Case Study - Amazon Book Review

Cece Liu, Imaobong Undieh, Nisera Vaughan, Sichen Tong, and Yijie Xu

Big Data Analytics (BYGB-7990-001)

Gabelli School of Business, Fordham University

Executive Summary

The significance of online reviews has grown significantly in today's e-commerce landscape, prompting increased attention to the analysis. This report delves into the connection between sentiment in online reviews and various factors such as ratings, helpfulness, and price. Additionally, different models for sentiment analysis were evaluated. The identified sentiment analysis model offers a helpful tool for businesses and researchers.

A critical analysis finding reveals a positive correlation between sentiment and rating, except for reviews with a rating of "4," which surprisingly showed more negative sentiment. Further investigation is recommended to understand the unique dynamics surrounding reviews with a rating of "4", which may indicate a more neutral sentiment where reviewers were neither overly impressed nor disappointed.

The relationship between review sentiment and product price was also explored. Although a weak correlation was observed, higher-priced items tended to elicit more neutral reactions. This could be attributed to consumers' greater sense of investment in expensive products and a desire to justify their expenditure.

Overall, this report contributes to understanding the dynamics of online reviews and provides valuable implications for businesses aiming to analyze and leverage sentiment in online consumer feedback.

Business Problem

The Amazon Book Review is an online resource and website that enlists the help of Amazon Books Editors to curate book recommendations for Amazon customers. Amazon boasts that the Editors are a small group; however, due to the small number, book reviews and recommendations tend to be biased.

To resolve the issue of biased recommendations, a predictive model will be constructed to analyze book reviews by utilizing real customer reviews.

Benefits of this research include:

- Unbiased and Diversified Book Recommendations
- Book Pricing
- Better Inventory Control of Books

Research Questions

I. **Are measures of book review sentiments strongly correlated with book ratings measured on a scale of 1 - 5?**
   - Review sentiment was plotted on a scale from -1 to 1, which aligns with negative to positive sentiment, and compared reviews with each rating score to assess trends or patterns.

II. **Is there an association between a book review's sentiment and its helpfulness rating?**
   - A helpfulness rating is a rating given by those who read a book review before purchasing the book and either found it helpful or had it contribute to their decision to purchase the book. To answer this research question, we plotted the helpfulness ratings on a scale from 0 to 1 and graphed them alongside review sentiment.

III. **How can book price factor into either an overall positive or negative review sentiment that can ultimately affect its ratings?**

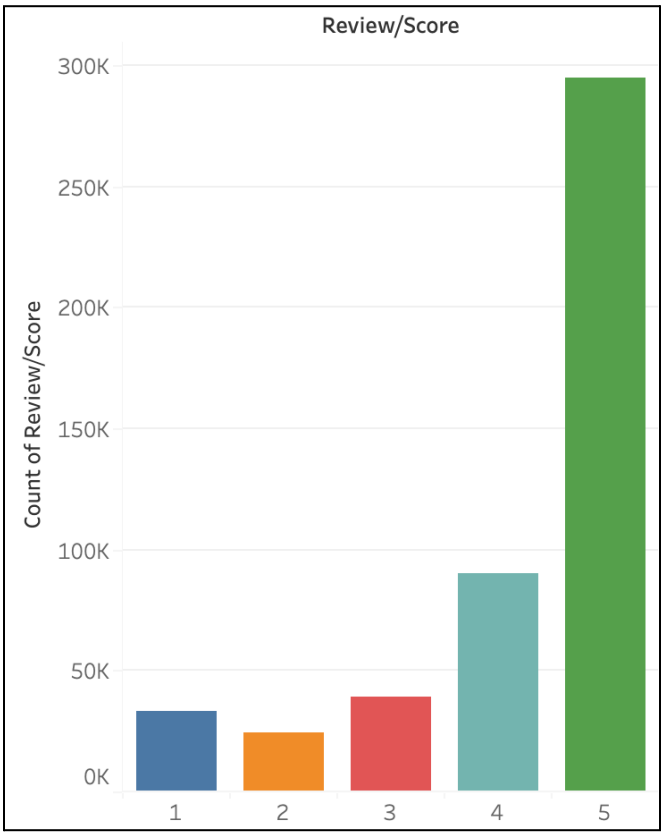
Dataset Description

The dataset was sourced from Kaggle and consisted of over three million records.

