Natural Language Processing on Amazon Books Reviews Dataset

Group 1 - Dhruv Shah, Jenn Hong, Santiago Mazzei, Setu Shah, Victor Floriano

1. Business Objective / Analysis Plan

Our chosen approach begins with performing clustering to reveal initial insights, particularly aimed at Market Segmentation and Profiling through the book reviews. To achieve this, we've applied KMeans and Hierarchical Clustering methods, organizing the reviews based on varied criteria such as price, votes, and the total number of reviews a book has received.

Building upon this foundation, we've expanded our exploration into the realm of text mining and sentiment analysis. This step not only deepens our understanding of the underlying sentiments expressed within reviews—categorizing them into positive and negative sentiments—but also provides summary statistics that reflect the overall sentiment associated with each book. Additionally, sentiment analysis acts as a form of concise summarization, offering users a snapshot of a review's sentiment, thereby enhancing their experience by allowing them to grasp the essence of reviews without engaging with the entire text. Such a feature is not just a nod to efficiency, it significantly enriches the quality of time users spend on our website.

2. EDA and Preprocessing

Our initial task involved downsizing the extensive dataset to a manageable scale. The original dataset comprised over 20GB, and 4GB of metadata. To address this, we selected a subset of ~21,000 samples from the initial dataset and cleaned the metadata to include relevant details to our analysis. The original metadata did not have a genre column, so we had to extract the main genre from a list of different categories for each book. Following this procedure, we merged the book reviews with the selected metadata based on the "ASIN" code. This merger produced a consolidated dataset that integrated a representative sample of book reviews with the essential metadata.

We identified and corrected multiple issues in the dataframe, including dictionary-format "style" columns, null values, formatting inconsistencies, and incorrect data types, stemming from "json" files. We handled missing values by setting "vote" to 0 where upvotes were missing, replacing "style" nulls with "unknown", and filling missing "price" with the median value. Additionally, we enhanced the dataset with feature engineering, adding variables for total reviews, and extracting "year", "month", and "day_of_week" from "reviewTime". To conclude this stage, we carefully selected the variables (List 1) most relevant for our subsequent clustering analysis.

3. Clustering Methodology

After standardizing our data, we applied PCA to the dataset and assessed the cumulative explained variance for each component. Our analysis revealed that retaining ~94% of explained variance required 6 of 7 components. (Figure 1) Given the minimal number of features in our dataset, we deemed the loss in interpretability from PCA unnecessary. Moving on to KMeans, we first used an elbow plot (Figure 2) to determine the optimal number of clusters. While the elbow was not clear, there was a definite kink at 5 clusters. Given that a subsequent silhouette plot (Figure 3) was also inconclusive, we determined 5 clusters to be an appropriate starting point.

Then, we explored hierarchical clustering, determining from silhouette score (Figure 4) that the best number of clusters for hierarchical clustering was 6 clusters. However, given we had used 5 clusters for KMeans, and the difference in silhouette score was not great between 5 and 6 clusters, we decided to stick to 5 clusters to stay consistent. Both KMeans and hierarchical clustering ended up yielding the same 5 clusters, showing that neither method was superior over the other.

3.1 Clustering Results

The analysis of our book review database reveals different characteristics across five clusters, offering insights into reader engagement, credibility of reviews, and book pricing. The most distinct was Cluster 1 and Cluster 5. Cluster 1 features books with a high average vote count of ~324, much higher than other clusters, suggesting that all the reviews that were deemed helpful by others ended up in this cluster. Meanwhile, Cluster 5 is unique for its small number of highly expensive books, with an average price of $690.15. This cluster demonstrates a niche market where expensive books consistently attract attention and favorable reviews. These two clusters hence show two different types of power users, one for usefulness and credibility, and one for high spend, in the data.

4. Sentiment Analysis Methodology

In our project, we conducted sentiment analysis, beginning with the preprocessing of our primary dataset's 'reviewText' column. This process involved converting all text to lowercase, eliminating non-alphanumeric characters, removing stop words, tokenizing the texts into individual words, and applying lemmatization to reduce words to their base forms. Following this initial data preparation, we established a methodology for our sentiment analysis based on the following steps:

4.1 Construction of Ideal Sentiment Vectors: We developed two benchmark sentiment vectors to serve as standards for comparison: one representing ideal positive sentiment and the other ideal negative sentiment. These vectors were compiled using lists of words sourced from an online database (see citations). Additionally, we selected two representative reviews from our dataset, one clearly positive and the other clearly negative, as detailed in the Appendix (List 2).

