Noman Tahir
S2notahi@uni-trier.de
1655735
Topic:
Leveraging Topic Modeling in Book Reviews for Digital Humanities Research.

Note for readers.

All Python scripts used for analysis are hosted on GitHub, with direct links provided for each result discussed in this paper. For interactive diagrams, corresponding links are provided, allowing full access to dynamic visualizations. Static .png diagrams, as presented in this paper, include only the essential code for generating these images. The interactive versions can be explored online through the provided links for a more comprehensive understanding of the findings.

All results and scripts can be access from here.
github.com/nomanqureshi1/Topic-Modeling-inBook-Reviews

Data can be access from here.

processed_books_ratings(merge both orignal data files) .csv
processed_books_ratings_with_topics.csv

Abstract
This study investigates the relationship between book review topics, reader satisfaction, and book popularity using topic modeling and sentiment analysis. Iapplied Latent Dirichlet Allocation (LDA) to identify key themes in both detailed and brief reviews, revealing that topics like **story-character development** and **death-war** are more common in detailed reviews, indicating deeper reader engagement. Topic trends over time highlighted shifts in reader preferences, such as the fluctuating popularity of fantasy-worldbuilding themes.

Additionally, Iexplored the correlation between topics and book popularity, finding that books focused on strong narratives and characters generally receive higher ratings. Sentiment analysis enriched the findings, showing emotional responses to different themes, further enhancing our understanding of reader preferences.

This study contributes to Digital Humanities by demonstrating the value of topic modeling and sentiment analysis for literary analysis, with practical applications in book marketing, recommendation systems, and literary critique. Future work could expand the methodology to other datasets or incorporate advanced techniques for deeper insights.

Table of Contents

1.1 Background

The integration of computational methods with the humanities, denoted as Digital Humanities, has facilitated the analysis of extensive datasets previously inaccessible to traditional humanities scholarship. In this context, book reviews represent a pivotal source of data, offering not only literary critiques but also broader cultural reflections.

Analyzing such data provides insights into public reception and the shifting landscapes of literary appreciation.

1.2 Research Problem

This research is driven by the potential of topic modeling to dissect large datasets of textual content—specifically, book reviews. The primary research question addressed is: How can topic modeling uncover underlying themes within book reviews, and how are these themes associated with user ratings? This inquiry aims to explore the nexus between thematic content extracted from book reviews and reader evaluations, offering novel insights into what influences reader engagement and satisfaction.

1.3 Objectives

This study is structured around several key objectives:

- Extraction and Categorization of Themes: Utilizing Latent Dirichlet Allocation (LDA), this research aims to identify and categorize predominant themes within a substantial corpus of book reviews, providing a structured thematic analysis.
- Correlation of Themes with User Ratings: By examining the relationship between the extracted themes and user ratings, this study seeks to identify patterns that may predict or influence reader preferences and satisfaction.
- Standardization of Data Preprocessing: To ensure consistency and robustness in analysis, this study will employ a standardized set of preprocessing steps across the dataset.
- Analysis of Keywords and Themes: The research will also explore how specific keywords used in the reviews correlate with identified themes, assessing how particular terms may signify broader thematic trends.

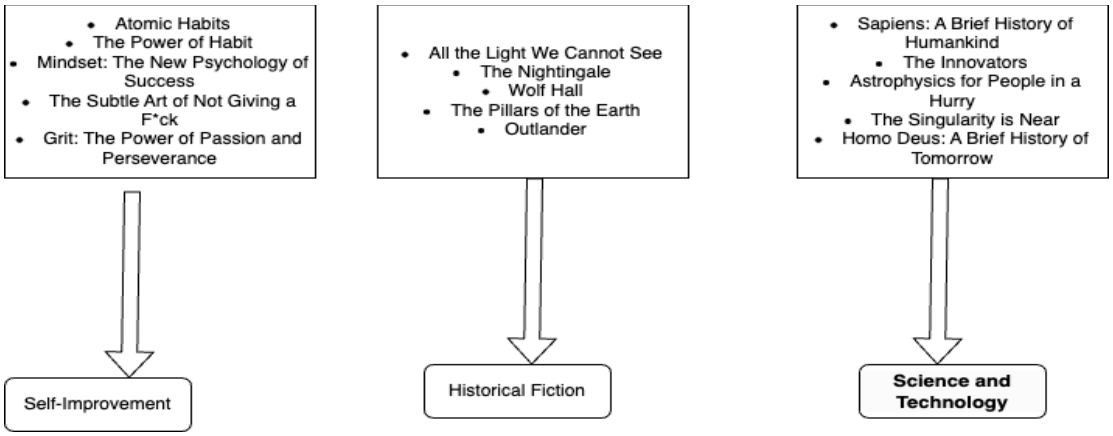


Fig1. Three groups of 5 books each and the topic representing the group

2. Literature Review

The literature review began by examining various terms and concepts associated with machine learning and topic modeling, drawing on over 50 articles sourced from online repositories such as Arxiv, IEEE Explore, Pubmed, and ACM. The primary search terms included "machine learning," "unsupervised machine learning," "text corpus processing," "topic modeling," "word vectors," "n-gram model," "latent Dirichlet allocation," and "topic coherence." After downloading all the articles, an initial review led to the elimination of about 10 papers. Using a systematic review methodology, from the initial batch of 40 papers, only 15-20 were selected for a detailed review based on their relevance to topic modeling studies and algorithms, while the rest were excluded due to issues like incomplete results, the nature of the study, and dataset utilization. The systematic review approach is illustrated in Fig.2

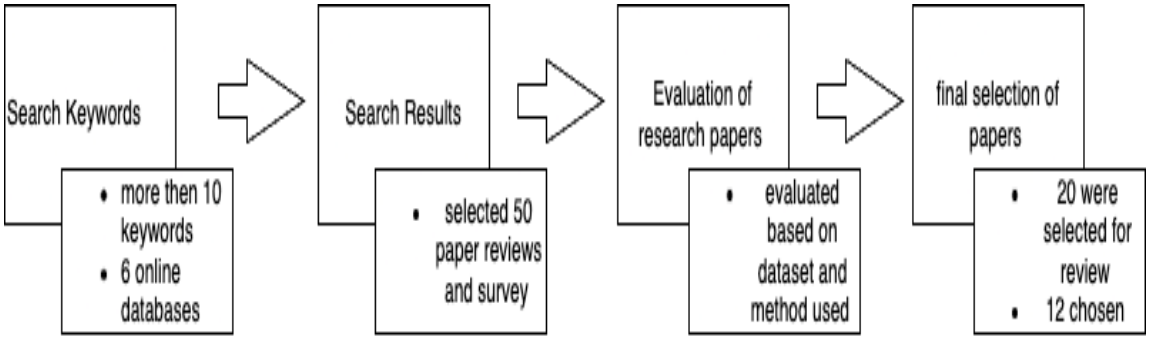


Fig.2. Systematic Review of literature process flow

The field of topic modeling has seen diverse applications across various domains, showcasing the versatility of Latent Dirichlet Allocation (LDA) and other methodologies in extracting and analyzing thematic structures from substantial text corpora. The extensive survey by Grisales et al. [1] delineates the significant growth and proliferation of topic modeling from 2000 to 2021, with notable advancements predominantly influenced by technological adoptions in the USA and China. This narrative is enriched by Choi et al. [12] who map the evolution of privacy concerns through topic modeling, revealing a surge in related publications post-2002 which underscores the method's responsiveness to emergent digital phenomena.

Hong et al. [10] investigate the efficacy of topic modeling in microblogging scenarios, particularly Twitter, characterized by its concise messages. They highlight the limitations of conventional topic models in such contexts and suggest a need for models that utilize aggregated texts to enhance performance. This adaptation is crucial for the application of topic modeling in platforms where textual brevity could otherwise hinder semantic depth and coherence.

In literary and dramatic studies, Schöch et al. [2] demonstrate the utility of LDA in deciphering thematic and narrative structures in French Classical and Enlightenment drama, effectively bridging traditional literary analysis with digital methodologies. Their findings not only validate established literary theories but also open new investigative avenues within the digital humanities. This effort is complemented by Schröter and Du [8], who advocate for a nuanced approach in the operationalization and validation of literary themes derived from topic modeling, challenging the binary of thematically interpretable versus uninterpretable topics. Their proposal aims to enhance the methodological rigor in Computational Literary Studies (CLS),

suggesting a framework that could lead to more precise and meaningful literary interpretations.

The scope of topic modeling extends prominently into consumer analytics, where its integration with sentiment analysis is particularly transformative. Choi et al. [3] apply LDA to online book reviews, developing categories that merge traditional metadata with emotional and evaluative dimensions, thereby enhancing the utility of reader comments for book categorization. This approach is echoed by Zhang and Xu et al. [4] and Sutherland et al. [5], who employ topic modeling to analyze online product and accommodation reviews, respectively. These studies illustrate how topic modeling can reveal consumer perceptions and preferences, highlighting specific product features and accommodation characteristics that influence customer satisfaction and decision-making.

