprocessed review text and provides an immediate and insightful overview of the aspects that matter more and weigh more to customers, offering valuable perspectives for our analysis.

**Actions Implemented**:

- **Sentiment Analysis:** We use the VADER sentiment analyzer on pre-processed *reviewtext*. The 'compound' score is normalized and weighted across neutral, positive, and negative scores, ranging from -1 to +1. We categorize these scores into five emotional tones: 'Very Positive,' 'Positive,' 'Neutral,' 'Negative,' and 'Very Negative.' The scores above 0.6 are labeled as ‘Very Positive’, scores between 0.2 and 0.6 as ‘Positive’, scores between -0.2 and 0.2 as ‘Neutral’, scores between -0.6 and -0.2 as ‘Negative’, and scores below -0.6 as ‘Very Negative’.This classification helps businesses understand customer sentiment towards products, enabling them to enhance offerings and satisfaction. Analyzing trends in feedback allows businesses to identify areas for improvement and tailor marketing strategies to meet consumer needs, potentially increasing brand loyalty and market share.
- **Clustering Optimization:** We conducted a thorough selection process to determine the optimal number of clusters for the K-means algorithm. Additionally, we employed visualization tools such as PCA to effectively illustrate the cohesive grouping of similar reviews within these clusters. Our next step was to conduct an in-depth analysis of each cluster to identify prevalent sentiments and significant feedback patterns.
- **Integration with Rating Data:** We integrated our sentiment analysis results with the overall rating data to validate the effectiveness of our sentiment model. Any disparities identified were thoroughly investigated to understand the limitations of sentiment analysis. Additionally, we explored the potential for predictive models that leverage sentiment scores and textual features to predict numerical ratings. This effort aimed to establish a quantitative connection between textual sentiment and rating behavior.

**Results**:

- The histogram(Figure 4) indicates a notable concentration of reviews falling within the range of 0.2 to 1.0, suggesting predominantly positive sentiment among the analyzed reviews. Few reviews exhibit 'Very Negative' sentiment scores (below -0.6), while a moderate number falls within the neutral sentiment range (around 0). This distribution

visually portrays the overall sentiment polarity within our dataset, aiding in discerning the prevailing level of customer satisfaction or dissatisfaction with the reviewed products.

- Results indicated that the '5.0' rating exhibits the highest percentage of matches, with the majority (15,434) classified as 'Very Positive', accounting for approximately 35.02%. This suggests a strong alignment between the 'Very Positive' classifications and the highest rating. However, match percentages for other ratings (1.0 to 4.0) are notably lower, indicating fewer matches between sentiment classifications and actual ratings. This analysis provides insight into how well the sentiment analysis model's classifications correspond with reviewers' actual ratings. A high correlation between these two metrics would validate the model's effectiveness in accurately capturing review sentiment.

**Challenge**:

- **Processing Limitations with Full Dataset:** Handling the whole dataset proves challenging given its substantial size, consisting of 883,636 records. The limited RAM capacity of Colab necessitates working with a smaller sample size. Consequently, we had to employ a random formula to reduce the dataset to 5% of its original size (44,182 records). However, this reduction may compromise accuracy, as some data may be lost in the process.
- **Constraints of using Word2Vec:** Initially, our plan involved applying the Word2Vec method to analyze the distinction in wording among positive, neutral, and negative reviews. The expectation was that words such as 'love,' 'like,' and 'favorite' would cluster together for positive reviews, while words like 'hate,' 'dislike,' and 'hard' would cluster for negative ones. However, the visualization process required dimension reduction using PCA. Due to the large dataset, limitations in Colab hindered achieving the desired display.
- **Clustering Constraints:** With PCA, we condensed the dataset into 1078 components to capture around 80% of its variance. However, when attempting to present our findings, the extensive number of components has significantly hindered our ability to extract meaningful insights from the clustering results. This challenge has made it exceedingly difficult to derive actionable conclusions or discern patterns effectively.

**Conclusion**:

- The sentiment classification accuracy of reviews, achieved through the VADER sentiment analyzer, reveals a moderate alignment between the highest customer ratings ('5.0') and the sentiment expressed in their reviews. This reflects the reasonable effectiveness of the sentiment analysis method in capturing the nuances of customer emotions.
- In the realm of machine-learning applications and associated challenges, the project endeavors to utilize unsupervised machine-learning techniques for textual analysis. The goal is to unearth trends and determinants that sway purchasing decisions in the fashion sector.
- Regarding sentiment analysis, research into Amazon customer reviews has proven valuable, providing a deeper understanding of customer contentment and pinpointing disparities within the reviews. These insights are pivotal for Amazon to craft strategies that resonate with current market trends and leverage consumer feedback for product refinement.

