UNIVERSITY OF MARYLAND 18 56

ROBERT H. SMITH
SCHOOL OF BUSINESS

Enterprise Cloud Computing and Big Data

BUDT758J

Professor John Silberholz

Team Members:

Anushka Jain (ajain10)

Loveleen Kaur (lkaur)

Megha Jahnavi Mudigonda (jahnavim)

Manasvi Surasani (maivy)

Palak Kishore (kish0507)

Section 1: Introduction

In today's fast-evolving publishing industry, catering to reader interests is crucial for driving engagement and impact. With an abundance of book choices and the rise of platforms like Google Books and Hardcover, readers actively share feedback on what they enjoy, dislike, or abandon. However, traditional publishing strategies are yet to integrate this rich qualitative feedback into decision-making.

This project is designed to support our decision makers, the publishing houses — specifically, editorial directors and marketing strategists — in identifying high-potential genres and reader engagement drivers using a data-driven approach. The central question we tackle is:

> **What do reader reviews and ratings reveal about shifting genre trends and engagement factors that publishers can use to guide future releases?**

Our goal is to mine reader sentiment and topics across hundreds of reviews to surface which genres captivate audiences and what themes or pain points consistently emerge. These insights can help publishers refine their catalogs, target promotional strategies, and plan more reader-aligned book launches.

Section 2: Data Sources & Techniques

We relied on two core data sources to balance reader reviews with structured metadata:

Source	Details	
Hardcover (Webscraped)	~300 reviews across 300 books (2014–2024) which provides rich qualitative text	https://hardcover.app
Google Books API using Google Play Links	Metadata for books across 2014-2024, including average ratings & count of ratings	https://www.googleapis.com/books/v1/volumes

Each review on Hardcover includes the review text, book title, author, and post date. Google Books adds an additional numeric layer with rating count, average score, and genre classifications.

Data Pipeline & Integration:

- Cleaning: Reviews were cleaned using spaCy (stopwords, lemmatization) and non-English reviews were retained and tagged to capture multilingual trends.

- Sentiment Analysis: vaderSentiment was used to score each review for polarity.

- Topic Modeling: We used BERTopic, which combines Sentence Transformers with UMAP (for dimensionality reduction) and HDBSCAN (for clustering).

Key software/tools:

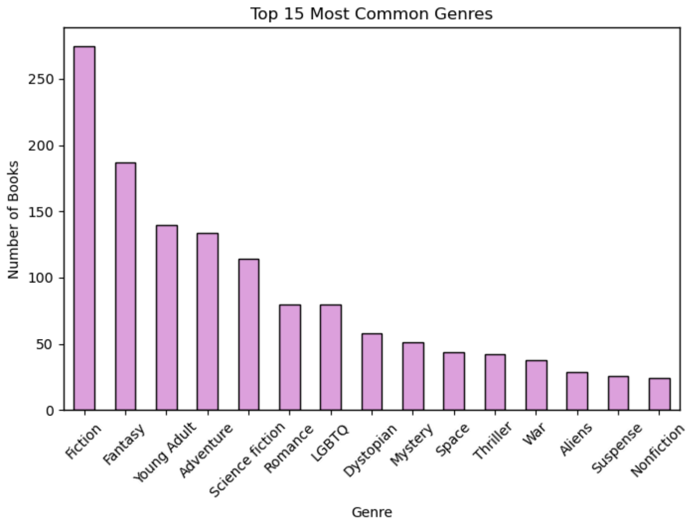
- Data wrangling & scraping - *pandas, BeautifulSoup*
- Text preprocessing & sentiment scoring - *spaCy, vaderSentiment*
- Topic modeling - *BERTopic, umap-learn, hdbscan*
- Visualization - *matplotlib, seaborn, WordCloud*

Section 3: Results

Our multi-stage analysis produced valuable insights into genre trends, reader sentiment, and the thematic diversity present in book reviews. These findings provide actionable evidence for publishers seeking to optimize their editorial and marketing strategies.