Variable	Type	Description	Example
Id	string	The ID of the book	1882931173
Title	string	Book title	Dramatica for Screenwriters
Price	float	The price of the book	17.46
User_Id	string	ID of user that rated the book	A2KSXSRTMD3ZJ4
profileName	string	Name of user who rated the book	Johnny Appleseed
review/ helpfulness	float	Helpfulness rating of the review	5.0
review/score	float	Rating from 0 to 5 for the book	2.0
review/time	float	Time the review was given	1335108095.98

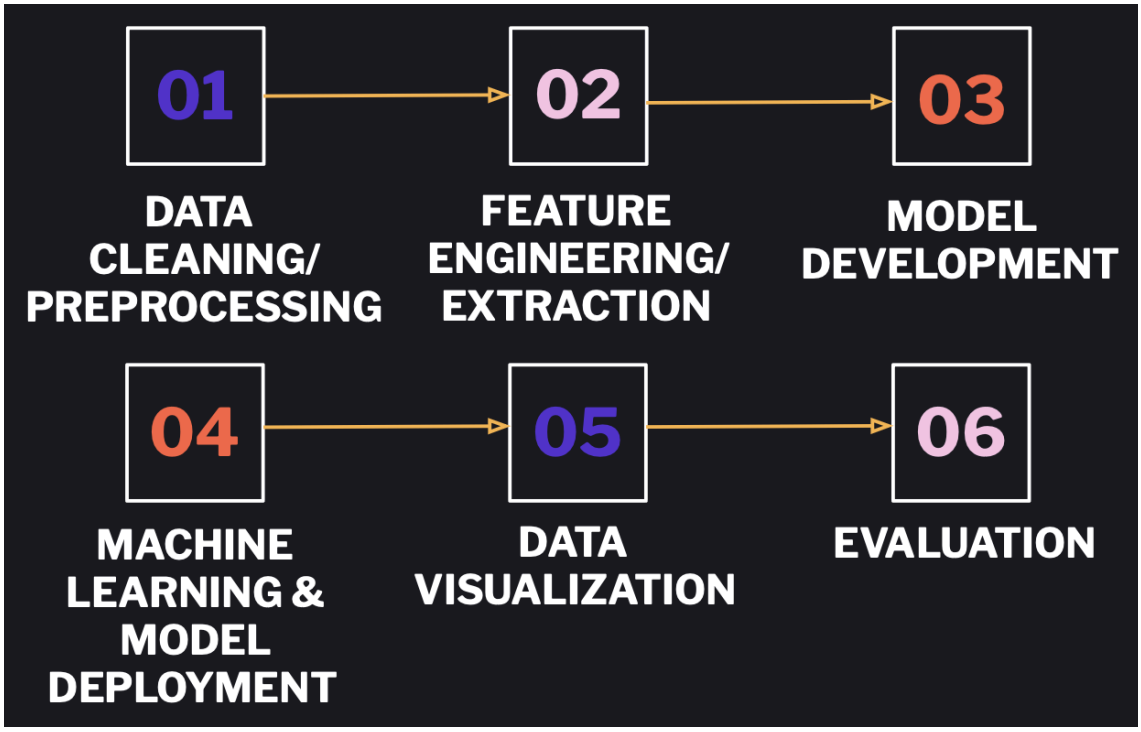
review/ summary	string	Summary of the text review	“Exactly what I needed.”
review/text	string	The full text of the review	“This book answered all my questions about getting into law school in a plain and straightforward way...”

I. Rating Distribution



This histogram shows the distribution of the rating score. Rating scores a heavily skewed to the right.