**Contribution Sheet**

| | Coding | Wording | Contribution |
|---|---|---|---|
| ChihHsin (Olivia) Peng | EDA /<br>Data preprocessing /<br>NLP data cleaning /<br>Sentiment analysis - Distribution chart / Conclusion | <u>Colab Analysis:</u><br>Data preprocessing /<br>Relationship between original_summary and summary after pre-processing /<br>Verification status of a customer /<br>Challenge /<br>Slide | (5/20*2/3)<br>+(7/25 *1/3)<br>= 24% |
| YiCheng (Ethan) Chung | Data preprocessing /<br>NLP summary comparing /<br>Emotional tone comparison /<br>Top3 items by the tone /<br>Sentiment Score by Overall Rating and Tone | <u>Colab Analysis:</u><br>Data preprocessing /<br>Top 3 Most Reviewed Products by Sentiment Tone /<br>Verification status /<br>Slide | (5/20*2/3)<br>+(5/25 *1/3)<br>= 22 % |
| YaChu (Ya Ya) Hsu | Tone across verification /<br>Average Sentiment Score /<br>Sentiment Score by Review Length | <u>Colab Analysis:</u><br>Word Cloud /<br>Sentiment Analysis /<br>Distribution of Sentiment Scores /<br>Compare the classified emotional tone with the overall rating /<br>Average Sentiment Score and Review Volume Over Time /<br>Average Sentiment Score by Review Length /<br>Conclusion /<br>Slide | (3/20*2/3)<br>+(8/25 *1/3)<br>= 21 % |
| Gaurangi Agrawal | EDA /<br>Clustering /<br>PCA /<br>K-Means /<br>Sentiment Analysis - Classify tone and Score | <u>Colab Analysis:</u><br>PCA /<br>K-Means /<br>Business Relevance /<br>Google Document | (5/20*2/3)<br>+(4/25 *1/3)<br>= 21 % |
| JiaCheng Li | Sentimental Analysis - Prediction /<br>NLP Word Could Visuals | <u>Colab Analysis:</u><br>Sentimental Analysis /<br>Data Introduction /Google Document | (2/20*2/3)<br>+(3/25 *1/3)<br>= 12% |
| Total | 20 | 25 | |
| Weight | * 2/3 | * 1/3 | |

# Appendix

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 618545 entries, 87569 to 500283
Data columns (total 12 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   overall         618545 non-null  float64
 1   verified        618545 non-null  bool
 2   reviewTime      618545 non-null  object
 3   reviewerID      618545 non-null  object
 4   asin            618545 non-null  object
 5   reviewerName    618480 non-null  object
 6   reviewText      617689 non-null  object
 7   summary         618165 non-null  object
 8   unixReviewTime  618545 non-null  int64
 9   vote            55867 non-null   object
 10  style           213157 non-null  object
 11  image           20180 non-null   object
dtypes: bool(1), float64(1), int64(1), object(9)
memory usage: 57.2+ MB
```

**Figure 1. Dataset information**

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 44182 entries, 87569 to 847668
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   overall         44182 non-null  float64
 1   verified        44182 non-null  bool
 2   reviewTime      44182 non-null  object
 3   reviewerID      44182 non-null  object
 4   asin            44182 non-null  object
 5   reviewerName    44175 non-null  object
 6   reviewText      44122 non-null  object
 7   summary         44155 non-null  object
 8   unixReviewTime  44182 non-null  int64
 9   vote            3953 non-null   object
 10  style           15254 non-null  object
 11  image           1385 non-null   object
dtypes: bool(1), float64(1), int64(1), object(9)
memory usage: 4.1+ MB
```

