FA 550 – DATA VISUALISATION

PROJECT REPORT

TITLE: Kindle Reviews: A Data-Driven Analysis

Kirtana Sridharan

CWID: 20011879

**INTRODUCTION**

The surge in e-commerce platforms has revolutionized the way consumers interact with products and make purchase decisions. Within this landscape, online reviews have emerged as a crucial source of information, influencing consumer choices and shaping the reputation of products. This paradigm shift holds particularly true in the realm of books, where avid readers often turn to online platforms like Amazon for insights before deciding on their next literary adventure.

The abundance of book reviews on Amazon offers a treasure trove of information, capturing the sentiments, preferences, and experiences of diverse readers. Understanding this wealth of data is pivotal for both readers seeking informed choices and authors and publishers striving to comprehend and improve the reception of their work.

**Introduction to the Dataset:**

**DATASET:** <https://cseweb.ucsd.edu/~jmcauley/datasets/amazon_v2/>

The chosen dataset revolves around Amazon Kindle book reviews, providing valuable insights into customer sentiments and preferences regarding various literary works. It encompasses diverse information, including the book's ASIN (Amazon Standard Identification Number), overall rating, reviewer ID, review text, product-related details (such as language, format, and product rank), and additional metadata.

The dataset comprises Amazon Kindle book reviews, featuring the following:

* key fields such as reviewerID, asin (Amazon Standard Identification Number),
* reviewerName,
* helpfulness metrics,
* reviewText,
* overall rating,
* summary,
* and timestamps (unixReviewTime and reviewTime).

The meta data consists of the following columns:

* category – category of books
* description
* title
* brand
* rank – rank of the product in the Amazon Kindle catalog
* details – such as book length(pages), language, published date, book size(MB)
* asin

**Purpose of Choosing This Dataset:**

The selection of this dataset aligns with the goal of understanding customer sentiments and behavior in the context of book reviews. Analyzing Amazon book reviews allows for a comprehensive exploration of factors influencing the success of literary works, such as the impact of language, format, and other attributes on customer satisfaction. Additionally, the dataset enables the application of natural language processing techniques for sentiment analysis, offering a deeper understanding of the emotional tone expressed in reviews.

**Research Questions**

1. Is there a correlation between the overall rating given by reviewers and the sentiment labels assigned to the reviews?
2. Does the sentiment score of reviews vary across different book formats, and are there specific patterns within positive, negative, and neutral sentiment labels?
3. How has the sentiment of reviews changed over time, and are there noticeable trends? How has the volume of reviews changed over time?
4. Are certain book categories more prone to receiving positive or negative sentiments?
5. Can we visualize the distribution and density of reviews across different book categories?
6. What is the distribution of books across various languages?

**TOOLS USED**

In this project, Python and Tableau were used for an effective and comprehensive analysis of the dataset. The following summarizes the key steps taken to clean the data, extract relevant information, and create visualizations to answer specific research questions:

**Python:**

1. Data Cleaning:

* Used Python, along with pandas library, to load and explore the dataset.
* Identified and handled missing values, and duplicates, to ensure data integrity.

2. Text Analysis and Sentiment Classification:

* Employed spaCy, a natural language processing library, to tokenize and analyze the sentiment of reviews.
* Utilized a pre-trained sentiment analysis model from Hugging Face to perform sentiment analysis on reviews
* Classified reviews into positive, negative, and neutral sentiments based on their content.

3. Feature Engineering:

- Extracted additional features such as book length, book format and categories, to enrich the dataset.

4. Subset Creation:

* Created subsets of the data based on specific criteria, such as particular time periods, to focus on relevant aspects of the analysis.

5. Keyword Extraction:

* Utilized spaCy for keyword extraction from the 'reviewText' column, focusing on nouns and adjectives while excluding stop words and punctuation.
* Generated separate sets of keywords for positive, negative, and neutral sentiments.

**Tableau:**

1. Data Connection:

* Connected Tableau to the cleaned and enriched dataset created using Python.