4.2 Model Score Definition: To evaluate the effectiveness of our sentiment analysis models, we defined a unique metric known as the *Model Score*. This score is calculated by determining the sentiment of the positive review and subtracting the sentiment of the negative review. More technically, the *sentiment score* is derived from the difference in cosine similarities between the review text and the ideal sentiment vectors – one positive and one negative. This approach allows us to quantitatively measure, through *Model Score*, how effectively each model can separate the hand-picked positive review from the negative. (Equation 1)

4.3 Model Evaluation and Comparison: We assessed three distinct models: a Word2Vec model trained on our dataset, a pre-trained Word2Vec model based on Google News data, and a pre-trained GloVe (Global Vectors for Word Representation) model with 300-dimensional vectors. For our

custom Word2Vec model, we experimented with 27 different sets of hyperparameters to identify the optimal configuration. Based on *Model Scores,* our best performing model was the GloVe model, so we used this model to generate a *sentiment score* (as defined above) for each review in our dataset.

4.4 Sentiment Analysis Results

Upon applying our best model to our reviews dataset, we see that there are very little outright negative reviews and most reviews ended up in the 0.2 to 0.4 range. There is a clear positive linear relationship between 5-star ratings and sentiment analysis as well (Figure 5), and mixed media book styles performed stronger than text styles (Figure 7). The top performing genres are primarily non-fiction, while fiction books scored middle of the pack. Genres that encourage controversy (History, Parenting) and more dry genres (Engineering, Reference) scored on the lower end of the spectrum. (Figure 8)

5. Limitations And Next Steps

Many books have multiple versions that would have different ASIN numbers, making analysis by book title less effective as they would be multiple unique entries that all refer to the same book text. Furthermore, in the specific context of book reviews, we found aggregation of word sentiment to obtain overall review sentiment not to be very effective. Oftentimes, reviews that clearly indicated their dislike and lack of recommendation for a certain book still contained many positive words, especially if it included discussion about plot (Table 1), leading to miscalculation. Context hence was not distinguished. Having a model that could put weights on certain keywords or identify key sentences would thus be more effective at extracting overall sentiment. Future models could also weight the sentiment scores with the overall rating scores to create more scale between the highly rated reviews and the less well-rated reviews, as many of the books with the lowest sentiment scores actually had some of the highest ratings (4.7), indicating further issues with the model. Additionally, a lot of our sentiment scores were relatively close to each other, and the differences are likely not statistically significant. To form definite conclusions on if a genre performed better than another for concrete recommendations, we would likely have to scale this project up to include the full breadth of reviews included in the dataset to increase statistical power.

6. Next Steps/Business Implications

This project delved into the utilization of NLP techniques for analyzing Amazon book reviews. Through clustering, we identified two groups of power users, those with high verification and those who purchase books at high prices, which can be utilized for customer segmentation and targeting. Via sentiment analysis, we identified top genres, book styles, and titles, which can be fed into recommendation algorithms for marketing strategies, content recommendations, and product development. Moreover, these insights could influence publishing houses' production decisions regarding the styles of books they choose to produce. Further exploration into specific book categories or sentiment patterns could yield deeper understanding and actionable recommendations. These sentiment scores can also be used in the review interface to give prospective buyers more insights on the book overall.

References

“Amazon Review Data.” Nijianmo.github.io, nijianmo.github.io/amazon/. Accessed 3 Mar. 2024.

Jianmo Ni, Jiacheng Li, Julian McAuley Empirical Methods in Natural Language Processing (EMNLP), 2019

Minqing Hu and Bing Liu. "Mining and Summarizing Customer Reviews." Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004), Aug 22-25, 2004, Seattle, Washington, USA,
Bing Liu, Minqing Hu and Junsheng Cheng. "Opinion Observer: Analyzing and Comparing Opinions on the Web." Proceedings of the 14th International World Wide Web conference (WWW-2005), May 10-14, 2005, Chiba, Japan.