Genre Insights

Our data analysis revealed that Fiction was the most prevalent genre, mapping to 138 topics. It was followed closely by Fantasy, Young Adult, and Adventure. These genres not only dominated in frequency but also demonstrated significant topic diversity, suggesting deeper engagement and broader thematic exploration by readers.



The distribution of topics across genres aligned closely with genre frequency, offering a form of validation for our BERTopic model. Genres that appeared more often in the dataset also exhibited a higher number of unique topic clusters. This consistency underscores the representativeness of our dataset and indicates that highly prevalent genres also inspire richer and more varied discussions, which is critical for understanding reader engagement at a deeper level.

Topic Modeling Insights

Applying BERTopic, we initially identified 314 raw topics, which were reduced to 141 well-defined clusters through dimensionality reduction (UMAP) and density-based clustering (HDBSCAN). The refined topics revealed a wide spectrum of reader conversations, offering insight into emotional response, content format preferences, recurring narrative structures, and even cultural or linguistic segmentation. These patterns go well beyond conventional rating metrics, surfacing the elements of storytelling that consistently drive engagement.

Several key insights emerged:.

- Series- and sequel-related themes featured heavily in discussions, especially within Fantasy and Science Fiction genres. These clusters reflect a deep engagement with long-form storytelling, serial characters, and extended worldbuilding, suggesting high retention and repeat readership.

- Multilingual clusters, particularly in Spanish and Dutch, were prominent and coherent. Their presence highlights the platform's global readership and signals potential for localized content strategy and international marketing expansion.

- One cluster focused on reviews deliberating between star ratings—specifically between 3.5 and 4 stars. These reviews revealed subtle forms of dissatisfaction or tempered praise, pointing to nuanced reader sentiment that may not be fully reflected in numerical scores.

- Several clusters captured emotional reactions to endings, both positive and critical. Phrases like “I didn’t expect that,” or “what an ending” dominated, indicating that the conclusion of a narrative is a key driver of lasting reader sentiment.

- A distinct cluster highlighted feedback on audiobook experiences, with readers commenting on narration quality, pacing, and listening habits. This suggests that format plays a significant role in how content is received and remembered.

- Perhaps most notable were fan-driven clusters focused entirely on specific authors and series—particularly *Brandon Sanderson’s Cosmere* and *Martha Wells’ Murderbot Diaries*. The emergence of such detailed, author-centric clusters underscores the influence of loyal fan communities and offers opportunities for publishers to invest in franchise-building and targeted promotional strategies.

Collectively, these insights reflect a rich and diverse reader ecosystem, where engagement is driven not just by genre, but by emotional resonance, delivery format, and the ability to foster long-term investment in story worlds. For decision makers, these topic clusters offer a roadmap for where and how to deepen audience connection—from optimizing plot structure and endings, to exploring audiobook markets, to identifying breakout series with fanbase momentum.

Sentiment Trends

Sentiment analysis using VADER revealed strong correlations between review polarity and overall book ratings, as expected. Conversely, several books with solid average ratings exhibited lower sentiment scores, typically due to consistent criticisms around unsatisfying endings or erratic pacing. This analysis helped surface both the themes that readers appreciated and those they responded to less favorably.

Platform Comparison

A comparative assessment of our two primary sources — Hardcover and Google Books — highlighted key differences in data richness and usability:

- Hardcover emerged as the more valuable platform for this analysis. It offered diverse, high-quality reviews and a greater depth of sentiment and thematic variation, making it ideal for extracting reader engagement signals.
- Google Books, while useful for structured metadata (ratings, counts, genres), lacked the granularity and emotional nuance necessary for meaningful text-based insights.