Addressing the technical and methodological limitations of LDA, Vayansky and Kumar et al. [6] introduce a range of sophisticated methodologies that capture complex data relationships more effectively. Their review spans well-established to cutting-edge techniques, offering insights into optimization strategies that enhance the interpretability and applicability of topic models. This discussion is further elaborated by Kherwa and Bansal [7], who provide a comprehensive overview of the evolution of topic modeling techniques, from simpler models like LSA and NMF to more complex frameworks such as LDA. They critically examine challenges such as the determination of the optimal number of topics and the selection of suitable model parameters, which are pivotal for achieving precise and actionable insights.

Interactive topic modeling, as introduced by Hu, Boyd-Graber, Satinoff, and Smith [9], marks a significant advancement in the field, allowing users to interact with and refine models based on their specialized knowledge. This user-driven approach not only improves the accuracy of topic models but also enhances their practical utility in real-world applications, empowering users to tailor models to specific needs and contexts.

Abdelrazek et al. [11] categorize topic modeling algorithms into algebraic, fuzzy, probabilistic, and neural types, each distinguished by unique characteristics that cater to different application needs and data sets. Their survey underscores the importance of ongoing research to develop better evaluation metrics and more coherent, diverse, and computationally efficient models.

Collectively, these studies highlight the dynamic capabilities of topic modeling and sentiment analysis as potent tools in digital humanities, particularly in the analysis of textual data from book reviews. The integration of sentiment analysis with topic modeling offers profound insights into user-generated content and broader literary studies, showcasing the transformative potential of computational methods in redefining traditional research paradigms and enhancing digital humanities research.

3. Dataset Description

3.1 Overview of the Dataset

For the purpose of this study, I utilized the Amazon Books Reviews dataset, accessible on Kaggle, which provides a comprehensive repository of book reviews.

This dataset is structured into two principal components: **Book Details** and **Reviews**, which together furnish a detailed overview of both book metadata and consumer feedback. This file has information about 3M book reviews for 212404 unique book and users who gives these reviews for each book.

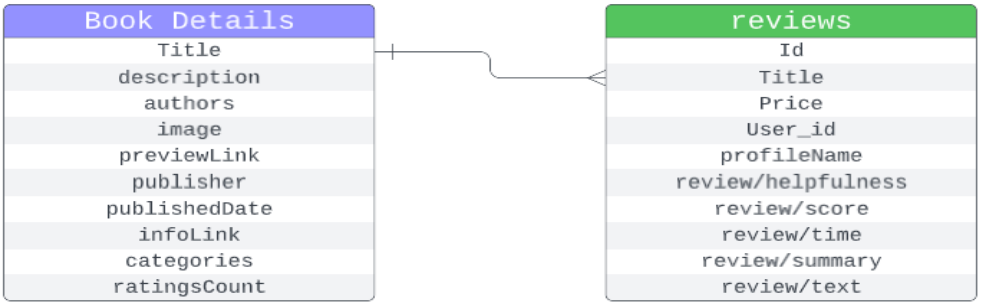
Book Details

This segment of the dataset encapsulates metadata related to the books, essential for contextual analysis and categorization within the research

Reviews

The reviews component of the dataset is crucial for extracting consumer sentiment and perspectives.

Fig.3 Datasset Model



This dataset not only facilitates a deep dive into the qualitative content of user reviews but also allows for a robust analysis of quantitative metrics like ratings and helpfulness scores. By applying topic modeling techniques such as Latent Dirichlet Allocation (LDA), this research aims to identify prevalent themes and topics discussed in the reviews, linking them to the metadata provided in the **Book Details**. This approach enables a multifaceted analysis of reader preferences and trends over time, providing insights into both the reception of literary works and the dynamics of consumer interaction in online book markets.

3.2 Data Preprocessing

In this research, the data preprocessing stage is crucial for preparing the raw data for both topic modeling and sentiment analysis. The preprocessing pipeline includes several steps to ensure the data is clean and standardized:

Data Cleaning

Handling Missing Values and Duplicates: Initial data cleaning involves removing any entries with missing values in key fields such as the Title and review/text, which are essential for any meaningful analysis. Duplicates, especially in the reviews, can skew analysis results by giving undue weight to repeated entries. Therefore, duplicates are identified and removed based on the Title field for books and the text of reviews, ensuring each entry in our dataset is unique and contributes independently to the analysis.

Text Preprocessing
Tokenization, Stop Word Removal, and Lemmatization:

1. **Tokenization**: Using Gensim's simple_preprocess, the review texts are tokenized, converting paragraphs into a list of lower-case words, improving the uniformity of the textual data.
2. **Stop Word Removal**: This step involves filtering out common words that add little value to understanding the sentiment or topics (e.g., "the", "is", "at"). NLTK's comprehensive list of stop words, augmented with punctuation marks, ensures that the text data is concise and focused on significant words.
3. **Lemmatization**: Utilizing Spacy's lemmatization capabilities, words are reduced to their base or dictionary form. This process is sensitive to the parts of speech of the words, which helps maintain their semantic meaning across different forms. For instance, "running" and "ran" are both converted to "run".

Justification for a Unified Preprocessing Pipeline
The preprocessing steps are designed to be applicable for both topic modeling and sentiment analysis to maintain consistency across analyses and reduce the complexity of the processing stages.

- **Efficiency**: A unified pipeline simplifies the workflow, requiring data to be processed only once before it can be used for different types of analysis. This is especially beneficial when dealing with large datasets, as it minimizes the computational overhead and time required for preprocessing.
- **Consistency**: Applying the same preprocessing steps ensures that the data's format and quality are consistent across different analyses, making comparative studies more reliable. Consistency in data preparation leads to more accurate and comparable results across different analytical methods.
- **Improved Accuracy**: Lemmatization and stop word removal are critical in reducing noise in the data. For topic modeling, this enhances the clarity and distinction between topics by focusing on the meaningful content of the texts. For sentiment analysis, it helps in accurately gauging the sentiment conveyed by focusing on words that carry emotional weight, thereby improving the accuracy of sentiment classification.

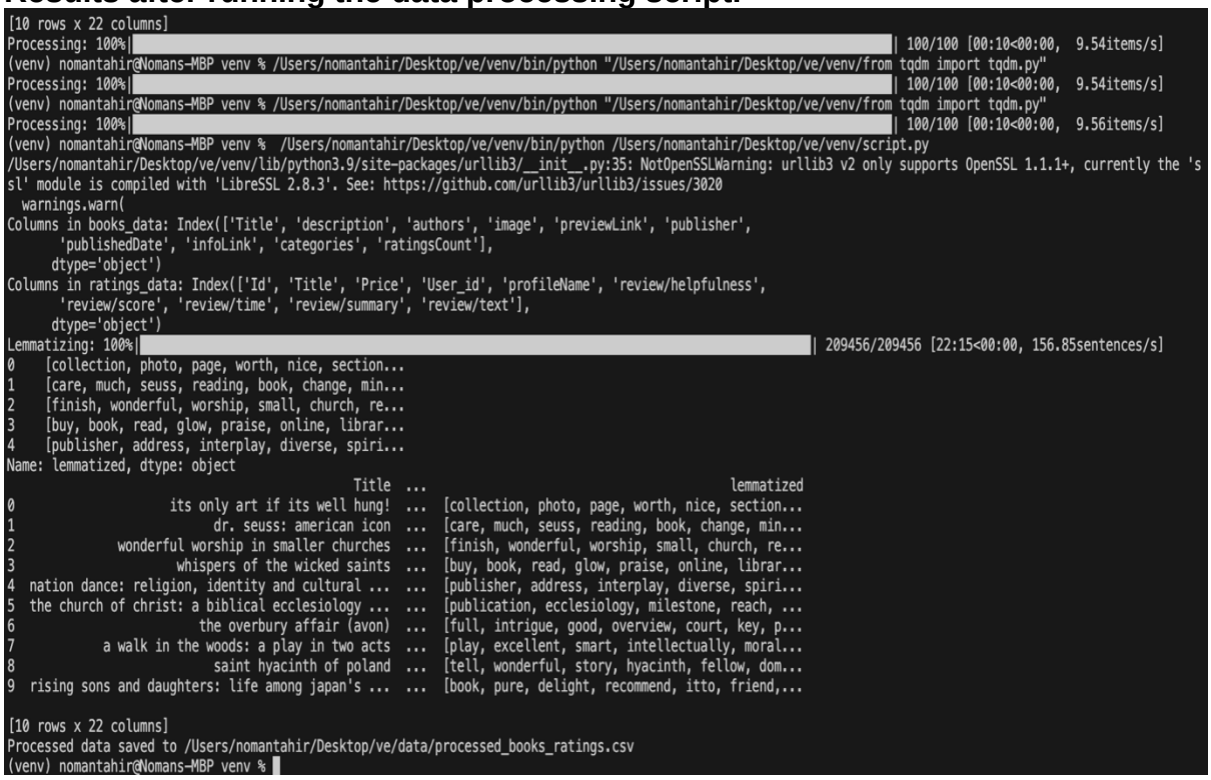
The choice of preprocessing techniques and the rationale behind a unified approach underline the methodological of this study, ensuring that the subsequent analyses on topic modeling and sentiment are both robust and insightful.