System Design



Data Preprocessing

I. **Basic Preprocessing**
   - Keep Only Necessary Information
     - As the original dataset consists of ten columns which include the 'User_id' and 'profileName' columns, which are unnecessary to our research questions, we removed those columns using the .drop() method.
     - For the data frame with the rest of the columns necessary to our research questions, any 'NaN' that occurs represents the row being useless and must be removed. Therefore, we removed these useless rows using the .dropna() method. After these two actions, the dataset was narrowed down from four million rows to 300,000.
   - Text Cleaning
     - To clean the review texts and keep only the informational context, we created a clean_text function which includes the sub-functions that remove all the non-alphabetic characters and convert all the texts to lowercase.

- Fraction Conversion
  - All the values under the column 'review/helpfulness' are in fractions such as 8/10, 1/1, and 7/11. To convert these fractions into decimals so that the values can be comparable in the model, a frac_to_dec function was created to extract the nominator and denominator from each row of the column and calculate them into decimals. Using this function, we could also handle the 0/0 and classify this value into 1 using the if-else statement.

II. Textual Level Preprocessing

- Tokenizing
  - After the Basic Preprocessing phase, the remaining textual data was tokenized. Tokenization is typically the first step of NLP or Natural Language Processing, which is the process of spitting text into tokens or splitting paragraphs and sentences into smaller units.
- Stemming
  - Next, the textual data was stemmed, which is the process of removing the last few characters from a word to get to the base form or root
- Lemmatization
  - Lastly, similar to stemming, the textual data is put through lemmatization which considers the context and converts the word to its meaningful base form or root word.

Model Development

I. Vectorizing

Vectorizing is the process of converting words into numeric equivalents, which will be helpful for the tasks of classification and sentiment analysis in the case study. The following techniques were employed in the study:

- Binary Vectorizer
- Word2Vec Vectorizer
- TD-IF Vectorizer

II. Splitting the Data

The data was then split into 70% training and 30% testing.

III. Model Selections

The following models were selected to be tested:

- Multinomial Naive Bayes
- Decision Tree
- Random Forest
- Logistic Regression

IV. Feature Engineering

The feature, sentiment/score, was created to calculate the sentiment rating based on the 'rating/score' variable.

- (-1) Negative **rating ≤ 2**
- (0) Neutral **rating = 3**
- (1) Positive **rating ≥ 4**

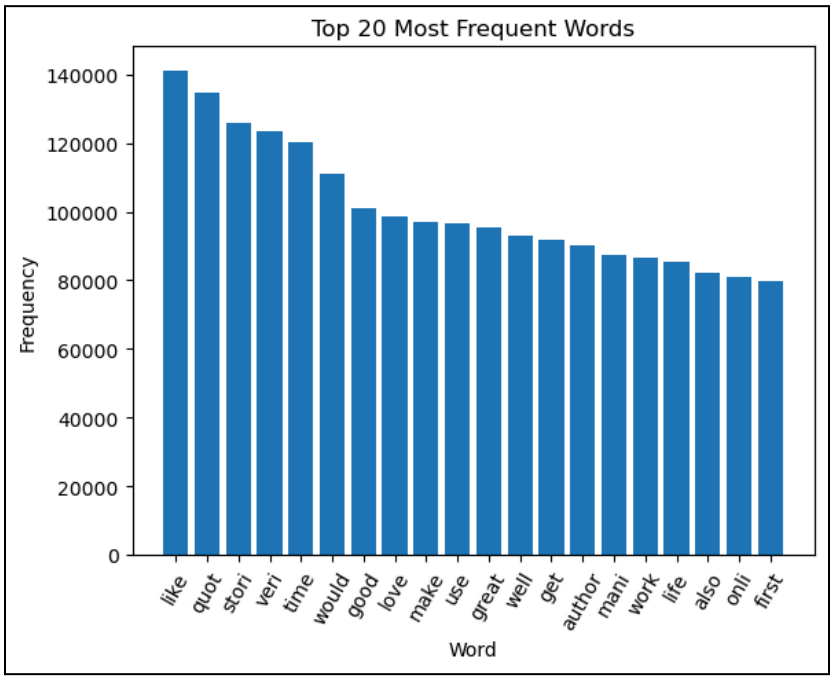
V. Handling Data Bias

To handle the skewness of the rating distributions, the dimensionality of textual data of positive and negative sentiment/score data was reduced to reduce bias.