2. Dashboard Creation:

* Designed dashboards in Tableau to visualize key insights to answer the listed research questions.
* Grouped visualizations logically to answer specific research questions effectively.

3. Exploratory Data Analysis (EDA):

* Created various plots, charts, and graphs to explore relationships and patterns in the data.
* Used Tableau's interactive features for dynamic exploration and understanding of the dataset.

4. Sentiment Analysis Visualizations and Temporal Trends:

* Employed Tableau to visualize sentiment analysis results, showcasing the distribution of positive, negative, and neutral sentiments.
* Investigated temporal trends by creating time-series visualizations in Tableau, analyzing how sentiments change over time.

6. Word Clouds:

* Integrated Python-generated keyword data into Tableau for word cloud visualizations, providing a visual representation of frequently used words for each sentiment.

7. Insights Presentation:

* Presented insights and findings through Tableau story, facilitating a clear and comprehensive understanding of the dataset.

By combining the strengths of Python for data manipulation, analysis, and preprocessing with Tableau's powerful visualization capabilities, a robust and insightful analysis was conducted to answer specific research questions effectively.

**EXPLORATORY DATA ANALYSIS**

The exploratory data analysis (EDA) of the dataset involved key visualizations to gain insights into various aspects of customer reviews.

1. Distribution of overall ratings provided by the customers.

The distribution of ratings was examined through a countplot, revealing the frequency of each overall rating. This provided a fundamental understanding of the distribution of customer sentiments.

A bar graph with different colored squares

Description automatically generated

Verified vs Unverified Reviews.

The distinction between verified and unverified reviews was explored, offering insights into the credibility of feedback. The countplot illustrated the distribution between these two categories.

A graph with a blue and orange rectangular bar

Description automatically generated

1. Review Length Distribution.

The distribution of review lengths was visualized using a histogram. This analysis showcased the variability in the length of reviews, providing valuable context for the subsequent sentiment analysis.

A graph of a review length distribution

Description automatically generated

**VISUALISATIONS TO ANSWER THE RESEARCH QUESTIONS**

1. **SENTIMENT OVERVIEW**

**QUESTION 1: Is there a correlation between the overall rating given by reviewers and the sentiment labels assigned to the reviews?**

A graph of a bar chart

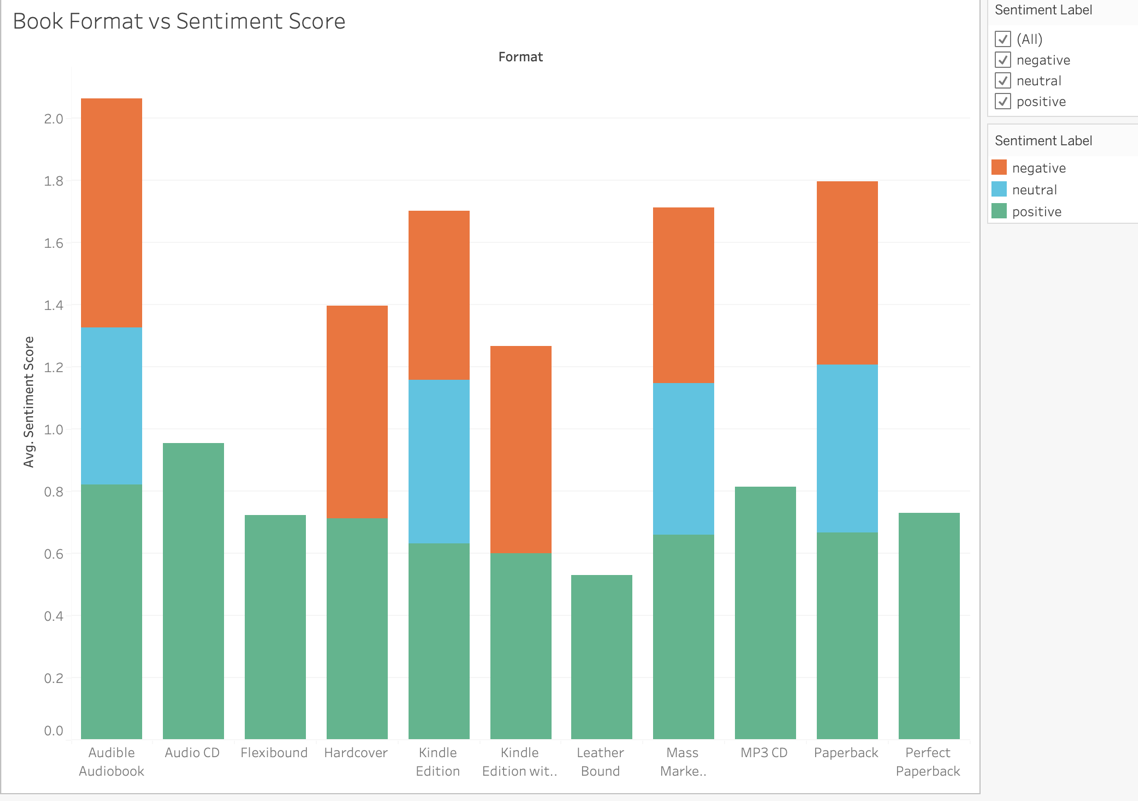
Description automatically generated with medium confidence