Generative AI Disclosure

For our project gen-AI was used in the following capacity:

- To do grammar checks, improve sentence structure, and summarization.
- Learn how to convert specific arrays in the original json files into DataFrame columns.
- Learn how to convert a DataFrame in Colab into an CSV file and export that file for later use.

Appendix

List 1: Variables selected for Clustering Analysis

- "overall": the review score (1 to 5)
- "verified": whether the review was verified or not (binary)
- "vote": how many upvotes did each review received
- "price": the price of the book being reviewed
- "total_review_count": the total number of reviews for the book being reviewed
- "year": year of the review
- "day_of_week": day of the week for the review
- "month": month of the review

Figure 1: PCA Cumulative Variance Plot

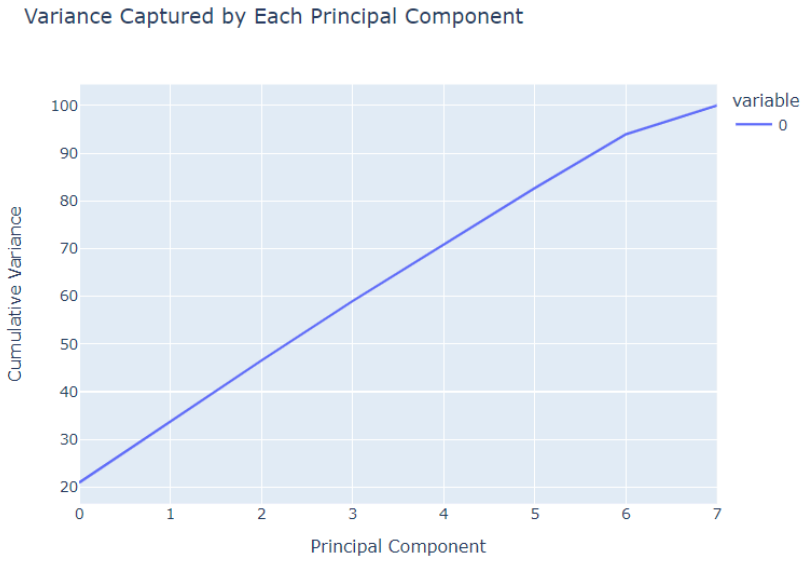


Figure 2: Elbow Plot

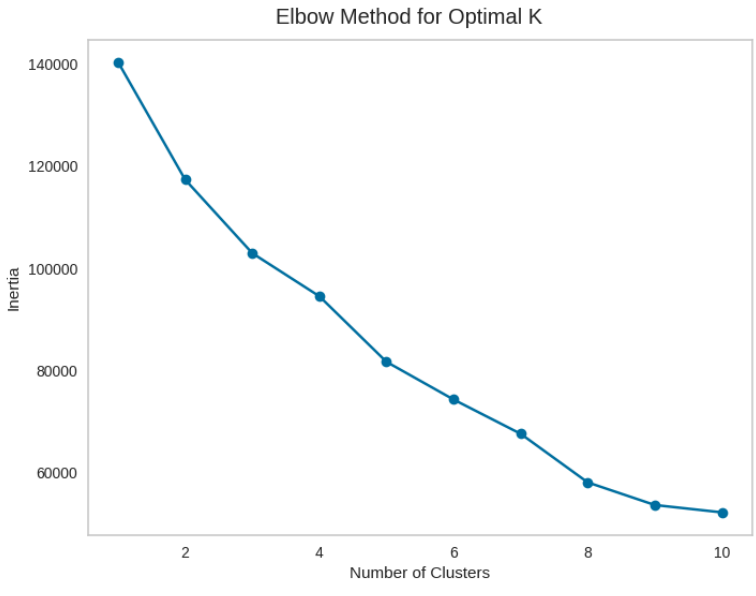


Figure 3: Silhouette Plot

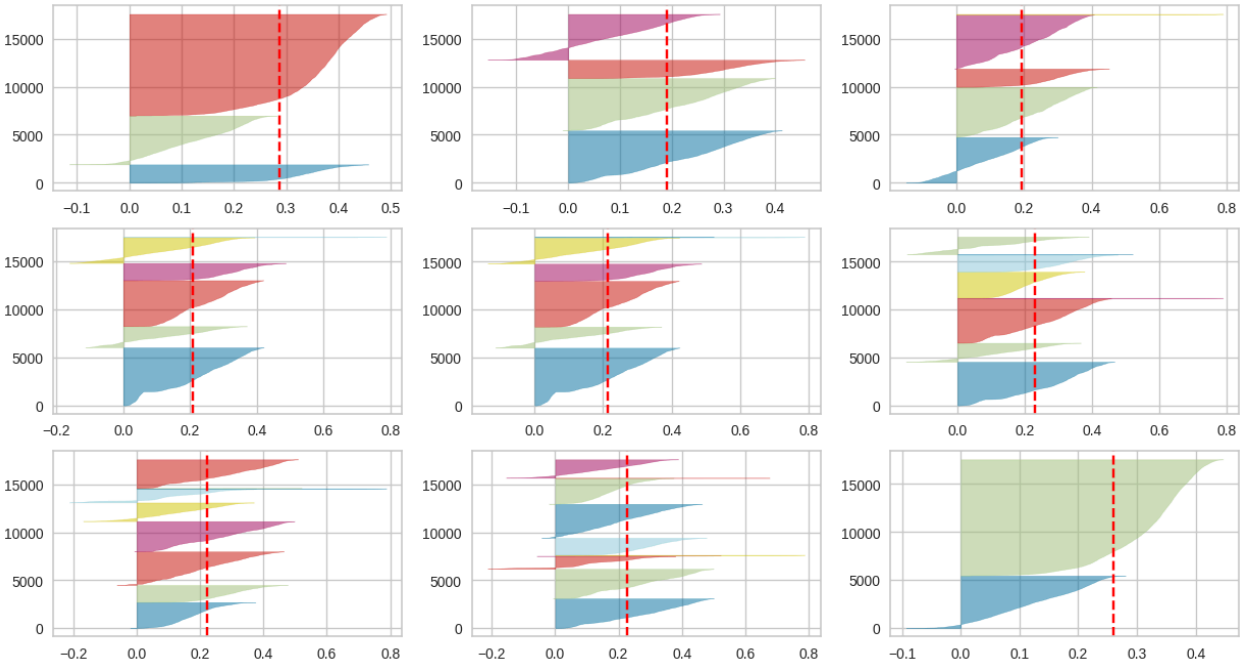
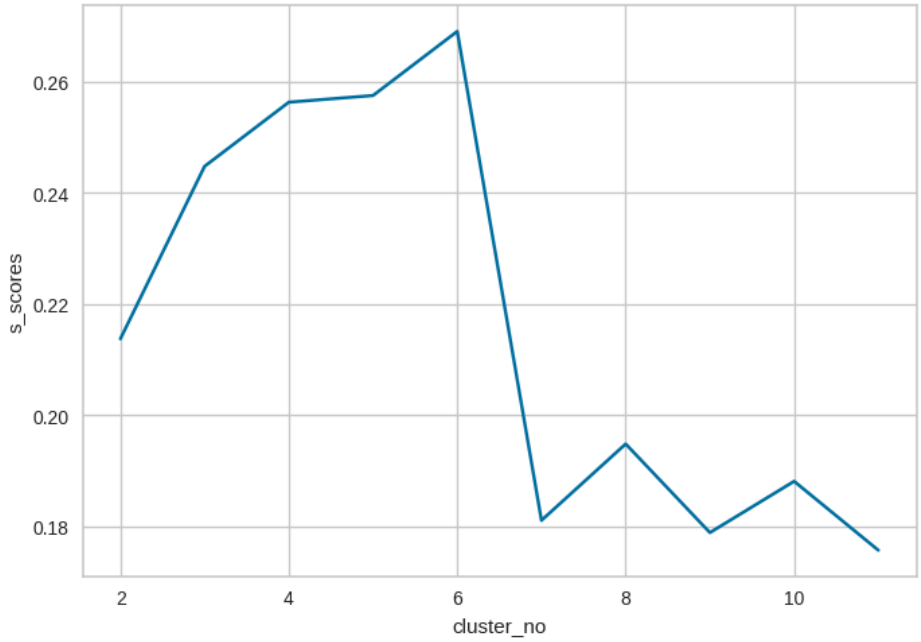


Figure 4: Silhouette Score Plot Per Cluster



List 2: Extracted Positive and Negative Benchmark Reviews from the Dataset

- **Positive Review:** “Love to read. Did not know this was now a movie until after I read the book. If you like this era this is a book you will not be able to put down. It is an easy read and absolutely wonderful.”
- **Negative Review:** “Unfortunately, The Fountainhead was ponderous, overly pretentious, preachy, self-indulgent and hollow. It is little more than a puffed-up facade with a whole bunch of really boring words.”

Equation 1: Formula for Sentiment Analysis Scoring

Model Score = *sentiment score*(Positive Review) - *sentiment score*(Negative Review), where

sentiment score = cosine similarity(Review, Ideal Positive) - cosine similarity(Review, Ideal Negative)

Figure 5: Mean Sentiment Score by Rating

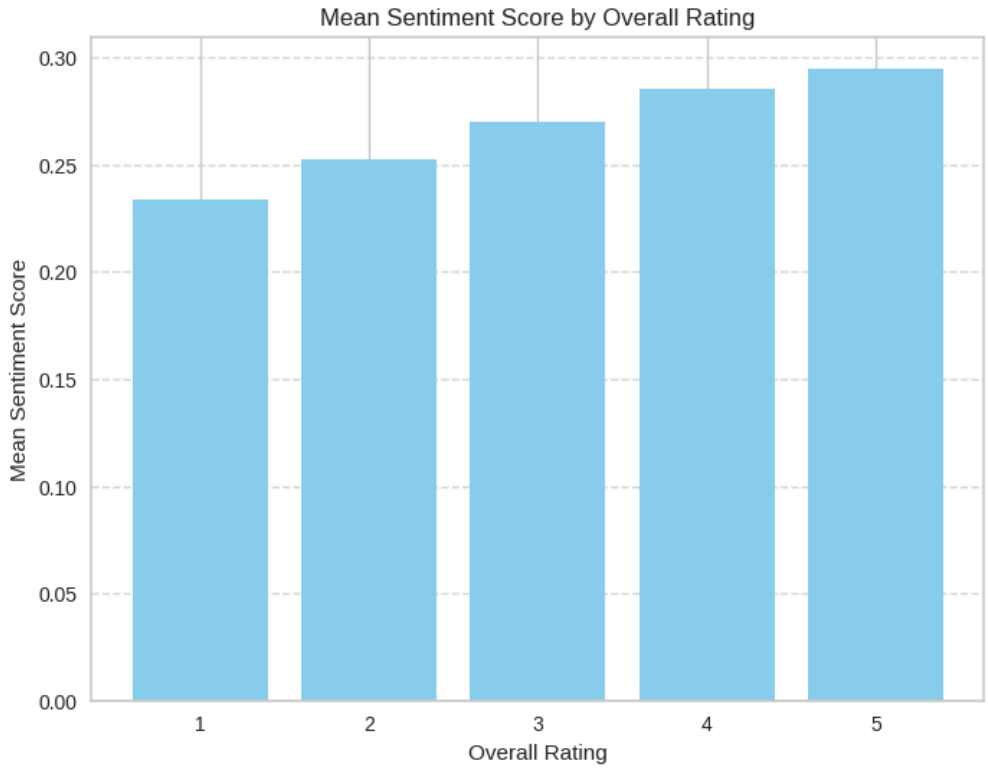


Figure 6: Sentiment Scores by Verification

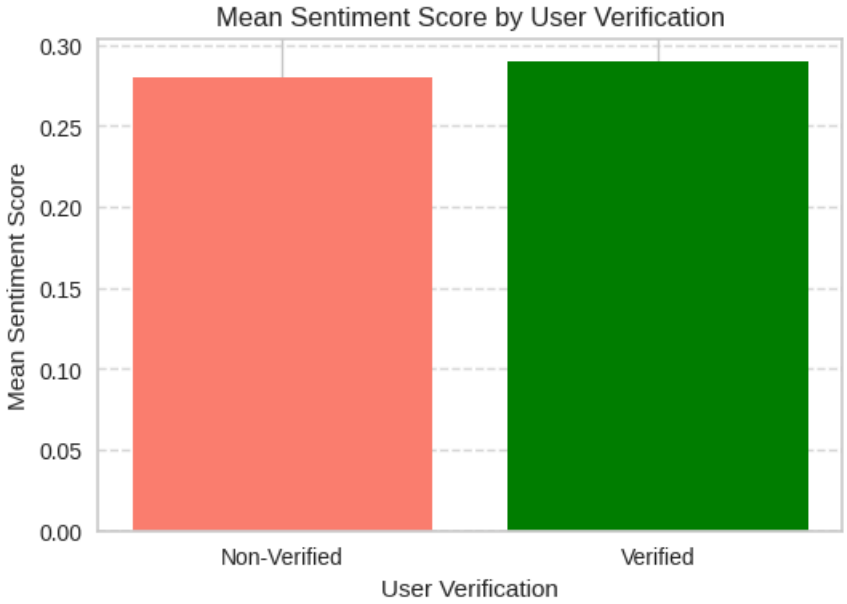


Figure 7: Sentiment Scores of All of the Book Style

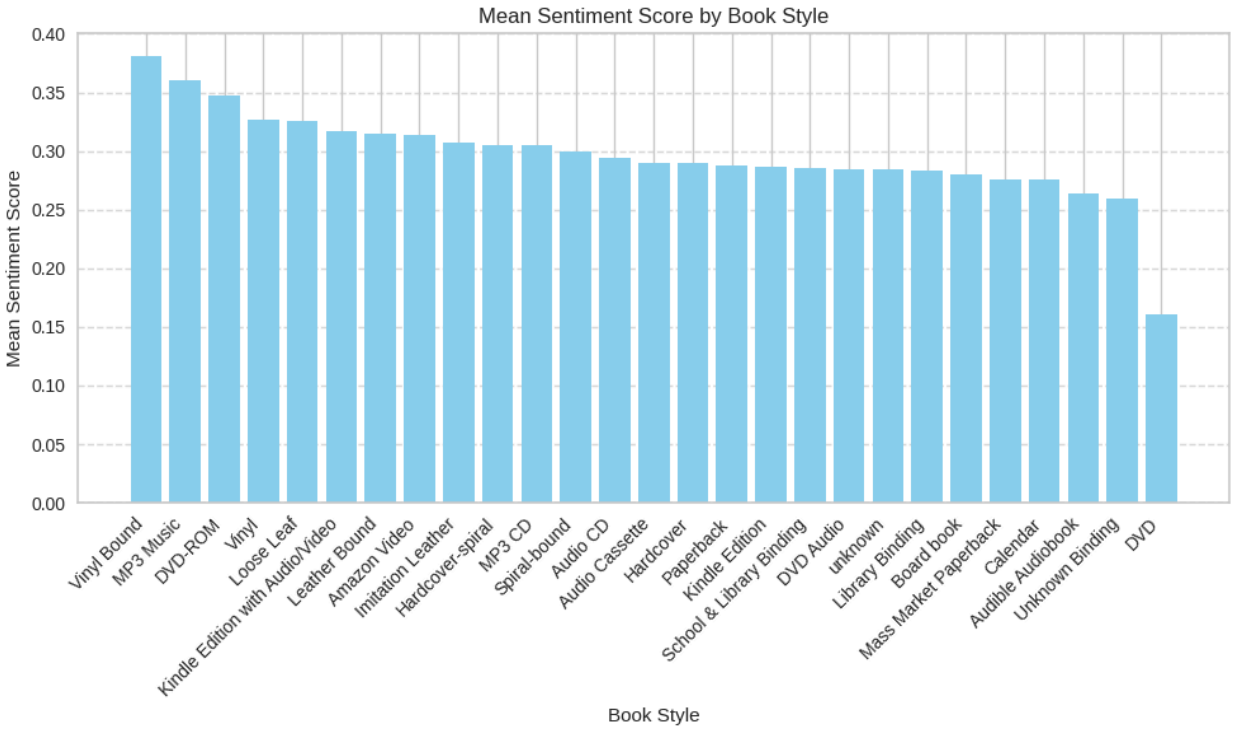


Figure 8: Sentiment Scores of All of the Genres

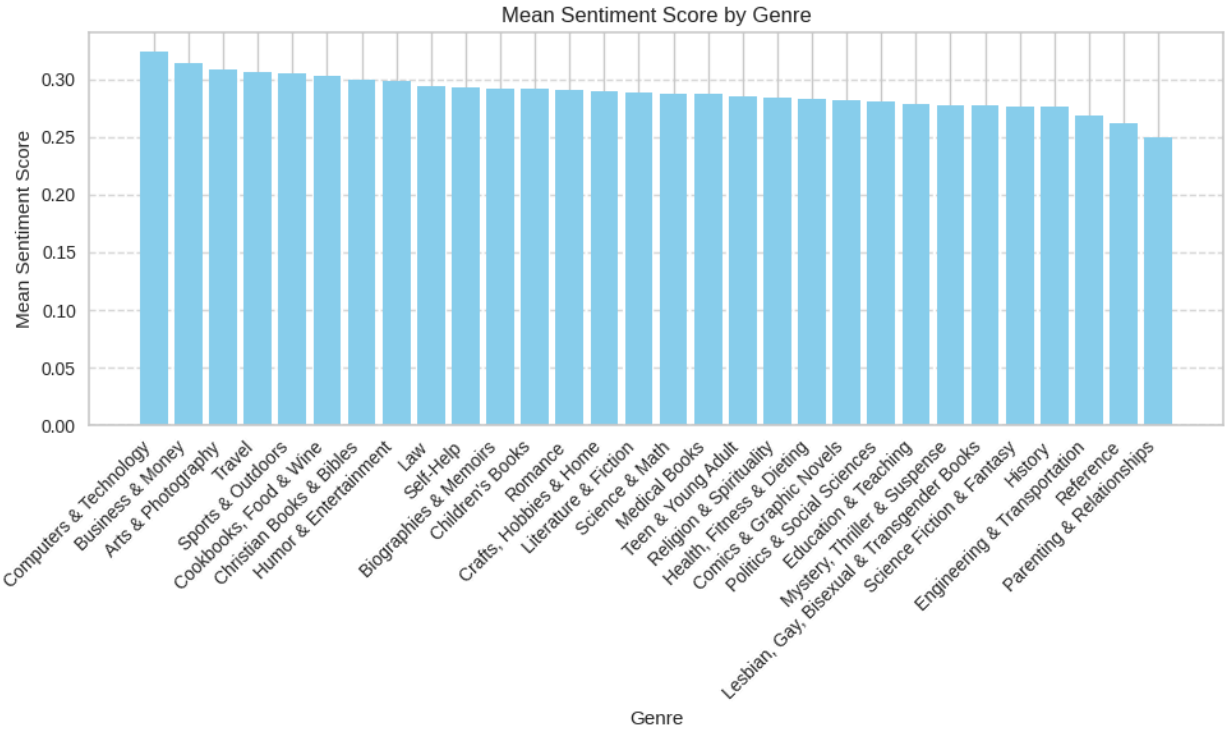
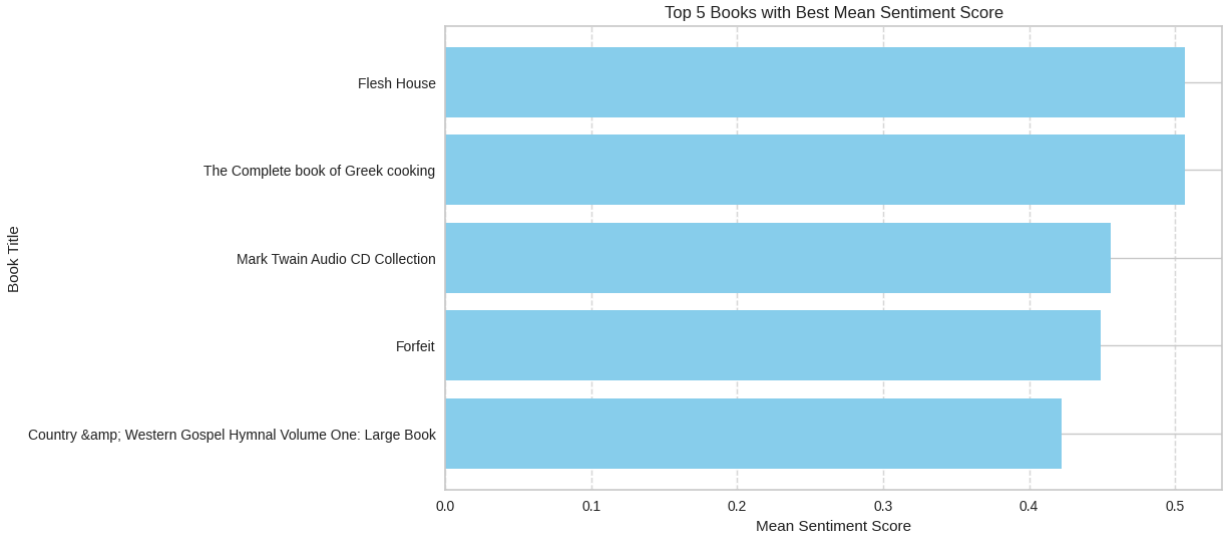


Figure 9: Top 5 Book with the Highest and Lowest Sentiment Scores

For specific titles, the best performing book was Flesh House by Stuart MacBride (0.51), The Complete book of Greek cooking (0.51), and Mark Twain Audio CD Collection (0.46). The worst performing books were Loyalties: A Novel of World War II (0.065), The Wisdom of No Escape and the Path of Loving Kindness (0.083), and The Ipcress File (0.099). (Figure 7) Note that even the worst performing book had an overall positive score.



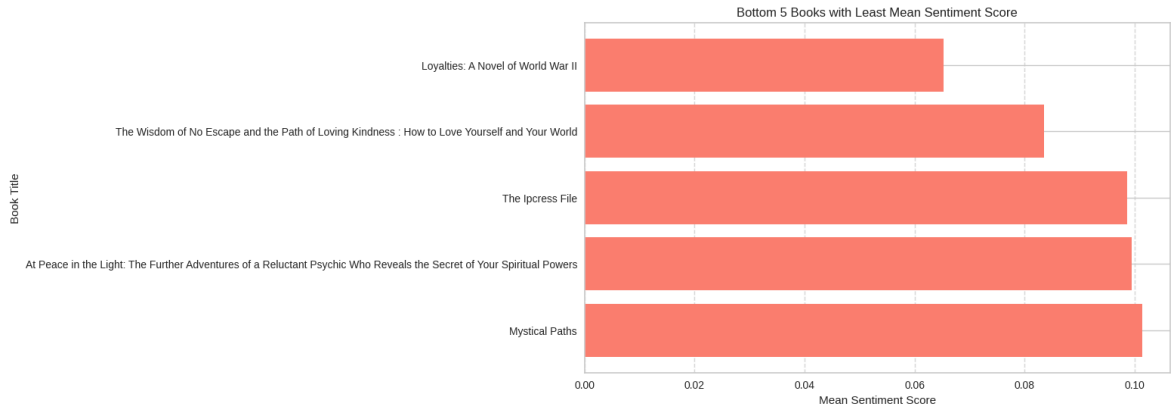


Table 1: Examples of Sentences

ReviewText	Sentiment	Number of opposite sentiment words in the review
Stuart MacBride's third installment in the Logan McRae series is one of the darkest, most gruesome, grisly books you will ever read.	Positive	3
This may be a classic, but it is very boring and hard to follow in the 21st century. Maybe if you are an English major and not an anthropologist as I am this may be more entertaining.	Negative	3
The first part of the novel, which takes place in our world, is excellent. The main-character contracts leprosy and Donaldson recreates the daily life of a leper (and the scorn he is subject to) in a truly convincing way. Highly original and interesting first 40 pages The problems start when Covenant is hit by a car and transported to the fantasy world. From here onwards the book just sucks big time.	Negative	8
If the company is in a feeling of doom then your emotions fall and the book drags but you just can't put it (the book) away. Brilliant, too bad Tolkien didn't live longer to write us more books.	Positive	6