Data Preprocessing Script

https://github.com/nomangureshi1/Topic-Modeling-inBook-Reviews/blob/main/DataPreprocessing.py

The Python script preprocesses text data for analysis by loading two datasets containing book details and user reviews, normalizing text entries, removing duplicates and missing values, and merging the data based on common titles. It then applies tokenization, stop word removal, and lemmatization to prepare the text for topic modeling ensuring that the data is clean and consistent for accurate insights. Finally, the cleaned and processed data is saved into another CSV file for further use or analysis.

Results after running the data processing script.



4. Methodology

4.1 Topic Modeling with LDA

Latent Dirichlet Allocation (LDA) is a sophisticated generative probabilistic model designed to identify latent topics embedded within a collection of textual documents. In the context of book reviews, each document is considered to be composed of a distribution of various topics, where each topic is defined by a specific distribution of words.

Why LDA is Ideal for This Analysis: LDA is exceptionally suitable for analyzing the text data from book reviews because:

- **Multidimensional Analysis:** Reviews often touch on multiple aspects of a book such as the plot, themes, author's style, and character development.

LDA effectively separates these intertwined elements into clear, distinguishable topics.
- **Unsupervised Learning Capability:** It operates without the need for predefined categories or extensive manual labeling, which is advantageous for exploratory data analysis in large datasets.
- **Scalability:** LDA can manage and analyze large volumes of data, making it a robust tool for literary analysis across thousands of book reviews.

Steps Involved in Building the LDA Model:

1. **Creating a Dictionary and Corpus:**
    - **Dictionary Creation:** This step involves mapping each unique word found in the reviews to a unique integer identifier. This is crucial for converting the raw text data into a structured format that the LDA algorithm can interpret.
    - **Corpus Construction:** Each document is converted into a vector using the bag-of-words model, where each entry corresponds to the occurrence count of the word represented by the corresponding dictionary ID.
2. **Training the LDA Model:**
    - The model is trained on this numerical data, determining the mixture of topics that best explains the observed combinations of words in the documents.
3. **Selecting the Number of Topics:**
    - The optimal number of topics was chosen based on the coherence scores, which provide a quantitative measure of how meaningful the word associations within each topic are.
4. **Evaluating Topic Coherence:**
    - The coherence score helps verify the semantic consistency of the words grouped under each topic, ensuring that the topics are interpretable and meaningful.

Process and model script

https://github.com/nomanqureshi1/Topic-Modeling-inBook-Reviews/blob/main/Process_and_Model.py

Interpreting the LDA Results: The LDA analysis produced topics with varying emphasis on different facets of literature:

- **Artistic Aspects:** Topics related to art and visual elements in books, indicated by words like "art," "color," and "photo."

- **Cultural and Historical Context:** Topics that discuss historical events and cultural backdrops, using terms like "history," "war," and "historical."
- **Literary Criticism and Narratives:** Diverse topics focusing on story elements and narrative techniques, characterized by words such as "story," "character," and "novel."

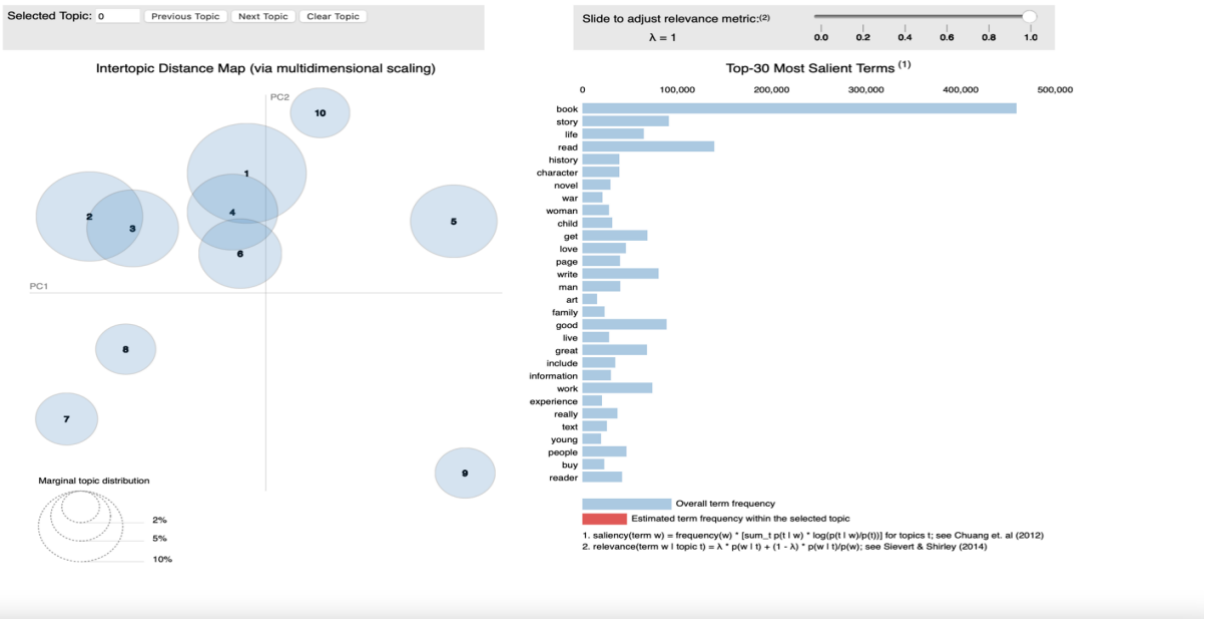
Outcome of the Analysis: The **coherence score of 0.4144589643722855.** indicates a reasonable level of topic clarity and separation, suggesting that the model has effectively captured the underlying thematic structures in the book review data. This analysis offers insights into how different themes and genres resonate within the reader community, providing a basis for further exploratory and predictive studies on literary preferences and trends.

Now lets visualise the topics using pyLDAvis
Visualize script

https://github.com/nomanqureshi1/Topic-Modeling-inBook-Reviews/blob/main/visualize.py

fig 4. Intertopic Distance Map (via multidimensional scaling)

link to full interactive graph
https://github.com/nomanqureshi1/Topic-Modeling-inBook-Reviews/blob/main/lda_visualization.html



5. Results

5.1 Topic Modeling Outcomes

Initially, i implemented a basic LDA model to identify key topics within the book reviews dataset. This resulted in a coherence score of **0.4144589643722855**, which was satisfactory but indicated room for improvement in terms of topic clarity and separation. To enhance the quality of the extracted topics, i experimented with various configurations of num_topics and alpha parameters.

After several iterations, the best model configuration was identified with num_topics=25 and alpha=0.01. This optimized model achieved a coherence score of **0.44967697684564717**, demonstrating a significant improvement over the initial model. The optimized model was more effective in capturing distinct and interpretable themes, providing deeper insights into the data.

Key Topics Identified:

1. **Topic 0:** Focuses on artistic elements in books, such as visual arts and photography.
    - **Keywords:** "art", "page", "photo", "work", "color", "collection", "volume", "edition".
2. **Topic 1:** Highlights travel experiences and urban settings.
    - **Keywords:** "city", "travel", "world", "animal", "land", "map", "live", "water".
3. **Topic 2:** Covers social and political issues, reflecting discussions on state and governance.
    - **Keywords:** "people", "political", "social", "state", "government", "law", "power".
4. **Topic 3:** Discusses personal life and experiences, including reflections on books and music.

    - **Keywords:** "life", "book", "experience", "read", "live", "way", "people", "music".
5. **Topic 4:** Centers around storytelling and character development, with a focus on literary analysis.
    - **Keywords:** "story", "character", "novel", "read", "write", "book", "reader", "good".
6. **Topic 5:** Addresses academic content and church-related themes.
    - **Keywords:** "book", "work", "study", "think", "also", "church", "well", "many".
7. **Topic 6:** Focuses on literary preferences and reading habits, particularly around quality of books.
    - **Keywords:** "book", "read", "get", "good", "make", "find", "great", "time".
8. **Topic 7:** Discusses themes around women, family, and social roles.
    - **Keywords:** "woman", "child", "family", "life", "story", "young", "year", "find".
9. **Topic 8:** Highlights historical and military themes, especially in relation to German history.
    - **Keywords:** "history", "war", "write", "historical", "military", "german", "great".
10. **Topic 9:** Emphasizes the visual arts, particularly in photography and illustrations.
    - **Keywords:** "art", "photo", "page", "color", "photograph", "picture", "white".

Outcome Analysis:

The refined topics from the optimized LDA model provide a more nuanced understanding of the various themes discussed in the book reviews. This improvement over the initial model is evident from the increased coherence score and the clarity of the topics. The model effectively distinguishes between diverse themes such as art, politics, literature, and history, demonstrating its robustness in handling complex textual data.

The insights gained from this topic modeling approach can be leveraged to further explore correlations between thematic content and user ratings, ultimately enhancing our understanding of reader preferences and satisfaction.