Section 4: Conclusion & Next Steps

This project illustrates the value of treating reader reviews not just as feedback, but as strategic signals. By combining text analytics with genre-level insights, we've laid the foundation for a more responsive, reader-centric publishing strategy. To act on these findings, we recommend the following steps:

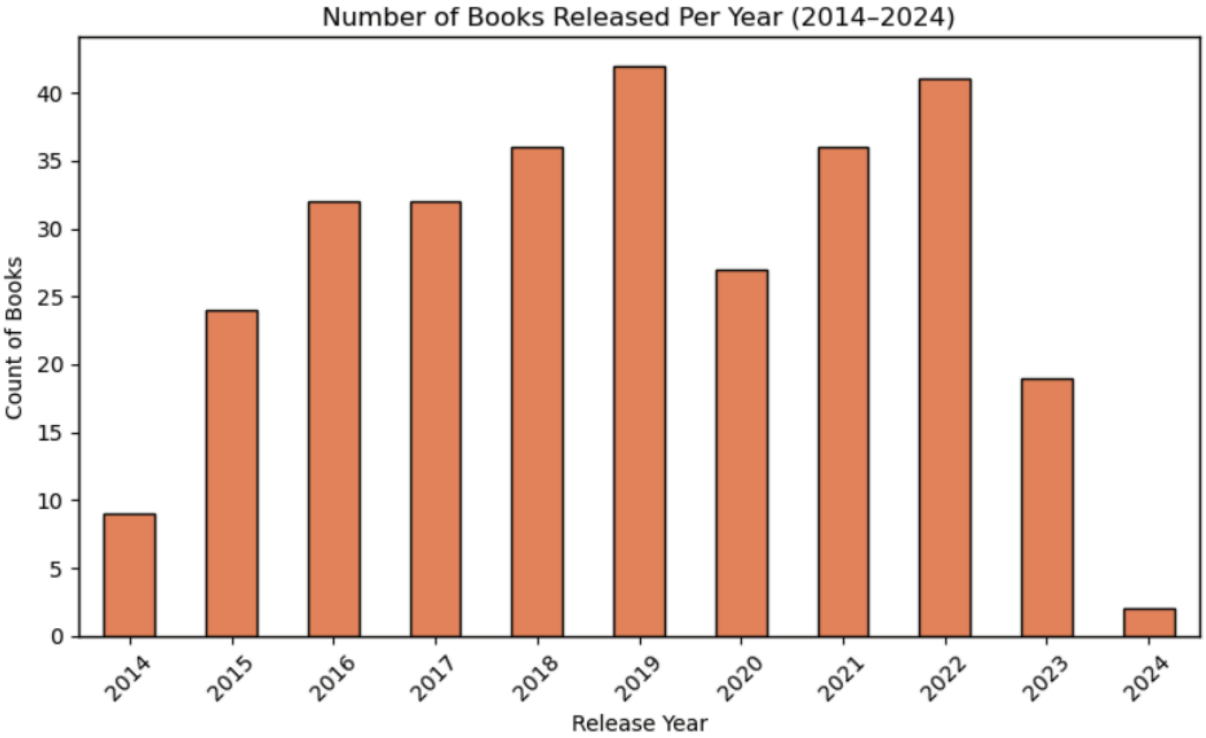
- Prioritize high interest genres like Fantasy, Young Adult, and Adventure and focus on key engagement drivers such as strong plot endings, compelling series, and high-quality audiobooks.
- Leverage Hardcover as a primary platform for ongoing sentiment and trend analysis, given its richer and more varied user feedback.
- Explore global reach by tapping into multilingual audiences, as evidenced by Spanish and Dutch review clusters.

Going forward, expanding to platforms like Goodreads and automating review monitoring can help publishers stay ahead of genre trends and reader expectations—placing audience insight at the heart of strategic planning.

Appendix

Part I: Data Collection

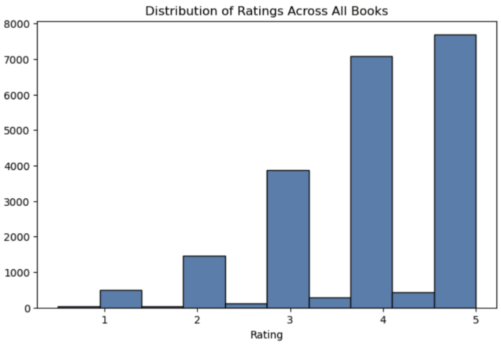
1. After analyzing various book review platforms, we finalized Hardcover and Google Books as our primary data sources. Review data was sourced from Hardcover, comprising over 21,000 user reviews across 300 books published between 2014 and 2024. We selected a 10-year time frame to ensure broad representation and avoid recency bias in our analysis. Python libraries like Pandas, BeautifulSoup and html were used.