There is a weak to moderate positive correlation between the overall rating given by reviewers and the sentiment labels assigned to the reviews. This means that reviews with positive sentiment labels tend to have higher overall ratings, while reviews with negative sentiment labels tend to have lower overall ratings. However, there is also a lot of scatter in the data, so it is important not to overstate the relationship. For example, there are many reviews with positive sentiment labels that have low overall ratings, and vice versa.

* **Reviewers may use sentiment as a factor when they give an overall rating.** In other words, reviewers may be more likely to give a high rating to a review that they feel is positive, and vice versa.
* **Sentiment labels may be a good indicator of the reviewer's overall experience with the product or service.**

**QUESTION 2: Does the sentiment score of reviews vary across different book formats, and are there specific patterns within positive, negative, and neutral sentiment labels?**

The chart you sent me shows the average sentiment score of reviews for different book formats, as well as the average sentiment score of each sentiment category (positive, neutral, and negative).



**Does the sentiment score of reviews vary across different book formats?**

No, the sentiment score of reviews does not appear to vary significantly across different book formats. The average sentiment score for all formats is around 1.2, and the scores for individual formats are all within a range of about 0.5-0.8points. This suggests that reviewers' overall feelings about the book are not influenced by the format in which they read it.

**Are there specific patterns within positive, negative, and neutral sentiment labels?**

There are some interesting patterns within the positive, negative, and neutral sentiment labels. For example:

* Positive sentiment labels: The percentage of positive reviews is highest for Perfect Bound, Audio CD and Flexi bound formats, and lowest for Leather Bound and Kindle formats. This suggests that people who prefer either reading soft cover books or listening to audio books, leave positive rating rather than the ones who opt for Kindle books.
* Negative sentiment labels: The percentage of negative reviews is highest for Kindle formats, and lowest for Paperback and Mass market formats. This suggests that people may be more likely to leave negative reviews for books that they find difficult to read or format.
* Neutral sentiment labels: The percentage of neutral reviews is highest for Audible and MP3 CD formats, and lowest for Audible and Paperback formats. This suggests that people may be more likely to have a neutral opinion of a book if they listen to it as an audiobook.
* Physical books may be seen as more valuable or special than audiobooks, so people may be more likely to leave positive reviews for them.
* Audiobooks may be seen as a more convenient way to consume content, so people may be more likely to leave neutral reviews for them.