**Figure 2. Dataset information after pre-processing**


**Figure 3. Word Cloud based on NLP**

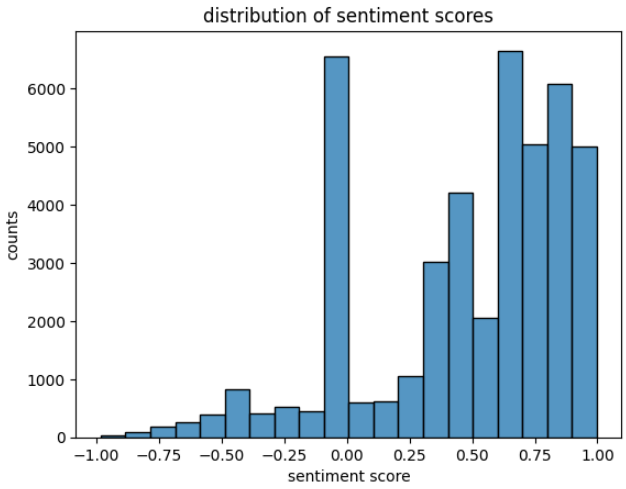


**Figure 4. Distribution of Sentimental Scores**

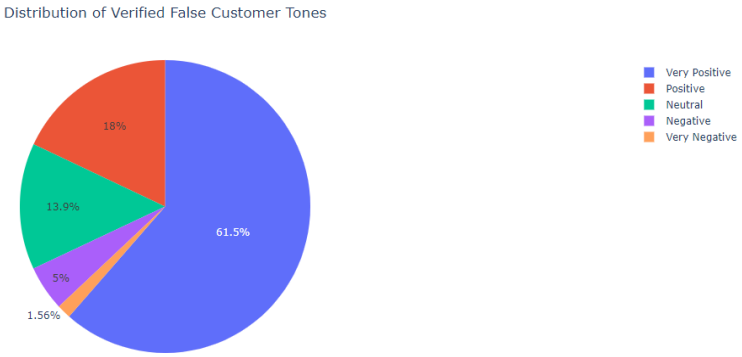


**Figure 5. Distribution of Verified False Customer Tone**

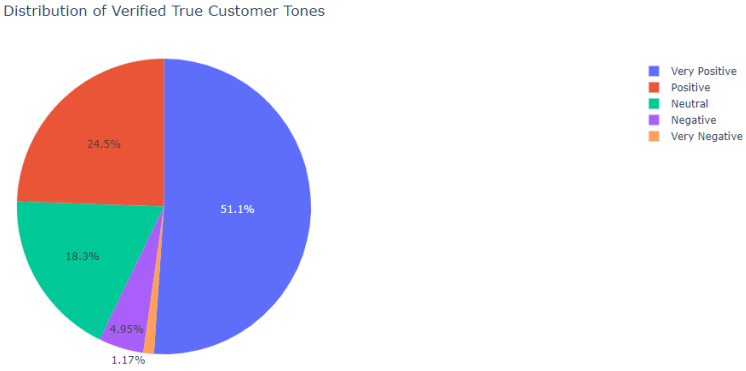


**Figure 6. Distribution of Verified True Customer Tone**

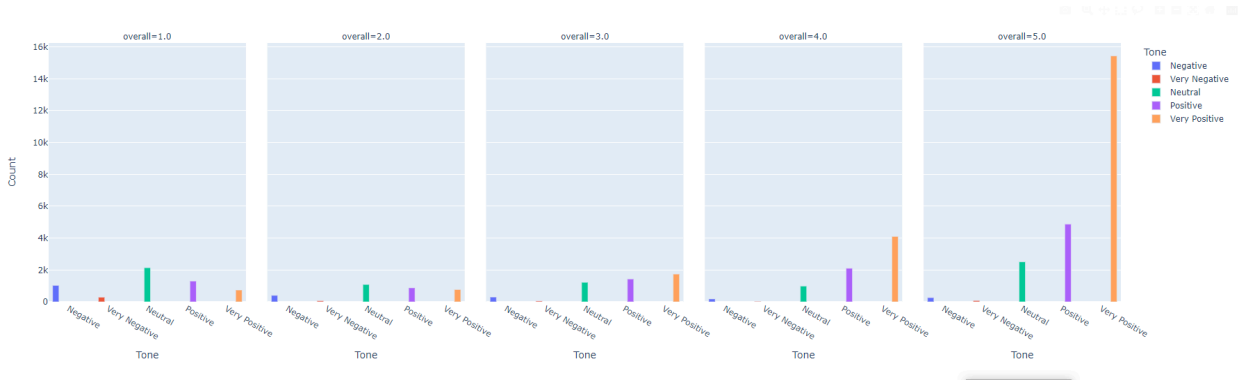


**Figure 7. Sentimental Analysis Based on Rating**