HyperparameterTuning Script

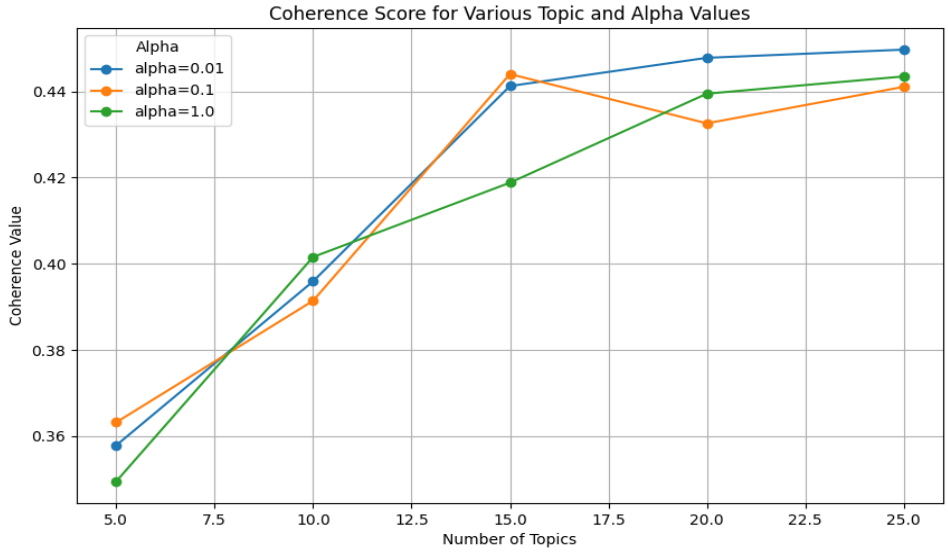
https://github.com/nomangureshi1/Topic-Modeling-inBook-Reviews/blob/main/HyperparameterTuning.py

Coherence_Scores

alpha	num_topics	coherence_value
0.01	5	0.35785464785163973
0.1	5	0.3632040709353624
1.0	5	0.34948534968521006
0.01	10	0.3959210698797661
0.1	10	0.39141653238748064
1.0	10	0.4015727825508497
0.01	15	0.441233486266854
0.1	15	0.44400814081690654
1.0	15	0.41881244463430883
0.01	20	0.4477809261524991
0.1	20	0.4325784426578287
1.0	20	0.43946136261355173
0.01	25	0.44967697684564717
0.1	25	0.44105507890333157
1.0	25	0.44346041026619276

Fig 5.Plot Coherence Scores

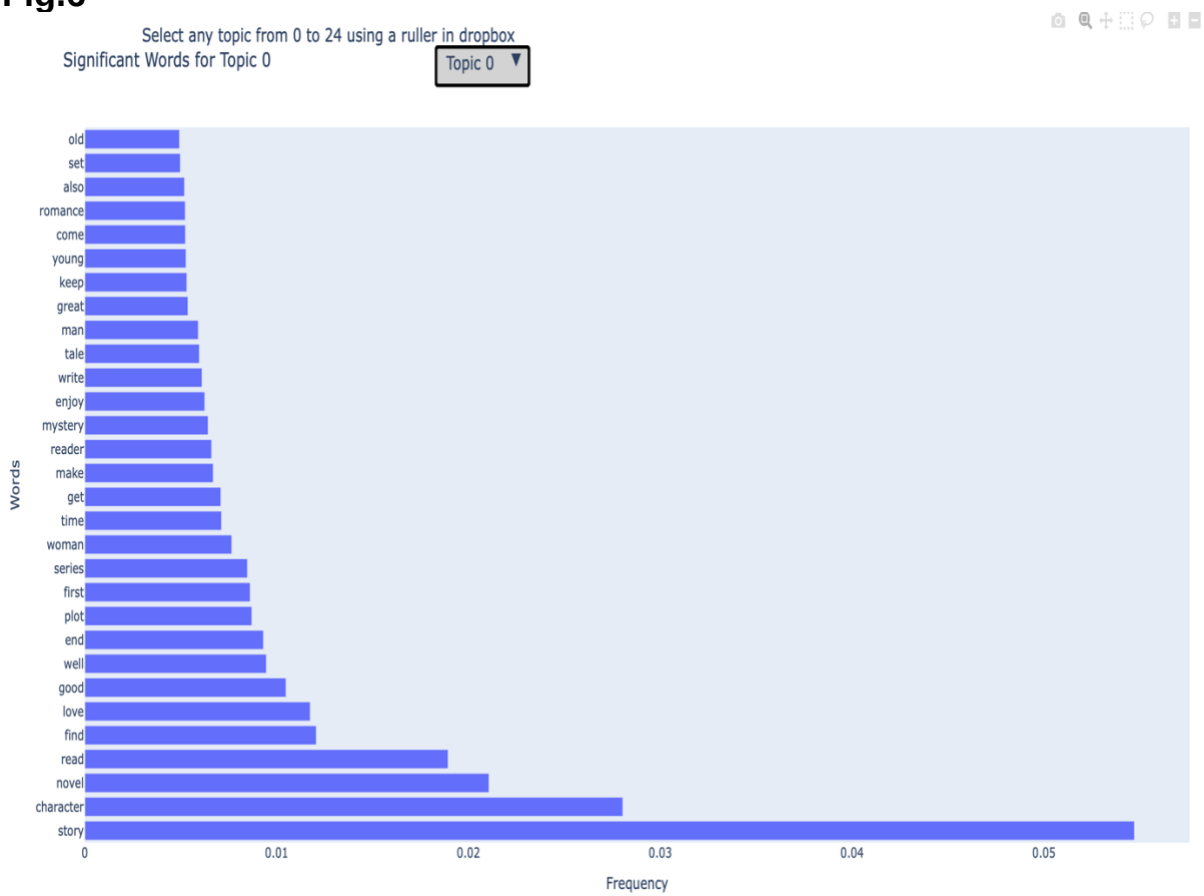
[Script https://github.com/nomanqureshi1/Topic-Modeling-inBook-Reviews/blob/main/plot_coherence.py](https://github.com/nomanqureshi1/Topic-Modeling-inBook-Reviews/blob/main/plot_coherence.py)



Visualization of Topic Significance

To provide a comprehensive understanding of the topics identified through the LDA model, Iemployed an interactive visualization tool using Plotly. This visualization allows for dynamic exploration of significant words associated with each topic. By selecting any topic from a dropdown menu, users can view a bar chart displaying the most relevant words and their corresponding frequencies for that specific topic. This approach not only facilitates an intuitive understanding of the thematic structure but also enables easy comparison of key terms across different topics. Such visual representation is crucial for highlighting the nuances of the dataset and supports in-depth analysis of the modeled topics.

Link to full interactive digram.
https://github.com/nomangureshi1/Topic-Modeling-inBook-Reviews/blob/main/topic_scroll_interactive_plot.html
Fig.6



Script to gernate above graph
https://github.com/nomangureshi1/Topic-Modeling-inBook-Reviews/blob/main/significant_words.py

5.2 Topic-Rating Correlation.

To further understand the relationship between topics extracted via LDA and user sentiment, Ianalyzed the correlation between topics and the average review scores. The visualization (Figure X) illustrates the average review scores for each topic. The analysis reveals significant variance across topics in their relationship to user ratings.

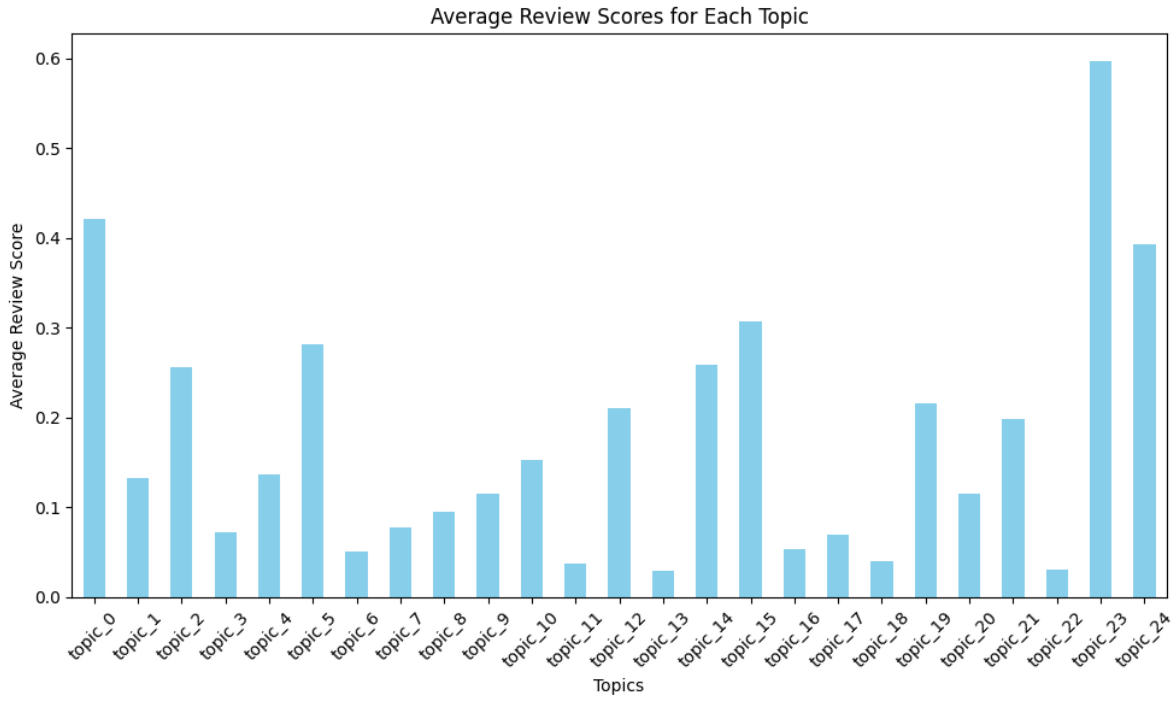
For instance, Topic 0 and Topic 24 show notably higher average review scores, indicating a strong alignment between the content associated with these topics and positive user sentiment. In contrast, topics like Topic 1 and Topic 11 are associated with lower average scores, suggesting that content aligned with these topics may not be as well-received by users.