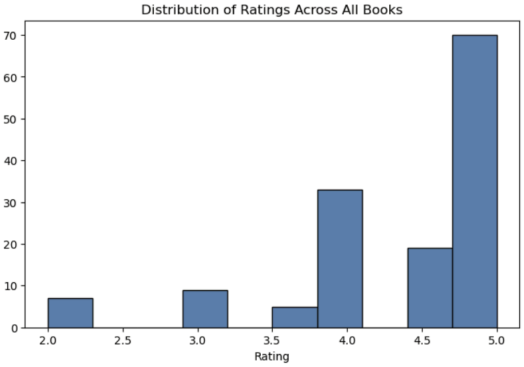
2. To scrape review and metadata information from Hardcover, we used an incremental data collection approach, beginning with a single book and progressing to sets of 50 and eventually 300 books. It allowed us to debug our scraping logic, manage resources efficiently, and verify the quality and consistency of the data before moving to large-scale extraction. It also helped us handle books of varied review quantity and structure, thus strengthening our data pipeline to be more scalable and robust.
3. Additional metadata (e.g., average rating, ratings count, publication date) was retrieved from the Google Books API using ISBN-13 numbers. We employed an incremental ISBN query strategy to obtain maximal metadata coverage from the Google Books API. Iterating over all ISBNs available for each book and storing the first hit with a valid rating, we significantly reduced missing values and maximized the external metadata completeness, which was critical for reliable platform comparison.
4. The final Google Books dataset included 143 matched books, highlighting the API's limited coverage for newer or niche titles.

Part II: Platform Comparison

1. Google Books' users' rating averages and reviews were compared to Hardcover's reviews.
2. Even though Hardcover provided a richer dataset in both size and polarity, Google Books was sparse, only 47% of the books included any rating metadata.
3. Google Books Review ratings were more positively biased because of filtering and low user engagement, whereas Hardcover had more spread and balanced ratings.



Ratings on Hardcover



Ratings on Google Books

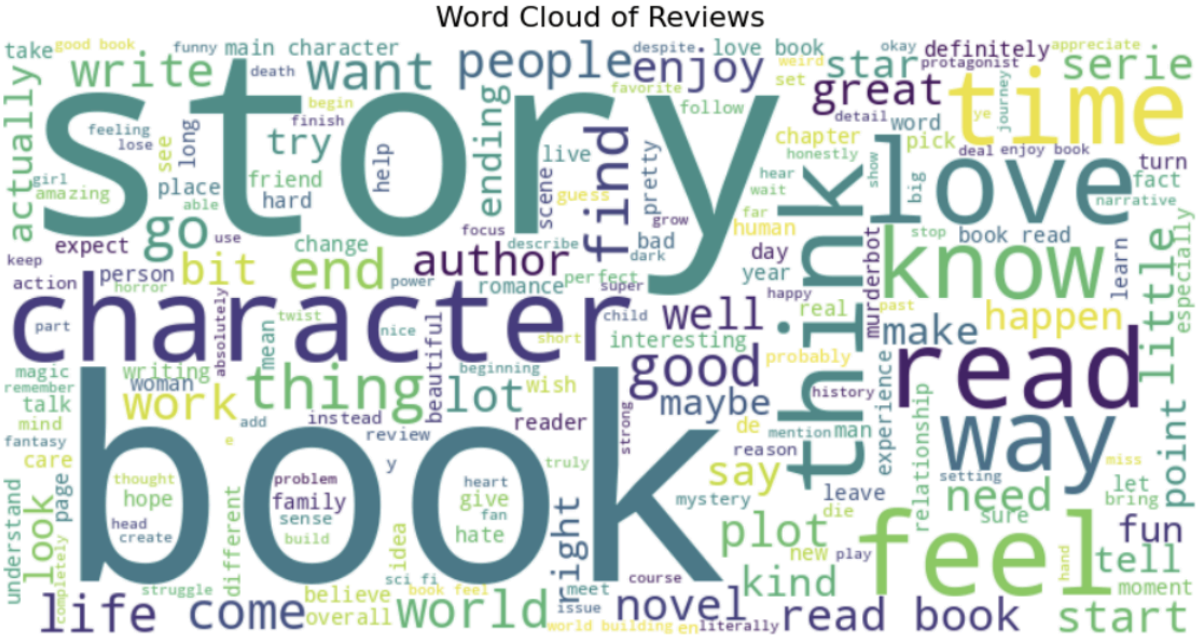
Part III: Data Preprocessing

1. Before textual analysis, we performed thorough data preprocessing using spaCy, a widely used natural language processing library. The aim was to standardize and clean the review text for meaningful analysis in downstream processes such as word cloud generation and topic modeling.
2. Text was converted to lowercase to prevent duplication in word frequency (for example, "Book" and "book").
3. We applied tokenization by utilizing spaCy's tokenizer to divide the text into individual tokens and subsequently lemmatization for converting each word to its root form (e.g., "reading" to "read"). This helped to combine word frequency across variations of the same word, improving accuracy in frequency-based analysis.
4. The SpaCy's default English stopword list was utilized to eliminate common but semantically weak words such as "the," "and," "is," etc. This ensured that the most critical and informative terms were highlighted in our visualizations.
5. Additional cleaning included punctuation removal, special character removal, and very generic words such as "book" and "read" that filled the dataset but offered no analytical value.

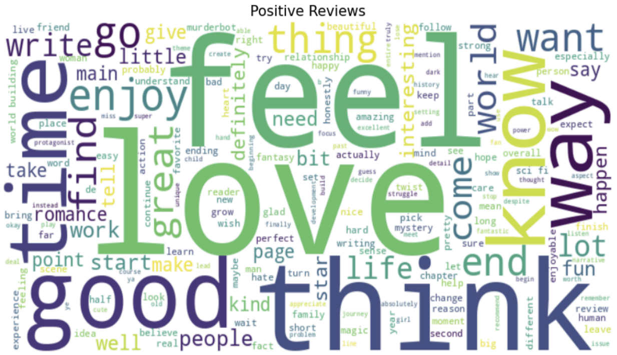
Part IV: Word Cloud Generation

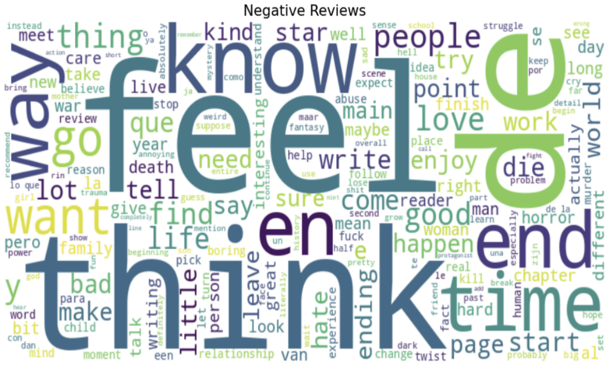
1. To begin our text analysis, we constructed an initial word cloud of all user reviews. We used this approach as a quick and easy method of visualizing the most frequent words within the dataset. Word clouds are well-suited to providing a general impression of text patterns through the display of words in relation to frequency of usage, and therefore make useful exploratory tools for spotting typical

language and recurring themes. But the first word cloud largely produced generic words such as "book," "read," "story," and "character," which, while to be expected, were not very heavy with interpretive significance for further analysis.



2. To achieve this, we advanced our approach by sentimentally dividing the reviews and then categorizing them as positive, negative, or neutral using VADER sentiment analysis. This allowed us to create individual word clouds for positive and negative reviews according to the variation in emotional and thematic content between them.
3. The final positive and negative word clouds resulted in distinct emotional flavors. The positive reviews were dominated by words like "beautiful," "favorite," "cry," "wow," and "amazing," which expressed strong emotional attachment and happiness. The negative reviews were dominated by words like "boring," "slow," "predictable," and "disappointing," expressing distinct reader discontent areas.





4. By combining spaCy's advanced NLP capabilities with sentiment filtering and word cloud visualization, we were able to get beyond surface-level term frequencies and uncover emotionally charged and theme-relevant linguistic patterns.

Part V: Sentiment Analysis

1. Sentiment scores were calculated using VADER from NLTK’s SentimentIntensityAnalyzer.VADER works by assigning a sentiment score to each word in the text based on its emotional tone,for example, words like “amazing” or “love” have high positive scores, while words like “boring” or “terrible” have strong negative scores.
2. Reviews were classified as:
   Positive: if compound score ≥ 0.05
   Negative: if ≤ -0.05
   Neutral: otherwise
3. 69% of reviews were positive, 17% negative, and 12% neutral.
4. This method was ideal for our project because it allowed us to quickly and reliably classify thousands of short user reviews into positive, negative, or neutral sentiment.

Part VI: Topic Modelling

1. We initially used BERTopic in its default configuration, which yielded 314 distinct topics from the review text. While this provided us with a wide range of clusters, the majority of the resulting topics were overly fragmented with overlapping topics or very general keywords, which made them meaningless. This indicated that the model was capturing noise and subtle differences in words as entirely different topics, reducing the output's clarity and usefulness.
2. To enrich the semantic value and consistency of the subjects, we enriched the BERTopic pipeline with more advanced modules:

Sentence Transformers:

3. We used the pre-trained model `all-MiniLM-L6-v2` of the Sentence Transformers library to generate contextual embeddings for each review. Unlike regular bag-of-words models, these embeddings consider meaning and context at the whole-sentence level. This allowed BERTopic to group reviews based on true semantic similarity, rather than word overlap.

UMAP (Uniform Manifold Approximation and Projection):

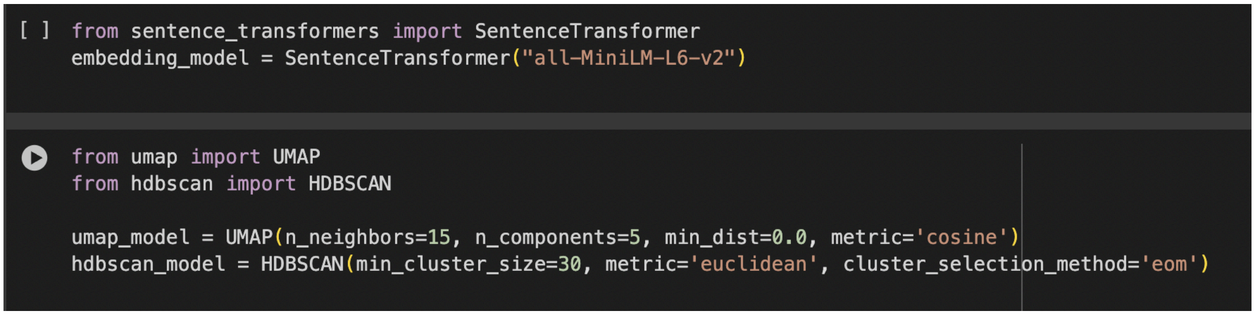
4. Sentence embeddings have a very high dimensionality (384+ features), so clustering would be inefficient. UMAP reduced these to 5 dimensions and preserved the structure of data. We selected:

- `n_neighbors=15` to take into account local groups of reviews that are similar
- `n_components=5` for the destination space to reduce
- `min_dist=0.0` to allow for tight clustering
- `metric='cosine'` for direction-based (not magnitude-based) similarity measurement between sentences

HDBSCAN (Hierarchical Density-Based Spatial Clustering):

5. We used HDBSCAN to cluster the dimensionally projected embeddings. It automatically recognizes clusters of any shape and size and filters out noise instead of compressing all points into a single group. We initialized:

- `min_cluster_size=30` in order not to have too small, less prominent clusters
- `metric='euclidean'` to calculate distance in the UMAP-reduced space



```
[ ] from sentence_transformers import SentenceTransformer
    embedding_model = SentenceTransformer("all-MiniLM-L6-v2")
```

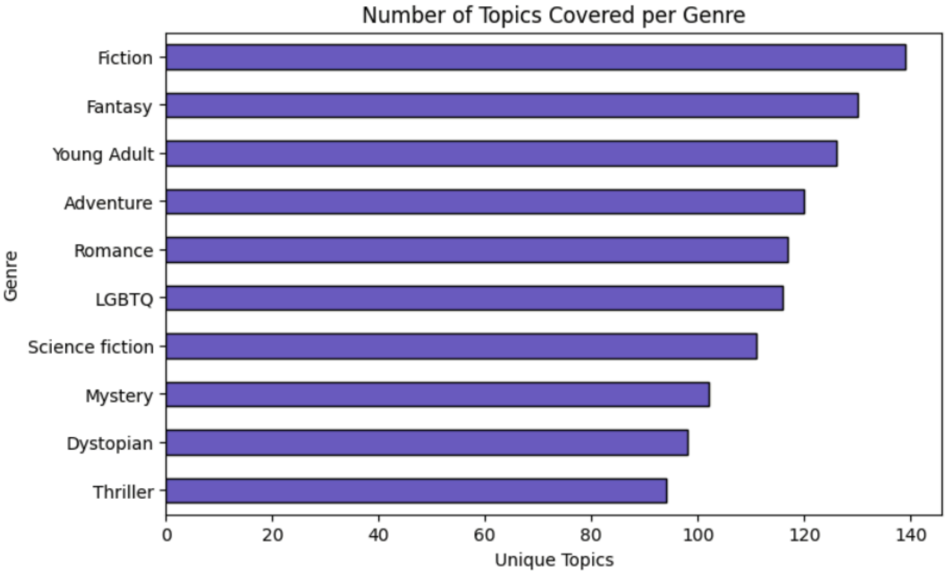
```
from umap import UMAP
from hdbscan import HDBSCAN

umap_model = UMAP(n_neighbors=15, n_components=5, min_dist=0.0, metric='cosine')
hdbscan_model = HDBSCAN(min_cluster_size=30, metric='euclidean', cluster_selection_method='eom')
```

6. With these enhancements, the count of topics dropped from 314 to 141, providing us with fewer but more solid topics. These enhanced clusters were more robust reader discussions and less confusing. The majority of topics fell neatly into a specific genre (like fantasy or romance), author enthusiasts (like Stephen King fans), or frequent themes (like audiobook narrations or plot twists), making our quality of insights much better as well as genre-topic mapping easier.

Part VII: Topic-to-Genre Mapping

1. Each review had one or more associated genres.
2. After topic modeling, we exploded genre lists and counted the genre distribution within each topic.
3. Each topic was labeled by its most dominant genre using normalized proportions.
4. This mapping allowed us to analyze which genres contributed most to topic diversity.



Part VIII: Additional Recommendations

This project underscores the long-term value of using reader reviews as strategic intelligence rather than passive feedback. To build on our findings, we recommend the following next steps:

1. Integrate insights into editorial workflows: Use topic and sentiment signals to support manuscript evaluation, genre positioning, and campaign messaging—ensuring alignment with what readers truly value.
2. Develop internal monitoring tools: Build lightweight dashboards or alert systems to track emerging reader themes, emotional tone shifts, and format preferences across reviews over time.
3. Move from reactive to proactive planning: Use real-time review analysis to anticipate trends—allowing publishers to spot rising subgenres, adapt to sentiment shifts, and plan more responsive catalog strategies.
4. Use reviews to validate creative decisions**:** Let reader-driven insights inform not just marketing, but the creative process itself—supporting decisions on story structure, character development, and series potential.

By embedding reader feedback into operational strategy, publishers can remain agile, audience-aligned, and better positioned to succeed in a competitive and fast-moving market.