THE BESTSELLER ALGORITHM

What the Data Says to Publish Next

Loveleen Kaur

Anushka Jain

Megha Mudigonda

Manasvi Surasani

Palak Kishore

OUR DECISION MAKER AND THEIR OBJECTIVE

DECISION MAKERS

Publishing houses focused on aligning their editorial strategy with data-driven insights into genre trends and reader engagement

KEY QUESTION

What can pattern in reader reviews and ratings tell us about shifting genre preferences and the factors that drive reader engagement?

WHY IT MATTERS?

Identifying high-potential genres and understanding the drivers of reader engagement is essential for optimizing catalog offerings, planning release calendars, and targeting marketing efforts effectively.

DATA SOURCES & TECHNIQUES

SOURCE	TECHNIQUE	INFO	TOOLS
Hardcover	• WEBSCRAPING • TOPIC MODELLING	• DATASET INFO: – 200~300 USER REVIEWS – 300 BOOKS • TIME RANGE: 2014–2024	• spacy wordcloud • vaderSentiment • bertopic • umap-learn
Google Play	GOOGLE BOOKS API USING GOOGLE PLAY LINKS	• DATASET INFO: AVERAGE RATING & RATING COUNT FOR 143 BOOKS • TIME RANGE: 2014–2024	

INSIGHTS

Fiction appears 138 times is the strongest-performing genre, based on the approximately 21,515 values we categorized from the dataset. Given its popularity, it's reasonable to infer that closely related genres like Fantasy and Young Adult are also likely to rank high in popularity.

Number of Topics Covered per Genre
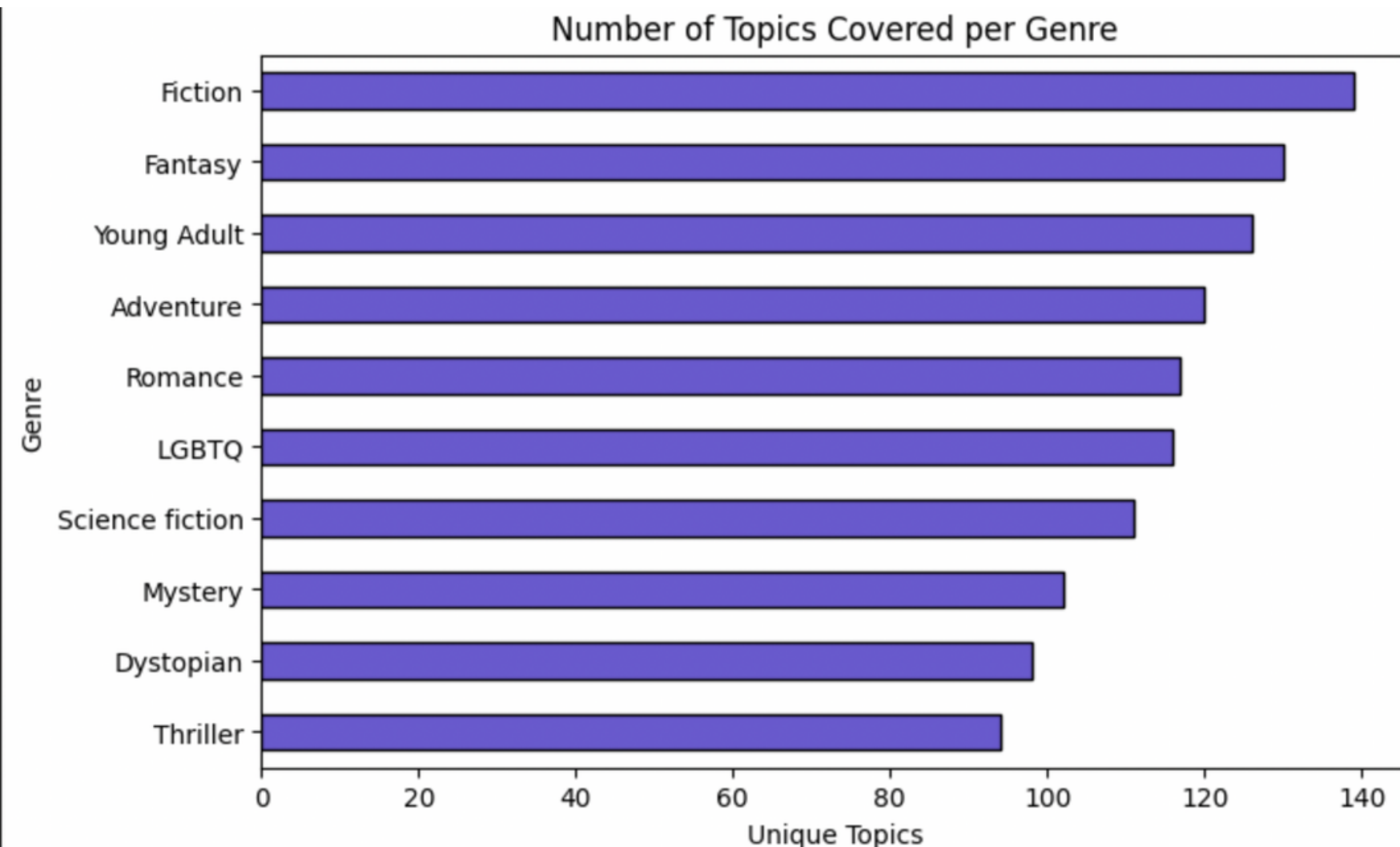
- The genres most common in our dataset (e.g., Fiction, Fantasy, YA, Adventure) also show the greatest topic diversity in BERTopic, validating both our EDA findings and the effectiveness of the topic modeling.

- Compared to Google Books, Hardcover offers broader coverage, more user engagement, richer and more reliable ratings, making it the superior platform for sentiment and rating-based analysis.

- Focusing on this genre and its related sub-genres can offer strong business potential for the publishing house while also creating a reliable and engaging space for readers.

- Enhanced topic modeling using Sentence Transformers, UMAP, and HDBSCAN reduced topics from 314 to 141, revealing clearer, more specific themes such as multilingual reviews (Spanish, Dutch), emotional reader feedback, and fan-favorite series like Murderbot Diaries and Brandon Sanderson's Cosmere, offering valuable insights into sentiment, engagement, and global reach.

CONCLUSION AND NEXT STEPS

Positive Reviews

Prioritize Popular Genres: Focus on Fiction, Fantasy, Young Adult, and Adventure — the most popular and emotionally engaging genres — for maximum reader interest and business impact.

Leverage Reader Insights: Focus on engaging series, strong plot endings, and high-quality audiobooks to align with reader sentiment and format preferences revealed through topic modeling.

Use Hardcover for Sentiment-Driven Insights: Hardcover's high user engagement and diverse rating distribution make it the more reliable platform for understanding reader preferences. This data can be used to fine-tune marketing strategies, identify breakout titles, and track evolving trends.

Global Reach Opportunities: Clusters identified Spanish and Dutch-language reviews. Signals potential for multilingual and international publishing strategies.

THE END