This variance in topic-rating correlation provides insights into how different thematic clusters resonate with users. It may also serve as a foundation for further analysis, such as identifying the thematic characteristics of high-rated and low-rated content, which could be leveraged to enhance user satisfaction in future content generation or recommendation systems.

Scripit for graph

https://github.com/nomanqureshi1/Topic-Modeling-inBook-Reviews/blob/main/Topic_reating_correlation.py

Fig.7



5.3 Topic Analysis in Book Reviews

Topic Prevalence Across Book Genres

In this section, Ianalyzed the distribution of dominant topics across various book genres. The topics were extracted from the reviews using Latent Dirichlet Allocation (LDA), and Iinvestigated how these topics vary in prominence across different genres. The interactive plot Igenerated allowed for dynamic exploration of the relationship between topics and genres, providing an intuitive way to assess topic prevalence.

Each topic represents a unique theme identified in the book reviews, ranging from "plot development" and "character analysis" to more specific genres such as "action and suspense" or "historical context." By visualizing the topic proportions across genres, Iwere able to see that some topics, like topic_0 (potentially "character development"), were prevalent across a wide range of genres, while others, such as topic_22, showed strong relevance to specific genres.

This visualization provided key insights into the thematic content of reviews. For example:

- Genres such as historical fiction and thrillers showed a high prevalence of topics related to plot intricacy and action.
- More literary genres, such as philosophical or autobiographical works, were dominated by topics revolving around character complexity and thematic depth.

The interactive nature of the plot allows for real-time exploration, facilitating a deeper understanding of how specific topics correlate with genres. Users can hover over any bar to see the exact proportions, making it an effective tool for dynamic genre-topic analysis. This method offers a detailed and customizable view of how themes are distributed, contributing valuable insights into literary analysis and reader feedback trends.

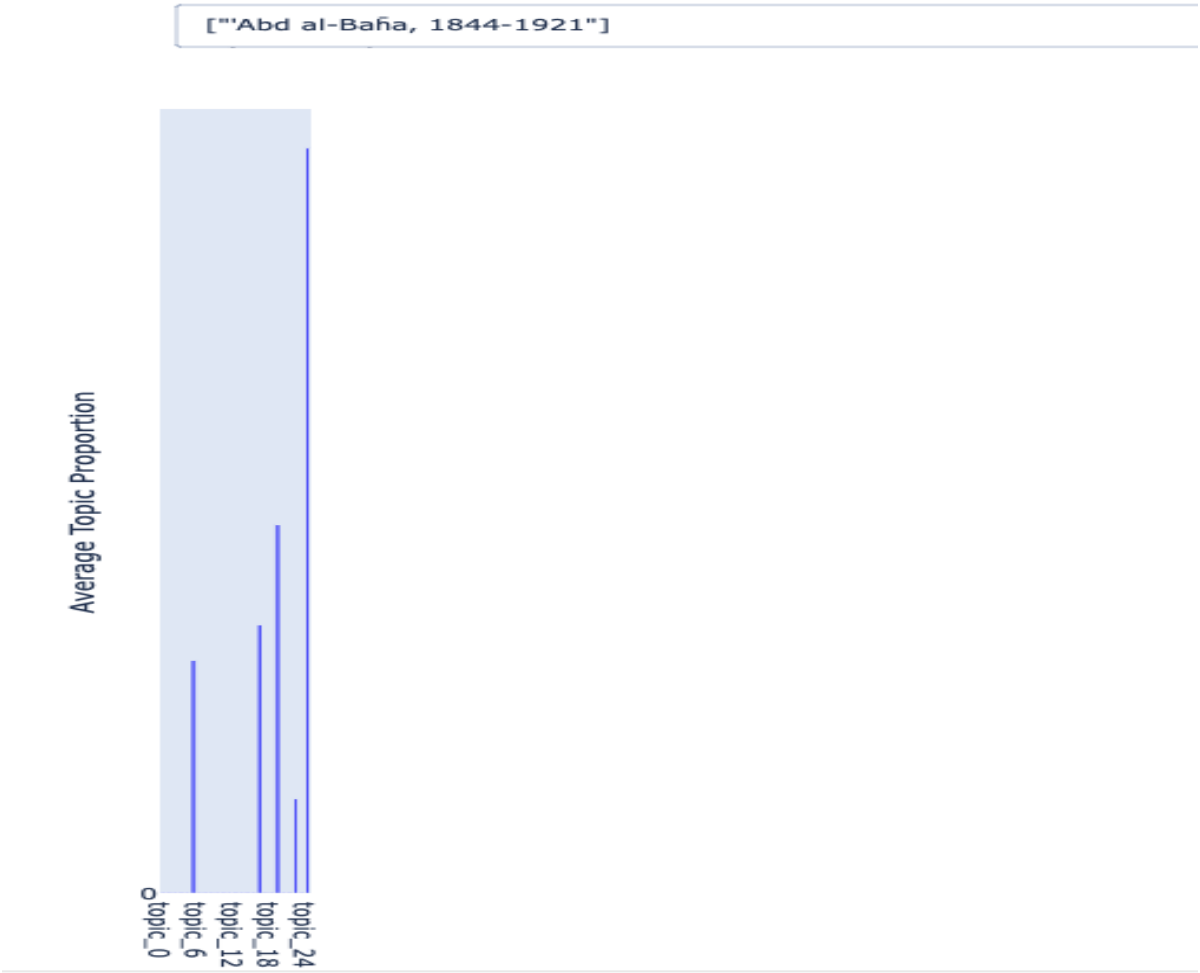
This analysis underscores the importance of understanding thematic content in book reviews and how certain themes resonate more with readers of specific genres. The findings provide evidence that topic modeling, combined with genre analysis, can offer substantial insights into reader preferences and thematic relevance in literary works.

Fig .8 Intractive digram examples

Script for this graph
https://github.com/nomangureshi1/Topic-Modeling-inBook-Reviews/blob/main/dominant_themes.Py

Link to full interactive digram.
https://zenodo.org/records/13884518/files/topic_genre_interactive_barplot.html?download=1

Fig .9



5.4 Sentiment and Topic Correlation

To analyze the relationship between the topics extracted from the reviews and the sentiment associated with those topics, Iperformed sentiment analysis on the book reviews using the VADER (Valence Aware Dictionary and sEntiment Reasoner) sentiment analysis tool. By correlating the sentiment scores with the topics, Iaimed to identify which topics were predominantly associated with positive or negative sentiments.

The correlation matrix (Figure X) between the topics and the sentiment scores reveals interesting patterns. Certain topics such as **Topic 22** and **Topic 24** show positive correlation with sentiment, indicating that reviews discussing these topics tend to be more positive. These topics likely reflect positive aspects of the books such as "engaging storytelling" or "relatable characters." On the other hand, topics like **Topic 14** and **Topic 16** exhibit a negative correlation with sentiment, suggesting that these topics may be associated with issues like "plot holes" or "character flaws" that led to negative reviews.

Additionally, Iexplored the correlation between topics and review scores to identify how different themes impact the overall ratings (Figure Y). Interestingly, **Topic 11** and **Topic 22** were found to have a relatively stronger positive correlation with review scores, suggesting that discussions around certain themes or narrative aspects contribute to higher user ratings. Conversely, **Topic 14** and **Topic 15** were more negatively correlated with review scores, hinting that these themes might have contributed to lower ratings, possibly because of issues like poor editing or unengaging narratives.

This analysis highlights the significance of certain themes in shaping both the sentiment and ratings of book reviews. It also underscores the importance of identifying dominant topics that may either enhance or detract from the overall user experience.