Results and Evaluation

I. Term Frequency

This bar graph describes the Term Frequency in the dataset after removing words with no sentiment. Words representing positive sentiment are dominant, corresponding to the lower ratings' heavy skewness.



II. Logistic Regression Model - Rating and Reviews

The dataset was partitioned into five segments based on ratings 1-5, and the Logistic Regression Model was trained on each partition. As one can observe, the accuracy remained high across all partitions. At the same time, the recall plummets on ratings 2-4, indicating that high accuracy is due to the high True Negative rate. Precisions of ratings 2-4 are all below 0.5, suggesting that among

our positive predictions, less than half were positive. A low recall tells us that among all the positive cases, only a small fraction was captured by models with ratings lower than 5. The recall of rating 1 bounces back because the reviews use strong negative words instead of moderate or ambiguous words, allowing the model to capture more True Positives cases.

Rating	5.0	4.0	3.0	2.0	1.0
Precision	0.793	0.495	0.479	0.376	0.714
Recall	0.868	0.080	0.070	0.048	0.340
Accuracy	0.780	0.814	0.918	0.948	0.945

III. Review Sentiment Analysis

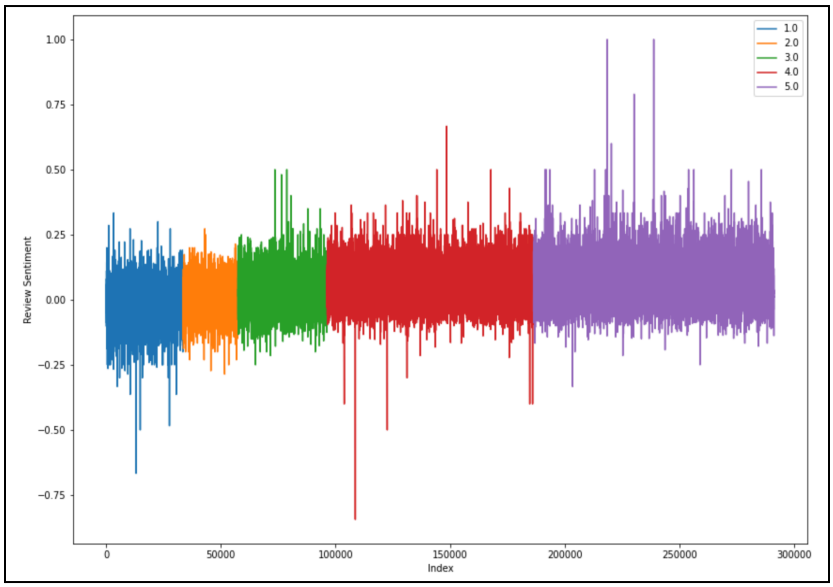
Since the raw data source contained more 5-star reviews than any other category, the results of the analyses would be biased, and to account for this skewness, rows of 5-star reviews were cropped. However, 4 and 5-star ratings were left as the most frequent response type to simulate real-world response patterns.

Following this, we created review sentiments based on the text responses of the user. Using NLTK's positive and negative word dictionaries, the overall sentiment of the text string was calculated for each review. Sentiment could be a value between -1 and 1, each being wholly negative and entirely positive, respectively, and zero being a neutral opinion.