1. **TEMPORAL TREND ANALYSIS**

**QUESTION 3: How has the sentiment of reviews changed over time, and are there**

**noticeable trends?**

A screenshot of a graph

Description automatically generated

Although the volume of reviews for each sentiment has diminished over time, the average sentiment score has exhibited an upward trend across the years. This indicates that, while there are fewer explicitly positive or negative opinions, the intensity of the sentiments in remaining reviews has increased. The combination of a decrease in the number of negative reviews and an increase in average sentiment scores signifies a positive shift in the overall sentiment landscape over the specified time frame.

**Overall, the number of categories of books reviewed increased over time between 2013 and 2015 and a sharp decrease from the years 2016-2018.** The graph shows a steady upward trend in the number of genres reviewed, from around 1600 in 2013 to over 3200 in 2015.

**Some categories have seen more growth than others.** The categories that have seen the most growth in the number of reviews include Literature & Fiction, Foreign Languages, Religion & Spirituality, and Children's eBooks. And these categories seem to have the highest number of reviews over the 6 years as compared to the other categories.

1. **CATEGORICAL INSIGHTS**

**QUESTION 4 & 5: Are certain book categories more prone to receiving positive or negative sentiments? Can we visualize the distribution and density of reviews across different book categories?**

**A close-up of a graph

Description automatically generated**

Categories with a higher average sentiment score for positive sentiment are likely well-received by readers. This could be attributed to engaging content, well-written narratives, or other positive factors.

Literatue & Fiction have the highest number of reviews, followed by Foreign Languages and books on Religion.

1. **LANGUAGE EXPLORATION**

**QUESTION 6: What is the distribution of books across various languages?**

**A graph with blue and white bars

Description automatically generated**

**Overall Distribution:**

The chart shows the number of books in each language, with the average score line drawn as a reference band. While German and English dominate, there are also a number of other languages with a notable presence.

* **German** has the most books by far, with close to **2,300 books**.
* Following German are **English**, **Spanish**, **French**, and **Portuguese**, all with over **600 books** each.
* The number of books then drops significantly for other languages, with most having less than **200 books**.
* Notably, there are **over 200 books** in **Afrikaans**, **Dutch**, **Chinese**, **Japanese**, and **Russian**.

It's important to acknowledge the rich tapestry of literature available in other languages. We do see a wide variety of books in various other languages, offering diverse perspectives, stories, and voices from around the world.

1. **WORD CLOUD VISUALISATION**

A group of circles with words

Description automatically generated

Advanced language analysis was employed to extract the most frequent words associated with positive, negative, and neutral sentiments in the reviews. This provided a window into the thematic and emotional undercurrents driving each type of response.

However, limitations inherent to sentiment analysis models must be acknowledged. These models may not always capture the full range of human emotions, and certain words can appear across different sentiment categories depending on context.

Taking the word “author” as an example, reviews expressing appreciation often highlight the author's masterful storytelling, as in the phrase "The author's masterful storytelling kept me captivated from the first page to the last." Conversely, disappointment may arise from the author's execution of the story, as in reviews criticizing "unfulfilled potential" or "characters lacking depth."

**CONCLUSION**

In summary, our exploration into book reviews has uncovered several valuable findings. We've delved into the correlation between sentiment labels and overall ratings, revealing intriguing patterns over time. The influence of book format, language, and length on sentiment has been examined, providing insights for publishers and authors.

Our keyword analysis has identified prevalent themes in positive, negative, and neutral reviews, offering a deeper understanding of reader sentiments. By harnessing both Python and Tableau, we've created interactive visualizations for a user-friendly exploration of the dataset.

By examining the relationship between language and sentiment labels, we've identified nuances in reader perceptions across different linguistic domains. The exploration of language provides an additional layer of understanding, allowing publishers and authors to tailor their content for diverse audiences. Whether it's discerning sentiment trends in English, uncovering unique patterns in other languages, or optimizing language choices for specific book categories, this report underscores the importance of considering linguistic factors in the literary landscape.

Through a harmonious integration of Python and Tableau, our analysis offers a comprehensive view of the dataset, facilitating informed decisions for stakeholders in the publishing industry. This report stands as a testament to the value of language-aware analytics in shaping the future of literature.