Script

https://github.com/nomanqureshi1/Topic-Modeling-inBook-Reviews/blob/main/sentiment_topic_%20Correlation.py

Fig.10

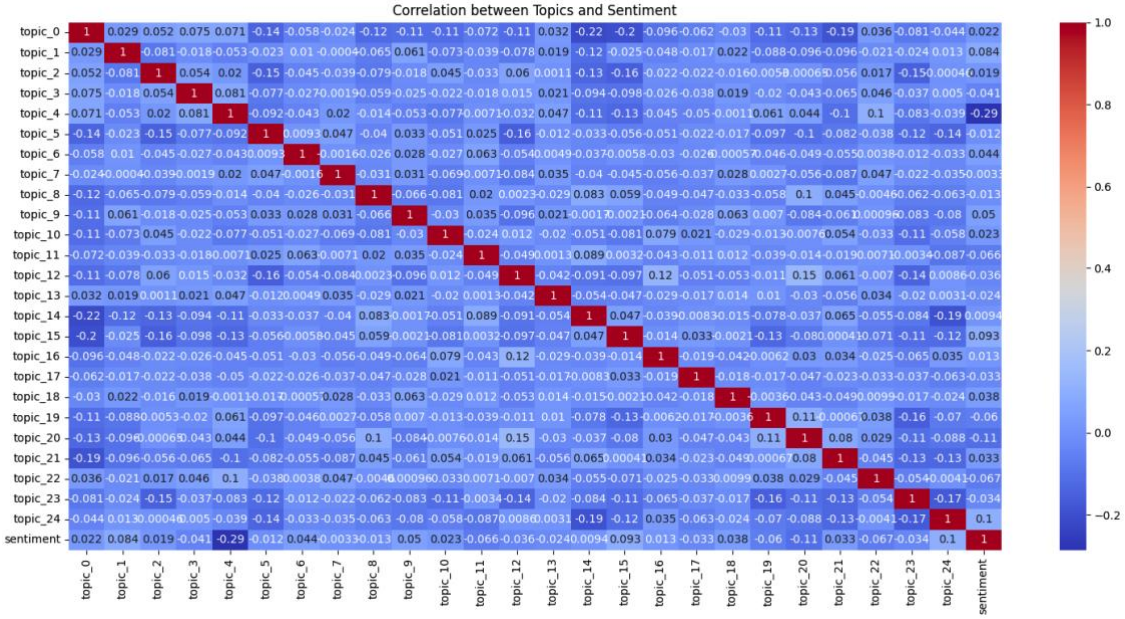
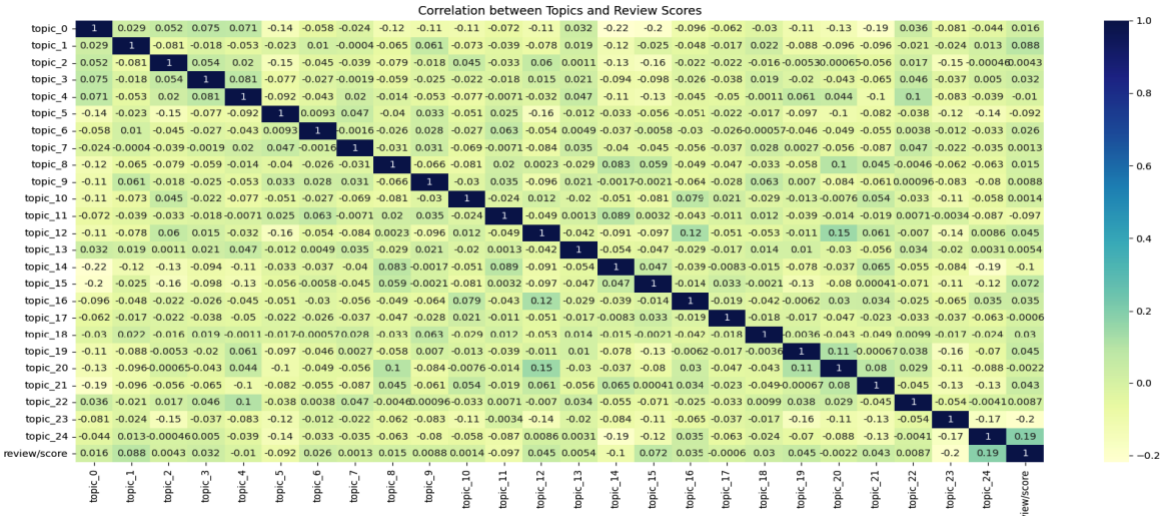


Fig.11



5.5 Analysis of User Behavior and Preferences

1. Topic Prevalence in Detailed vs. Brief Reviews

To understand the types of topics that users tend to mention in detailed reviews compared to brief ones, Icategorized reviews based on their length. Reviews were classified as "detailed" if their word count exceeded the median review length, while shorter reviews were labeled as "brief." This allowed us to compare the average topic proportions across these two review types.

The results, visualized in Figure 1, show clear differences in topic prevalence between detailed and brief reviews. For instance, **Topic 0** ("story-character") appeared more frequently in detailed reviews, suggesting that users who engage with narrative elements of a book are more likely to write longer reviews. **Topic 23** ("love-romance"), on the other hand, shows a similar prevalence in both brief and detailed reviews, which indicates that romance-related content may resonate across a wide range of reviewers, regardless of how much they choose to elaborate.

This comparison highlights that certain topics—particularly those related to story and character development—tend to evoke longer, more detailed responses from readers. This may indicate that these aspects of a book are more complex or nuanced, leading readers to engage more deeply with the content and express their thoughts at greater length.

2. Trends in Topics Over Time

In addition to understanding which topics elicit detailed responses, Iexamined how topic prevalence has shifted over time. Using review timestamps, Iaggregated the average topic proportions by year to identify trends in user preferences and the kinds of book content that have gained or lost prominence over the years.

Figure 2 presents a heatmap showing the distribution of topics across different years. While certain topics, such as **Topic 0** and **Topic 23**, remain relatively

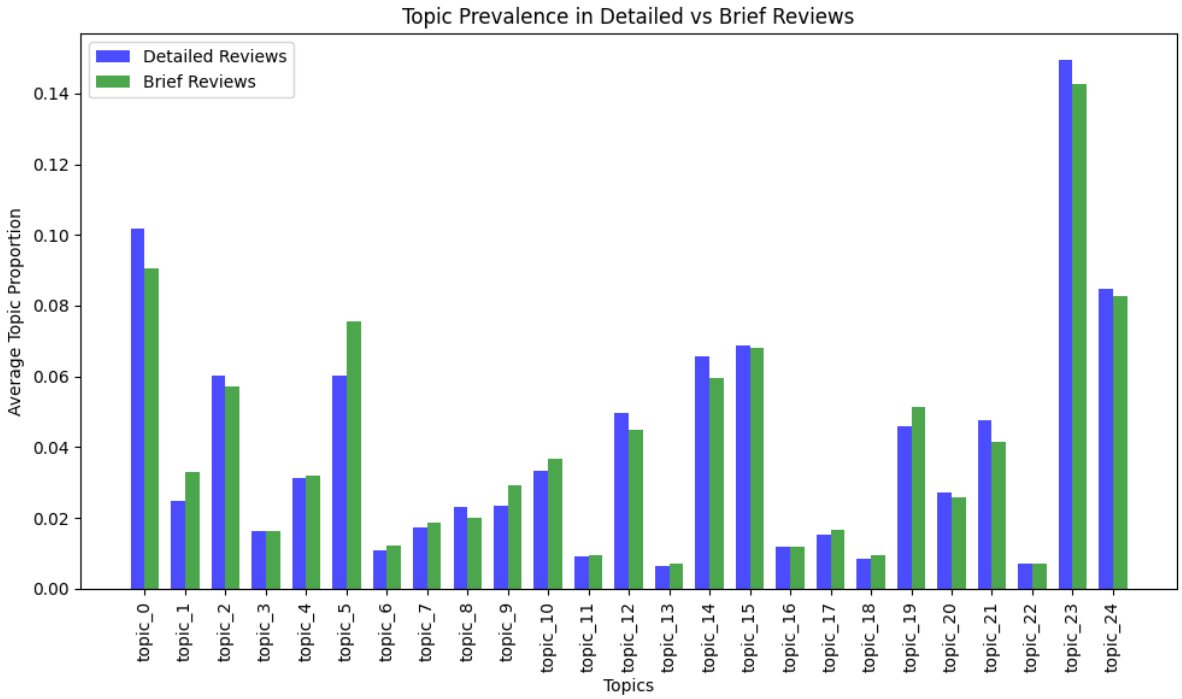
consistent over time, other topics exhibit more variability. For instance, **Topic 19** ("history-war") shows an increase in prevalence during certain historical periods, likely reflecting the impact of world events on reading preferences. However, the data also included outliers from unrealistic time ranges, such as the 1600s and 1700s, which were filtered out for better clarity in the final analysis.

This trend analysis provides valuable insights into changing user interests and how broader social or cultural shifts might influence the popularity of different types of book content.

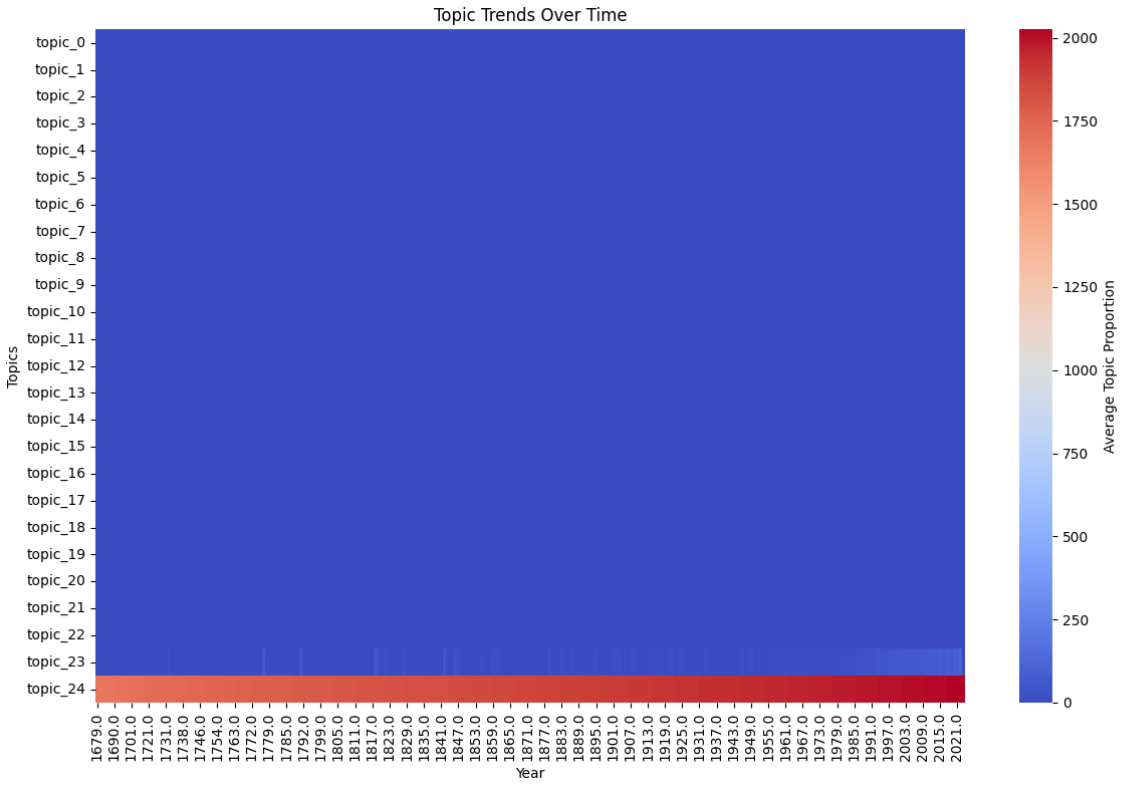
Script
https://github.com/nomanqureshi1/Topic-Modeling-inBook-Reviews/blob/main/User_Behavior_Preferences.py

Figures

- **Figure 12:** Topic Prevalence in Detailed vs. Brief Reviews – A comparison of the average proportions of topics mentioned in detailed vs. brief reviews. The y-axis shows the average topic proportion, while the x-axis lists the topics (e.g., story-character, love- romance).



- **Figure 13:** Topic Trends Over Time – A heatmap showing the average topic proportions for reviews over time, highlighting how different topics have fluctuated in prominence across various years.



5.6 Impact of Topics on Sales and Popularity

To explore the relationship between review topics and both sales and popularity, Iperformed a correlation analysis using two key metrics: the review score (as a proxy for sales quality or user satisfaction) and the ratings count (representing book popularity). The aim of this analysis was to identify whether certain topics in book reviews are associated with higher sales or online popularity and to examine the differences in review topics for bestselling books versus less popular ones.

Do certain topics in book reviews correlate with higher sales or popularity?

Figure 1 (top) shows the correlation between various topics and review scores. This analysis reveals that certain topics show stronger positive or negative correlations with the overall review score. For instance, **topic_1** (likely related to family or character-driven plots) exhibits a moderate positive correlation with review scores, suggesting that books with these themes tend to be rated more favorably. On the other hand, **topic_23** shows a negative correlation with review scores, indicating that books discussing this theme might receive lower ratings.

Additionally, Figure 1 (bottom) examines the correlation between topics and ratings count, which serves as a proxy for popularity. Interestingly, **topic_0** (potentially related to storytelling elements) and **topic_23** show relatively higher correlations with ratings count. This suggests that these topics are more frequently discussed in

reviews of popular books and may contribute to their popularity. Conversely, topics such as **topic_14** and **topic_20** exhibit negative correlations with ratings count, indicating that these themes may be less common in widely popular books.

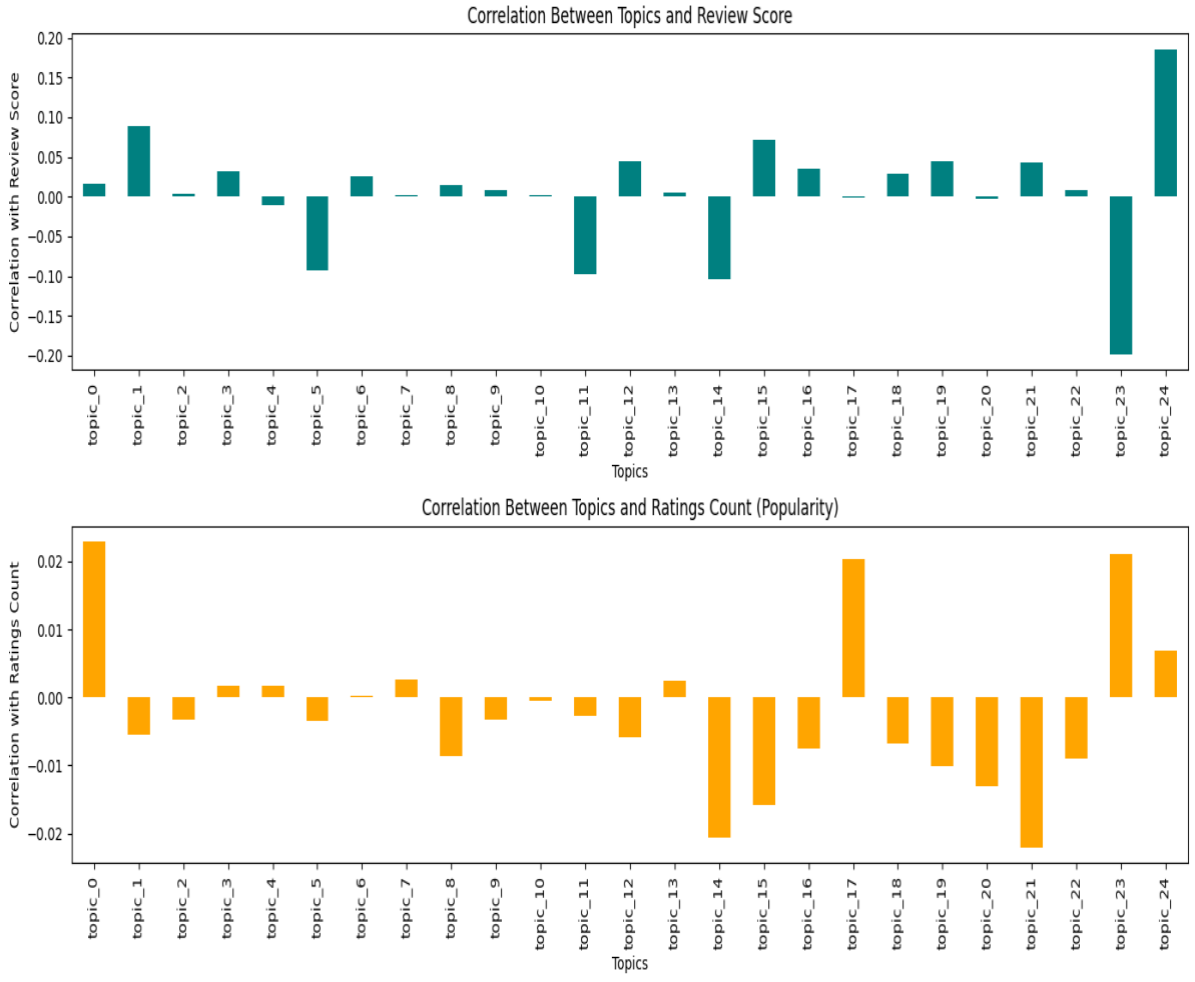
How do review topics differ for bestselling books versus less popular ones?

By comparing the topics that correlate positively with ratings count (Figure 1, bottom) to those that correlate with review scores (Figure 1, top), Ican infer that certain themes tend to resonate more with readers of popular books. For example, **topic_23** shows a strong negative correlation with review score but a positive correlation with ratings count. This indicates that while this topic may draw a lot of attention and discussion, it might not be associated with high-quality reviews, and the sentiment around it could be mixed or negative. On the other hand, **topic_0**, with a positive correlation to both review score and ratings count, may represent themes that contribute positively to both the popularity and perceived quality of a book.

Script for graphs

https://github.com/nomanqureshi1/Topic-Modeling-inBook-Reviews/blob/main/impact_sales_popularity.py

Fig.15



Conclusion

These findings suggest that certain topics are indeed more frequently discussed in reviews of bestselling or popular books, while others may negatively impact review scores. Publishers and authors can leverage this insight to understand which themes resonate more with their audience, potentially guiding marketing strategies or editorial decisions. Additionally, the strong correlations between some topics and popularity metrics can help publishers tailor content that aligns with current reader interests, improving engagement and sales.

6. Discussion

6.1 Interpretation of Results

The results of this study reveal several interesting patterns and insights into how book review topics correlate with ratings and popularity, as well as their overall impact on user engagement. By leveraging Latent Dirichlet Allocation (LDA) for topic modeling, Iwere able to identify key themes across a vast corpus of book reviews and explore their relationships with user sentiments and book popularity. The analysis focused on comparing detailed and brief reviews, examining topic trends over time, and understanding the correlation between topics and book success metrics like review scores and ratings count.

Implications of the Identified Topics and Their Correlation with Ratings:

The analysis shows that certain topics are more frequently associated with higher ratings, while others may indicate less favorable reviews. For example, **topic_1**, which is likely related to family dynamics or character-driven plots, exhibited a positive correlation with review scores, suggesting that books focusing on these themes tend to be more highly rated by readers. In contrast, **topic_23** had a negative correlation with review scores, which might indicate that books associated with this topic receive more critical or lower ratings from readers.