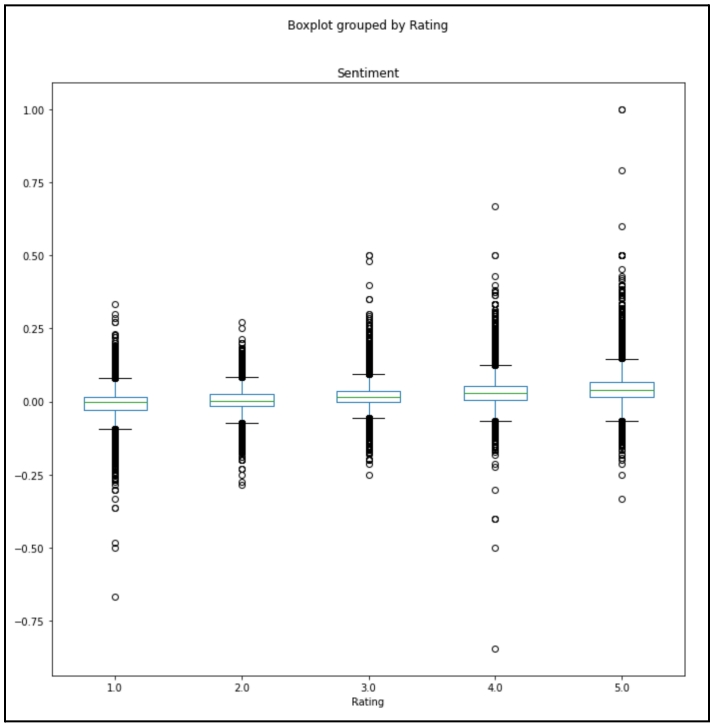
```
for i in range(len(mydf)):
#for row in mydf['Review']:
    row = mydf['Review'].iloc[i]
    total = len(row)
    pos = 0
    neg = 0
    for word in row:
        if word in positive_words:
            pos += 1
        if word in negative_words:
            neg += 1
    if total > 0:
        mydf['Sentiment'].iloc[i] = (pos - neg) / total
    else:
        mydf['Sentiment'].iloc[i] = 0
```

These sentiments were used with the rating frequencies to explore our data. The chart below shows the frequency of each rating by its overall sentiment. According to the chart, 4 and 5-star ratings are the most frequently occurring, while 2-star ratings are the

least occurring and have the most neutral reviews. On the other hand, 4-star ratings have the most negative reviews, followed by 1-star ratings. This can be used for a behavioral analysis: customers with issues with a book purchase are likelier to give a rating of 4 stars instead of 1 star; this may be because they see the issue as deserving of losing a single point only. As per our speculations, 5-star ratings have reviews with the most positive sentiment.

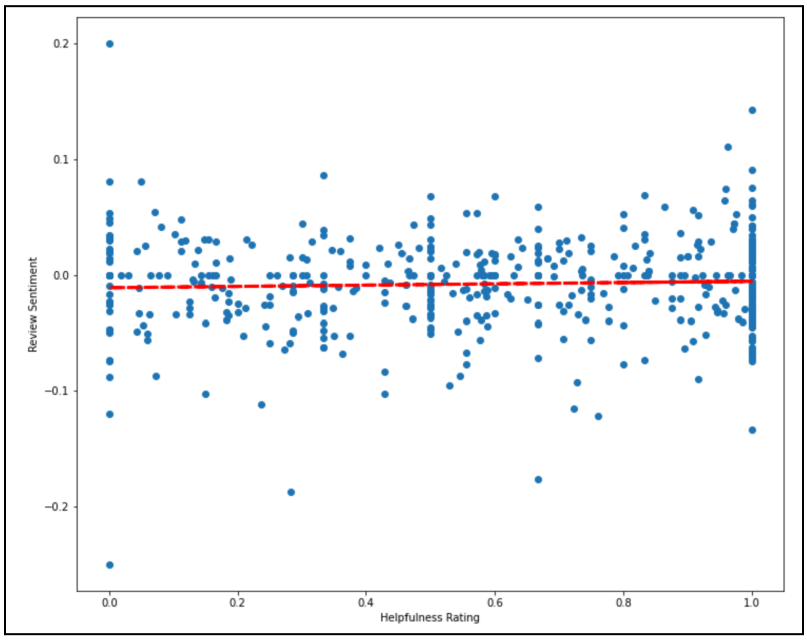


To give a clearer view of the sentiment distribution across different ratings, a Box-and-Whiskers chart was created that showed the sentiment of each review as part of a rating. As can be seen, all reviews have a median close to 0, with more reviews falling closer to the center. However, there are a few outliers that scale towards positive and negative. This graph shows that the sentiment value of the most positive rating is near 1.0 while the most negative is near -0.95.