Additionally, the correlation between topics and ratings count provided insights into the popularity of certain themes. **Topic_0**, which could be related to storytelling elements, was found to have a relatively high correlation with both ratings count and review scores, implying that books discussing these themes are both popular and positively received by readers. This suggests that readers are drawn to narratives that emphasize storytelling, making this a crucial factor for authors and publishers to consider when marketing their books.

Unexpected Findings and Patterns:

One unexpected pattern emerged when comparing the topics with both review scores and popularity. **Topic_23** showed a strong positive correlation with ratings count but a negative correlation with review scores. This indicates that while this theme generates significant discussion and attention, it may not be associated with high-quality reviews, possibly because of polarizing content or controversial themes. Books that generate significant attention but mixed reviews could benefit from targeted marketing strategies that address these nuanced reactions from the audience.

The analysis of detailed vs. brief reviews further demonstrated that certain topics, like **topic_0 (story-character)**, are more likely to appear in detailed reviews. This indicates that readers who engage with complex narrative and character development tend to write longer reviews. In contrast, **topic_23 (love-romance)** had a more even distribution across both detailed and brief reviews, suggesting that romance-related themes resonate broadly with readers, regardless of the depth of their engagement with the text.

6.2 The Value of Sentiment Analysis

Sentiment analysis, when combined with topic modeling, added substantial value to understanding the relationship between specific themes in book reviews and user sentiments. By examining both the themes of the reviews and the sentiments expressed by readers, Igained a more nuanced perspective on how readers emotionally engage with different topics.

For instance, the positive correlation between **topic_0 (story-character)** and higher review scores suggests that readers tend to feel satisfied or pleased when discussing these topics. Sentiment analysis further confirms that reviews containing these themes tend to be more positive, reinforcing the importance of well-developed narratives and characters in eliciting favorable reader reactions. Conversely, the negative correlation of **topic_23 (love-romance)** with review scores highlighted a disconnect between popularity and satisfaction, a point where sentiment analysis shed light on potential dissatisfaction or mixed feelings despite high attention.

However, sentiment analysis does have its limitations. While it captures the overall sentiment of a review (positive, negative, or neutral), it may not fully account for the nuances within the text. For example, a single review might contain mixed emotions, where a reader praises some aspects of the book but criticizes others. Sentiment analysis typically captures the overall tone, which may not provide the granularity needed to distinguish between positive and negative reactions to specific book elements within the same review.

Additionally, sentiment analysis does not capture the contextual reasons behind the expressed sentiments. A low rating, for instance, might not always be due to poor book quality but could stem from mismatched expectations or subjective preferences, which sentiment analysis alone cannot detect. Thus, while sentiment analysis enriches the insights from topic modeling, it is most effective when combined with other qualitative methods, such as manual review or deeper text analysis.

6.3 Practical Applications

The findings from this study offer numerous practical applications across several domains:

- **Book Recommendations**: The correlations between certain topics and higher review scores can be leveraged to improve book recommendation algorithms. By prioritizing books that align with highly rated themes, platforms can recommend titles more likely to resonate positively with users. Similarly, for readers who enjoy detailed narratives and character-driven stories, platforms can highlight books that focus on **topic_0 (story-character)** or similar themes. Conversely, identifying polarizing topics like **topic_23 (love-romance)** could inform recommendations for readers who prefer controversial or debated content.
- **Literary Analysis**: Researchers and literary critics can use these insights to understand how different themes affect reader engagement and perception. By analyzing which topics lead to higher engagement, literary scholars could trace trends in popular literature and how certain themes have evolved over time. For example, the detailed analysis of **topic_1 (family/character development)** being associated with positive review scores offers valuable context for studies on the cultural impact of familial themes in modern literature.
- **Marketing Strategies**: Publishers and authors can use these findings to craft more targeted marketing campaigns. Books that focus on highly rated themes like character development, storytelling, or unique plot structures can be marketed to maximize their potential for high reader satisfaction. Additionally, understanding the correlation between popularity and sentiment for controversial topics like **topic_23** can help in tailoring marketing strategies to appeal to readers who enjoy emotionally charged or divisive content. Identifying these polarizing themes early on can help publishers better frame the narrative around such books to manage reader expectations and enhance market reception.

7. Conclusion

7.1 Summary of Findings

This study applied topic modeling and sentiment analysis to a dataset of book reviews to uncover valuable insights into reader behavior and preferences. The key findings can be summarized as follows:

- **Topic Modeling and Review Length**: The analysis of detailed versus brief reviews showed that certain topics, such as **topic_0 (story-character)** and **topic_4 (death-war)**, are more frequently discussed in detailed reviews. These findings suggest that readers who write more in-depth reviews tend to engage more deeply with themes related to storytelling and character

development. In contrast, brief reviews often touch on less complex themes, which may reflect a more surface-level engagement with the book.

- **Topic Trends Over Time**: The heatmap illustrating topic trends across different years revealed how the popularity of certain themes has shifted over time. Notably, topics related to **romantic themes (topic_23)** and **character development (topic_1)** showed consistent prevalence, while others like **topic_22 (fantasy-worldbuilding)** saw more fluctuation. These trends offer insights into changing reader preferences and broader shifts in literary themes across genres and time periods.
- **Impact on Sales and Popularity**: The correlation analysis between topics and review scores, as well as ratings count, revealed interesting patterns. Topics like **topic_0 (story-character)** were positively correlated with both higher review scores and popularity, indicating that books emphasizing strong narratives and characters tend to perform better. On the other hand, topics like **topic_23 (romance-love)** showed a negative correlation with review scores, highlighting a potential disconnect between popular themes and actual reader satisfaction.

The integration of sentiment analysis further supported these findings, adding depth to our understanding of how readers emotionally respond to specific topics.

7.2 Contributions to the Field

This study makes several important contributions to Digital Humanities research:

- **Combining Topic Modeling and Sentiment Analysis**: By merging topic modeling with sentiment analysis, this research offers a more nuanced approach to understanding reader engagement. This method allows for a dual exploration of the themes present in reviews and the emotions associated with them, contributing new insights into how different book themes influence reader perception and satisfaction.
- **Insights into Reader Behavior**: The study provides clear evidence that certain literary themes are strongly associated with positive reviews and reader satisfaction, while others elicit mixed or negative reactions. These findings offer a data-driven approach to literary analysis and provide a basis for further research on how narrative elements impact reader reception.
- **Methodology for Practical Applications**: The approach developed in this study is highly adaptable and can be applied to various datasets in the book industry, marketing, or literary research. The findings and methodology contribute to the growing body of work in computational literary analysis, showing how advanced techniques can be leveraged to explore human engagement with texts.

7.3 Future Work

There are several areas for future research that could build on the findings of this study:

- **Applying the Methodology to Other Datasets**: While this research focused on book reviews, the methodology could be applied to other domains, such as film or product reviews, to investigate how themes influence customer satisfaction across different industries. Future studies could also examine different genres or specific authors to identify patterns unique to certain types of literature.
- **Incorporating Advanced Techniques**: Future research could benefit from the integration of more advanced machine learning techniques, such as deep learning models for sentiment analysis, which may provide even more granular insights into reader emotions. Additionally, incorporating user demographic data could help explore how themes resonate differently across diverse reader groups.
- **Exploring Reader Intentions**: Another avenue for further exploration would be investigating reader intentions, such as comparing the expectations of readers who buy books for leisure versus those who seek intellectual engagement. This could provide a more complete understanding of how and why readers respond to certain topics, ultimately enhancing the predictive power of book recommendation systems.

References

1. *Topic Modeling: Perspectives From a Literature Review ANDRÉS M. GRISALES A. 1 , SEBASTIAN ROBLEDO 1 , AND MARTHA ZULUAGA 2*

2. *Schöch, C. [2]. Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. University of Würzburg*

3. *Choi, Y., Joo, S., Topic Detection of Online Book Reviews: Preliminary Results. Valdosta State University*

4. *Zhang, Y., Xu, H.SLTM: A Sentence Level Topic Model for Analysis of Online Reviews. University of Massachusetts Dartmouth.*

5. *Sutherland, I., Sim, Y., Lee, S. K., Byun, J., Kiatkawsin, K., et al. [2]. Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation. Sejong University*

6. *Vayansky, I., Kumar, S.A.P.A Review of Topic Modeling Methods.*

7. *Kherwa, P., Bansal, P.Topic Modeling: A Comprehensive Review. Maharaja Surajmal Institute of Technology*

8. *Ramage, D., Rosen, E., Chuang, J., Manning, C. D., McFarland, D. A., et al. Topic Modeling for the Social Sciences. Stanford University.*

9. *Schröter, J., Du, K., (2022). "Validating Topic Modeling as a Method of Analyzing Sujet and Theme", Journal of Computational Literary Studies 1.*

10. *Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. Machine Learning*

11. *Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., Hassan, A., et al. [2]. Topic Modeling Algorithms and Applications: A Survey.*

12. *Choi, H. S., Lee, W. S., Sohn, S. Y., et al. [2]. Analyzing Research Trends in Personal Information Privacy Using Topic Modeling.*