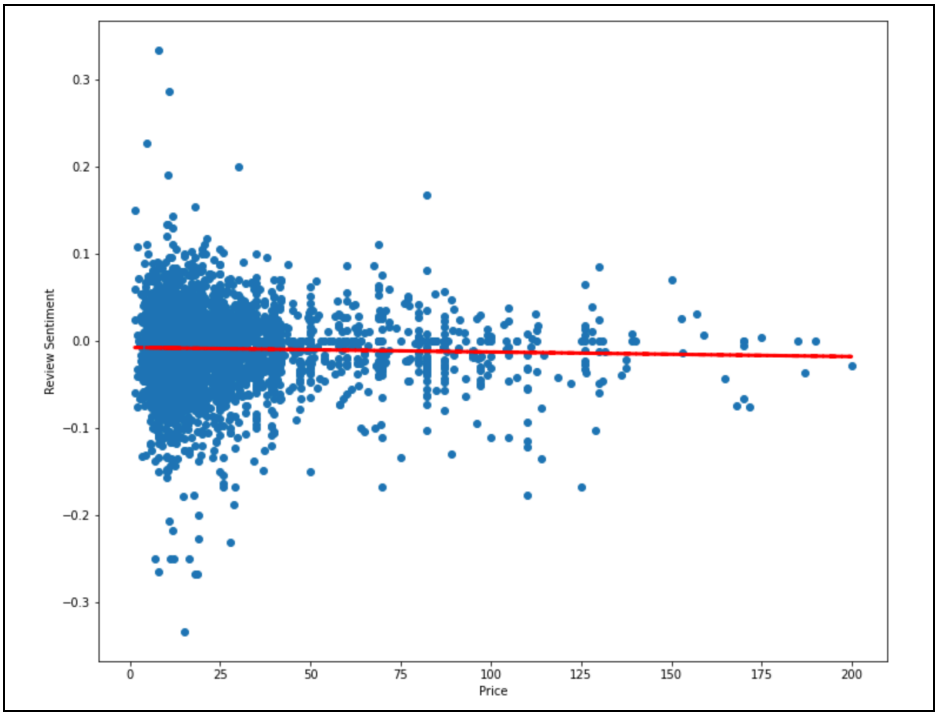
IV. Review Sentiment Analysis and Helpfulness Rating

We created a scatter plot and trend line of the sentiment scores by the helpfulness rating. According to the chart, there is no strong correlation between the review sentiment and the helpfulness rating. Sentiment dictionaries could be updated to include specific negative/positive words associated with book reading. The Helpfulness rating could be better defined. An issue that could have contributed to our results could have been our normalization process. Before normalizing the helpfulness ratings so that they are of a uniform measure, we could eliminate rows with helpfulness ratings that were difficult to infer meaning from, like “0/0”. This was normalized to 1.0 but could have also been normalized to 0.0. This uncertainty could be avoided altogether for our analysis.

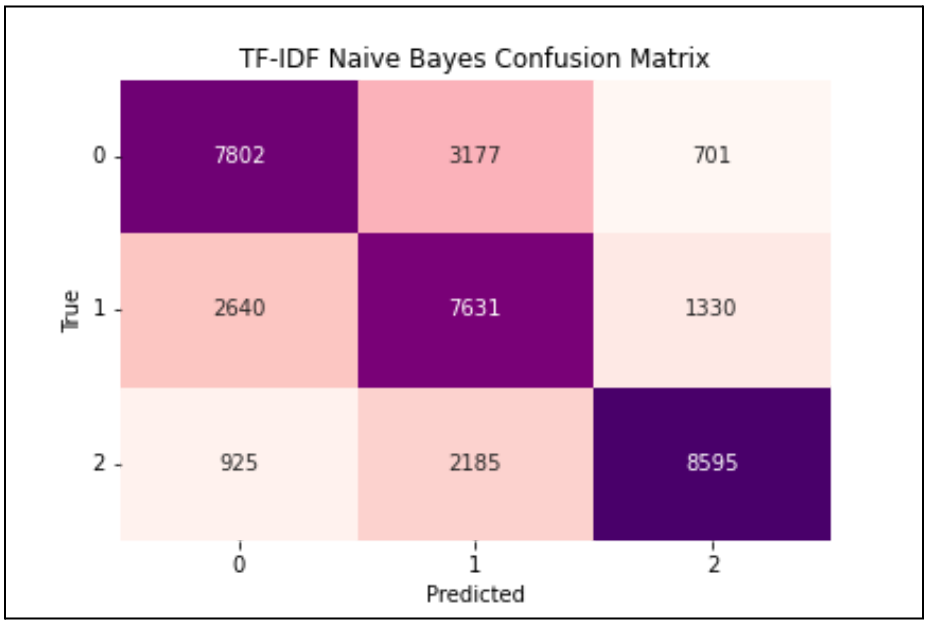


V. Review Sentiment and Price

On creating a scatter plot of review sentiment score by book price, there was no strong correlation between them, although there was a wide range of sentiments generated for cheaper books. This could be for several reasons: on the one hand, more expensive books tend to be more educational/scholastic, so opinions towards them are usually neutral and focused more on obtaining a particular result, or more people have access to less expensive books; thus it follows that the cheaper range of prices has more variation in the reaction towards them as they have more dots plotted. For future improvements, we noticed that the sentiment analysis did not pick up on context or meaning; thus, reviews speaking positively about a competing book would have positive sentiment scores. This could be resolved by including Noun Phrase analyses, Parts Of Speech processing, or a stronger word dictionary.



VI. Best Model Performer - Customer Reviews and Ratings



	Precision	Recall	Support	F1-Score
Negative	0.69	0.66	0.67	11754
Neutral	0.58	0.65	0.62	11671
Positive	0.80	0.74	0.22	11561
Accuracy	-	-	0.68	34986
Macro Avg.	0.69	0.68	0.68	34986
Weighted Avg.	0.69	0.68	0.68	34986

The TF-IDF, or Term Frequency, vectorizer in tangent with the Multinomial Naive Bayes was the model that performed the best. Based on the above Colored Confusion Matrix, one can see that the model predicts positive reviews with better accuracy than others. Also, due to the nature of the data, the overly positive bias, although balanced, the model still predicts positive sentiment the best.

Scope and Conclusions

Online reviews have become increasingly popular in today's world, particularly in the context of e-commerce and online shopping. As a result, the analysis of these reviews has gained significant attention in recent times. In this report, we have explored the relationship between the sentiment of online reviews and various factors such as rating, helpfulness, and price. We have also evaluated different models to find the ideal approach for sentiment analysis.

One of the key findings of our analysis is that there is a positive correlation between the sentiment score of a review and its rating, except for reviews with a rating of "4". Reviews with a rating of "4" were among the most negative, which is an interesting finding that warrants further investigation. The reviews rated "4" may represent a "neutral" sentiment, where the reviewer was not particularly impressed or disappointed with the product. Additionally, we found no significant association between the sentiment score and the helpfulness rating, which suggests that the helpfulness of a review is not necessarily related to its sentiment.

Another factor we explored was the relationship between the sentiment of a review and the price of the product being reviewed. Our analysis showed a weak correlation

between the two, with higher-priced books generally receiving neutral reactions. This finding suggests that consumers may be less likely to express extreme sentiment in their reviews for higher-priced products, possibly due to a greater sense of investment and a desire to justify the expense.

To identify the ideal model for sentiment analysis, we evaluated several approaches and found that using TF-IDF with MNB yielded the best results. This model achieved an accuracy of about 70%, with balanced Precision, Recall, Support, and F1 scores. TF-IDF (term frequency-inverse document frequency) is a common technique in natural language processing that helps identify the most essential words in a document by giving higher weight or scores to rare words that are more informative. MNB (Multinomial Naive Bayes) is a probabilistic algorithm that works well with sparse data, such as text data and has been widely used in sentiment analysis.

Overall, the findings of our analysis provide valuable insights into the relationship between the sentiment of online reviews and various factors such as rating, helpfulness, and price. By identifying the ideal model for sentiment analysis, we have also provided a valuable tool for businesses and researchers who wish to analyze online reviews. Future research could focus on further investigating the relationship between the sentiment of reviews and ratings of "4", as well as exploring the factors that influence the helpfulness of online reviews.

References

- Bekheet, Mohamed. “Amazon Books Reviews.” *Kaggle*, 13 Sept. 2022, www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews.