Beyond the Stars: A Text Mining Analysis of Reader Sentiment and Book Features in Amazon Reviews

Rohan Narasayya and Justin Huang
STAT 425 Final Project – June 17th 2025

Abstract

This study explores the linguistic and metadata-driven factors that influence book review scores on Amazon. By analyzing over 3 million reviews using natural language processing techniques such as sentiment analysis, topic modeling, and classification, we uncover patterns in review content, genre preferences, and rating behavior. Our work also evaluates how book features such as genre, author, and price correlate with review sentiment and reader satisfaction. Using Latent Dirichlet Allocation (LDA), classification models, and a recommendation system, we provide a comprehensive look at the drivers of book ratings and highlight the thematic content behind both highly rated and poorly rated books.

1. Introduction

Millions of readers rely on online reviews to guide book purchases. Amazon, in particular, serves as both a retail marketplace and a repository of reader sentiment. By merging structured book metadata with unstructured review text, we aim to answer key questions: What language patterns distinguish positive from negative reviews? Which authors and genres earn higher ratings? Can we predict reader satisfaction or recommend books based on review behavior?

2. Data Description

We utilized two datasets:

- **Amazon Book Review Dataset:** Over 3 million reviews with fields like `review/score`, `review/text`, `review/time`, etc.
- **Book Metadata Dataset:** Structured features for each title, including `title`, `authors`, `categories`, `price`, `publisher`, and `publishedDate`.

After merging on title, we cleaned and sampled subsets for different analyses.

Table 1: Codebook for Amazon Reviews Dataset

Feature	Description
id	The ID of the book
Title	Book title
Price	The price of the book
User_id	ID of the user who rated the book
profileName	Name of the user who rated the book
review/helpfulness	Helpfulness rating of the review (e.g. 2/3)
review/score	Rating from 0 to 5 for the book
review/time	Time of the review
review/summary	Summary of the text review
review/text	Full text of the review

Table 2: *

Codebook for Book Metadata Dataset

Feature	Description
Title	Book title
Descripe	Description of the book
authors	Name(s) of book author(s)
image	URL for book cover
previewLink	Link to access book on Google Books
publisher	Name of the publisher
publishedDate	Publication date
infoLink	Link for more information on Google Books
categories	Genres of books
ratingsCount	Number of ratings for the book

3. Research Questions

1. **Review Sentiment vs. Score**
    - What are the most common words in 1-star vs. 5-star reviews?
    - How do polarity and subjectivity vary by rating?
2. **Book Metadata and Ratings**
    - Which genres and authors receive the highest scores?
    - Does price affect score or sentiment?
3. **Topic Modeling and Themes**
    - What themes emerge in high vs. low-rated books?
    - Are themes genre-dependent?

4. **Classification and Recommendations**
   - How well can models predict whether a review is positive?
   - Can we build a basic book recommendation engine?

4. Sentiment and Word Analysis

4.1. Top Words in One-Star vs Five-Star Reviews

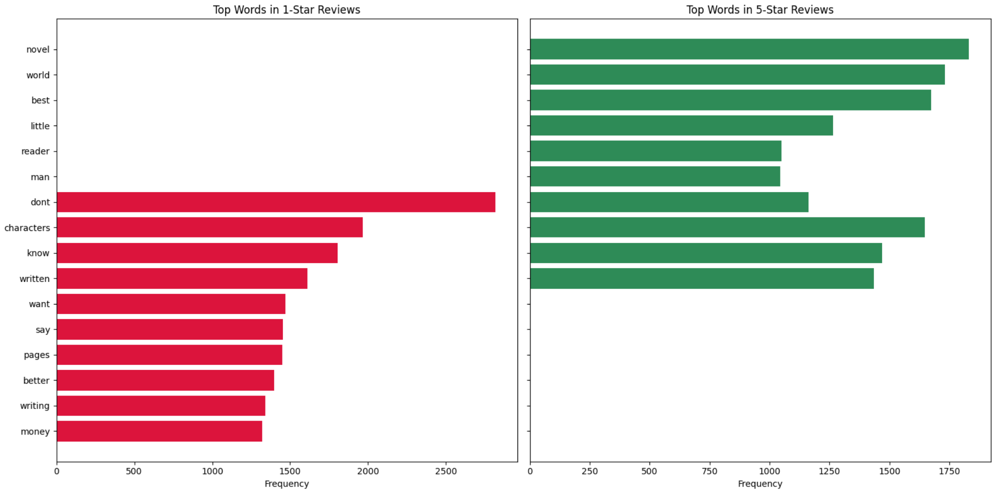

Figure 1: Top words in 1-star vs. 5-star reviews

Using tokenization and stopword removal, we created the side-by-side plot above showing the top 10 words for both one- and five-star reviews. We can see that there is actually some overlap between the two types of reviews. Words like don't, character, know, and written are in the top 10 for both. This makes sense since these words can be used in both a positive and a negative context.

4.2. Polarity vs Subjectivity

Polarity measure the emotional tone, meaning how positive or negative the text is. Subjectivity measures how opinionated or emotional a piece of text is. Using the TextBlob library, we made this plot of polarity and subjectivity vs. review score. We can see that polarity increases with review score, which makes sense since we would expect higher scoring reviews to be more positive. On the other hand, subjectivity is higher in extreme reviews, which also seems reasonable since extreme reviews will be more opinionated.

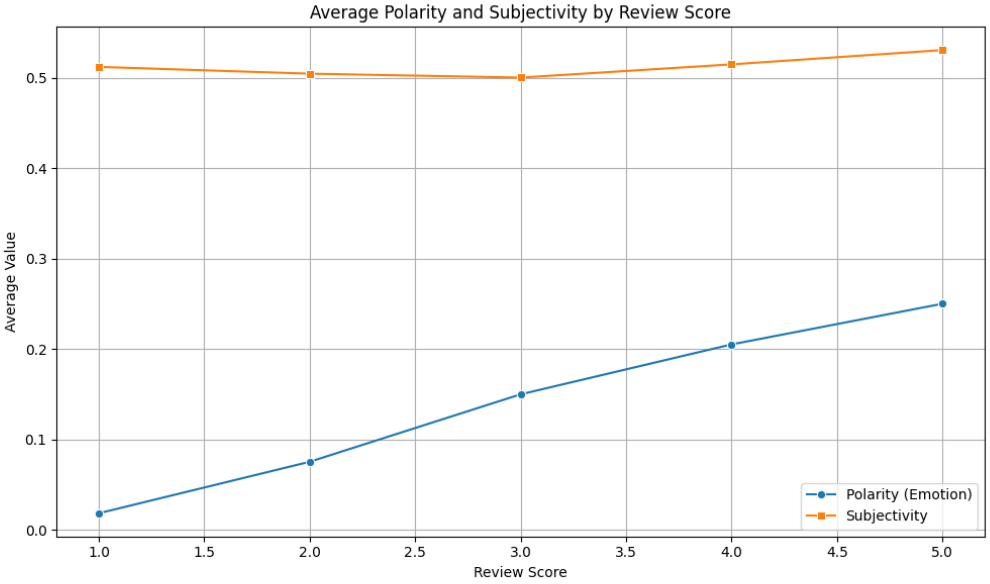

Figure 2: Average polarity and subjectivity by review score

5. Metadata Analysis

5.1. Genres

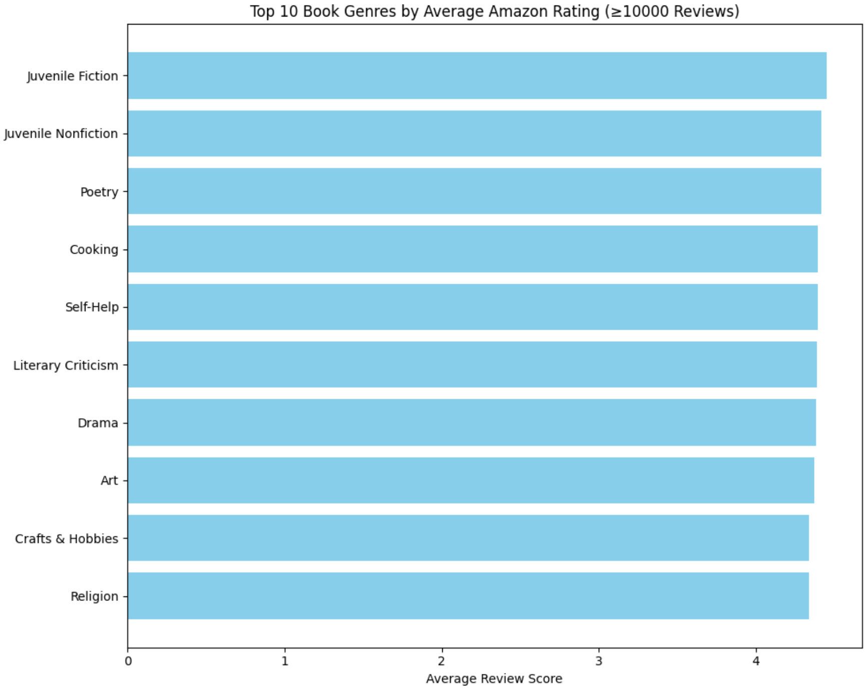

Figure 3: Average review scores across different genres

This plot above shows the top 10 genres by average review scores. There doesn't seem to be a clear pattern in which genres typically score higher as there is solid diversity among the top 10. One small pattern is that Juvenile Fiction and Nonfiction are the top scoring genres suggesting that books for a younger audience score better.

5.2. Publishers

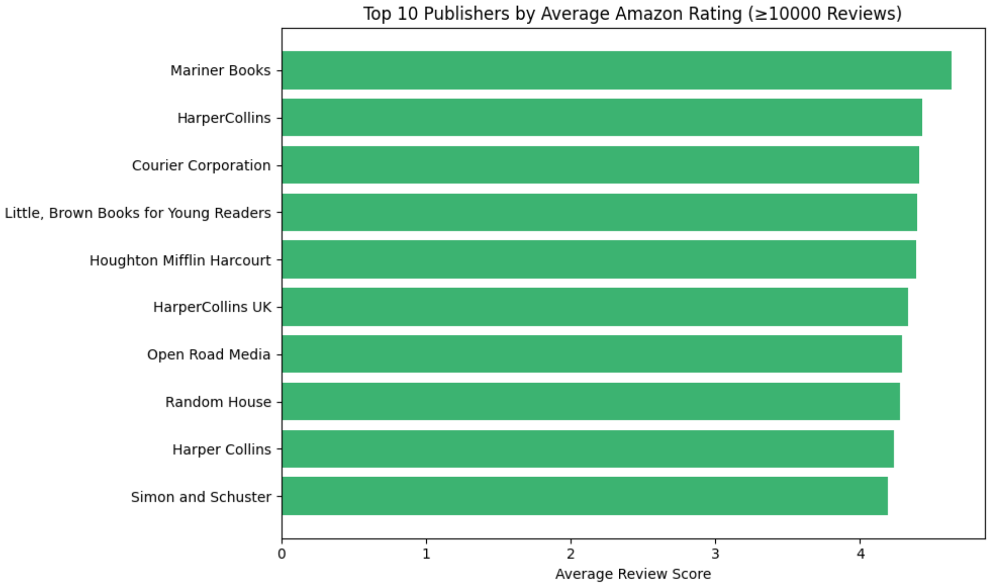

Figure 4: Average review scores by publisher

We can see that Mariner Books has the highest average rating by a sizable margin. One reason for this is because they have published iconic books like The Lord of the Rings, The Hobbit, 1984, Animal Farm, and The Handmaid's Tale. These books are beloved and are most likely to have high-score reviews that boost the publisher's rating.

5.3. Authors

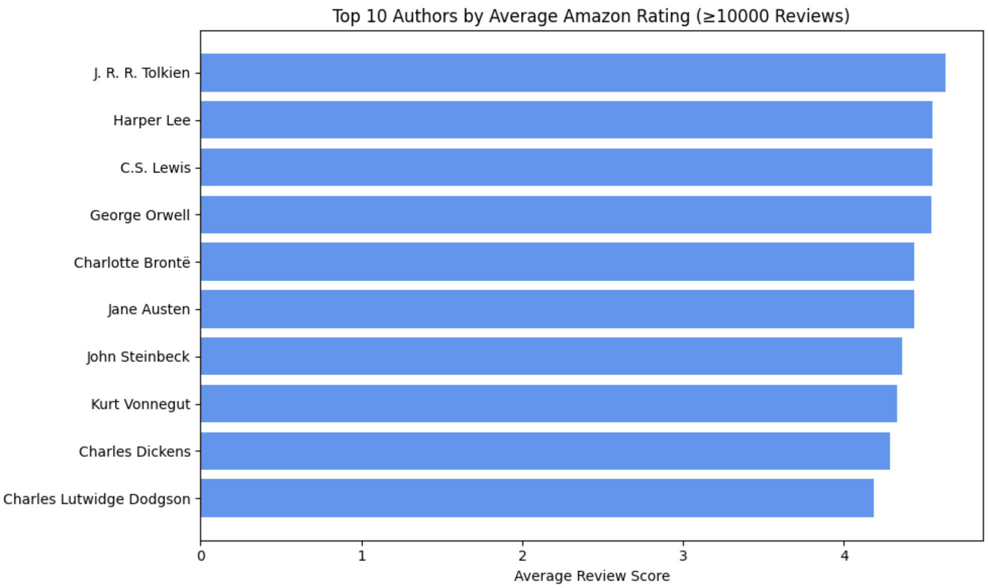

Figure 5: Average review scores by author

These authors are associated with beloved fantasy series and classic literature, which may contribute to their strong reception across audiences.

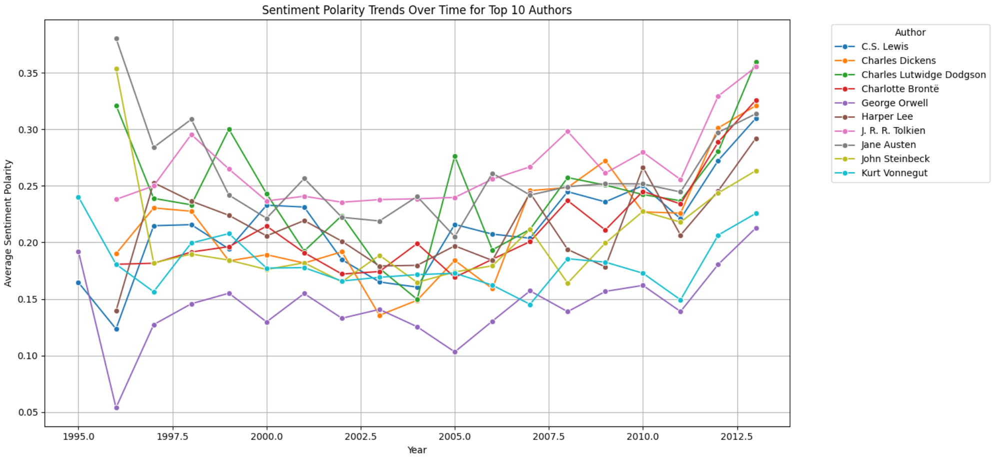

Figure 6: Average review scores of top authors over time

The plot above shows the sentiment polarity of reviews for these top 10 authors over time. Over time, sentiment polarity for most top authors exhibits a gradual upward trend, particularly after 2010. This may reflect shifting reader attitudes, where newer reviews are more generous or optimistic in tone—possibly due to increased fan engagement, nostalgia, or the influence of media adaptations that reignited interest in these works. For instance, authors like J.R.R. Tolkien and Jane Austen show noticeably rising sentiment, which could correlate with high-profile movie releases or cultural resurgence. Meanwhile, some authors exhibit more stable or fluctuating sentiment, suggesting diverse reader interpretations across time.

5.4. Price

The plots below showed book price has negligible correlation with both review score and sentiment.

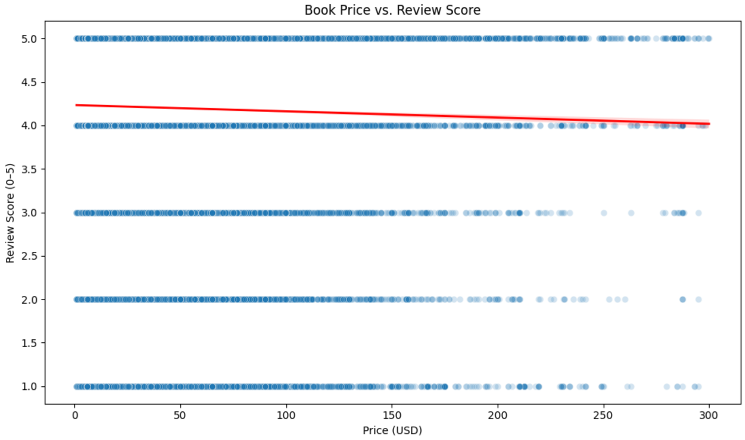

Figure 7: Relationship between book price and review score

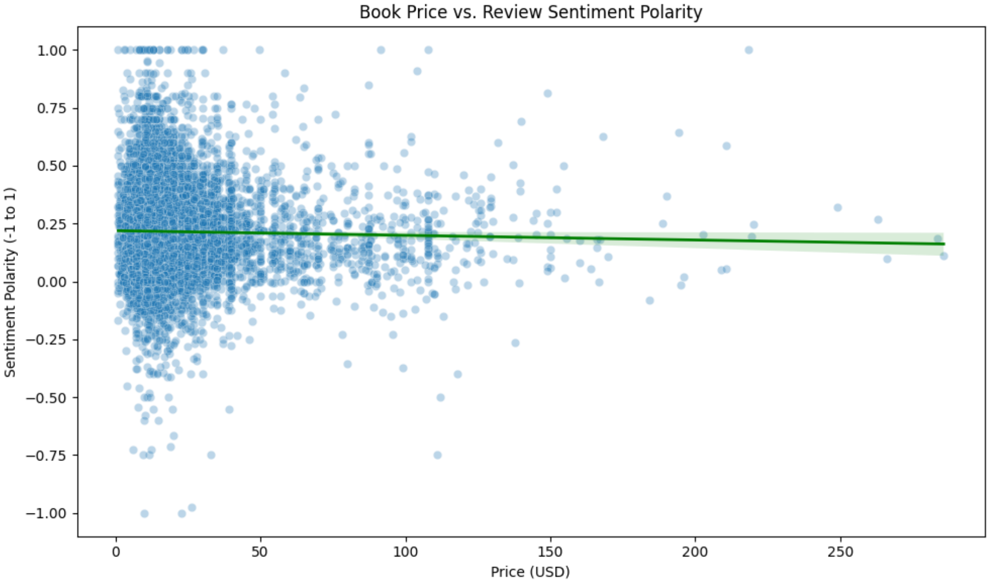

Figure 8: Relationship between book price and review sentiment polarity

This implies that readers are primarily influenced by content and quality rather than the book's listed price.

6. Topic Modeling (LDA)

We applied LDA to 5,000 samples each from 1-, 3-, and 5-star reviews and 3,000 reviews per genre.

6.1. Themes by Rating

One-Star Reviews

- **Topic 1: Frustration and Unmet Expectations** — *know, want, going, didnt, money, series, little, say, written, school*
- **Topic 2: Religious and Controversial Content** — *god, world, believe, fact, history, man, say, jesus, christian, bible*
- **Topic 3: Poor Format and Edition Quality** — *version, edition, page, buy, text, amazon, pages, kindle, money, copy*
- **Topic 4: Weak Plot and Writing Style** — *plot, writing, written, boring, character, bad, better, say, didnt, pages*
- **Topic 5: Technical Content and Confusion** — *code, information, real, written, use, little, know, lot, examples, used*

Three-Star Reviews

- **Topic 1: Character and Family Dynamics** — *man, know, family, old, young, didnt, woman, bad, girl, mother*
- **Topic 2: Informative but Uneven Reading Experience** — *information, interesting, writing, reader, better, little, looking, chapter, pages, page*
- **Topic 3: Middling Fiction Plots and Characters** — *plot, series, character, little, end, didnt, interesting, better, stories, thought*
- **Topic 4: Historical and Philosophical Themes** — *world, war, history, man, american, human, god, men, reader, society*
- **Topic 5: Technical or Instructional Writing** — *use, need, lot, different, know, better, little, want, used, edition*

Five-Star Reviews

- **Topic 1: Beloved Series and Characters** — *best, series, know, end, character, loved, things, world, want, stories*
- **Topic 2: Informative and Well-Written Nonfiction** — *information, excellent, easy, history, use, need, world, best, understand, knowledge*

- **Topic 3: Beautiful Editions and Practical Content** — *best, ive, written, black, bought, buy, edition, copy, beautiful, recipes*
- **Topic 4: Classic and Reflective Literature** — *little, world, old, lewis, man, reader, war, soul, day, jane*
- **Topic 5: Powerful Human and Social Themes** — *god, man, family, women, children, war, human, world, history, school*

Topic modeling revealed clear thematic distinctions across review scores. One-star reviews often expressed frustration, dissatisfaction with formatting, or objections to controversial content. Three-star reviews reflected mixed experiences, combining family and historical themes with neutral technical commentary. In contrast, five-star reviews were strongly associated with praise for beloved characters, informative nonfiction, and aesthetically pleasing editions. This progression highlights how sentiment in reviews shapes the dominant topics and signals different types of reader engagement at each rating level.

6.2. Themes by Genre

Self-Help Reviews

- **Topic 1: Mental Health** — *help, women, anger, men, helpful, understand, personal, positive, depression, difficult*
- **Topic 2: Personal Growth Through Life Stories** — *stories, things, know, want, best, ive, need, better, children, helped*
- **Topic 3: Inspirational and Reflective Writing** — *writing, world, stories, written, want, reader, page, believe, feel, little*
- **Topic 4: Mindfulness and Spiritual Development** — *mind, power, spiritual, human, world, body, help, tolle, thoughts, need*
- **Topic 5: Advice and Social Relationships** — *person, principles, advice, know, friends, use, business, better, say, best*

Cooking Reviews

- **Topic 1: Breadmaking and Baking** — *recipes, bread, baking, ive, recipe, bought, making, tried, wonderful, machine*
- **Topic 2: Culinary Cultures and Food Writing** — *food, wine, cuisine, world, history, interesting, little, french, writing, know*
- **Topic 3: Comprehensive Cookbooks** — *cooking, recipes, cookbook, cook, edition, food, pages, information, joy, use*
- **Topic 4: Dessert Recipes and Ingredients** — *recipes, recipe, cookbook, cook, ingredients, use, cake, food, chocolate, easy*

- **Topic 5: Family-Oriented and Easy Meals** — *recipes, cookbook, easy, cooking, food, cook, recipe, ive, family, ingredients*

Juvenile Fiction Reviews

- **Topic 1: Family and Coming-of-Age Themes** — *know, girl, things, world, father, character, end, want, family, young*
- **Topic 2: Tolkien-Inspired Fantasy Adventures** — *hobbit, bilbo, fantasy, lord, tolkien, adventure, rings, series, world, dwarves*
- **Topic 3: Illustrated Stories for Young Readers** — *little, stories, illustrations, young, family, children, movie, wonderful, child, loved*
- **Topic 4: Childhood, School, and Everyday Fun** — *old, year, kids, children, fun, loved, child, son, school, got*
- **Topic 5: Magical Worlds and Heroic Journeys** — *harry, potter, jonas, world, giver, community, boy, best, school, know*

Thematic analysis by genre revealed that reader expectations and review content are deeply tied to the type of book. Self-Help reviews focused on emotional healing, spiritual development, and practical advice, often using introspective and motivational language. Cooking reviews were rich in descriptive terms related to recipes, ingredients, and global cuisines, reflecting a blend of technical instruction and cultural storytelling. In Juvenile Fiction, topics centered on family, adventure, and school life, with distinct subthemes such as Tolkien-style fantasy and magical realism. These findings show how genre shapes not only the content of books but also the language readers use when evaluating them.

7. Classification Models

After exploring our dataset, we decided to implement classification models to predict whether a book would be recommended based on sentiment. We hypothesized that there is a correlation between the sentiment score (calculated using AFINN) and the recommendation label. This assumption was supported by initial exploratory analysis, including visualizations that showed a noticeable trend between higher sentiment scores and positive recommendations. The models we selected for this task were Logistic Regression, Random Forest, and XGBoost. Although we initially considered Support Vector Machines (SVM), we ultimately excluded it due to its long runtime and similar performance to the other models.

Model	Precision	Recall	Accuracy	F1 Score
Logistic Regression	0.50/0.81	0.07/0.98	0.80	0.45
Random Forest	0.37/0.82	0.15/0.94	0.78	0.54
XGBoost	0.36/0.80	0.15/0.94	0.80	0.51
Neural Network	0.40	0.50	0.80	0.45

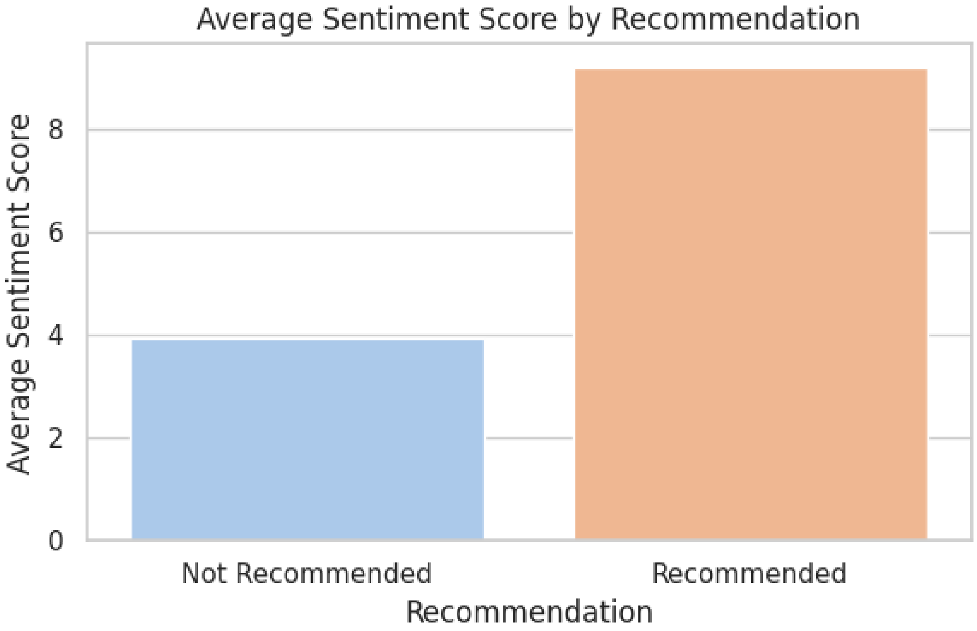

Figure 9: Relationship between sentiment scores and recommendation (4.0 >= *stars*)

7.1. Classification Models with SMOTE

During our initial model evaluations, we observed that recall values were relatively low across all classifiers. This is a common issue when working with imbalanced datasets, as is the case with Amazon reviews, which are heavily skewed toward positive ratings (See figure 10 below). As a result, our models tended to overpredict the positive class, leading to a high number of false positives and a lower F1 score. A distribution plot of the recommendation labels highlights this imbalance. To address this, we applied SMOTE (Synthetic Minority Oversampling Technique), which generates synthetic examples of the minority class to balance the training data. By incorporating SMOTE, we aimed to improve the model's ability to detect the minority (non-recommended) class, thereby increasing recall and improving the overall F1 score. This adjustment allowed for a more robust evaluation across both classes.

7.1.1 Smote results

Model	Precision	Recall	Accuracy	F1 Score
Logistic Regression	0.27 / 0.87	0.68 / 0.55	0.58	0.53
Random Forest	0.35 / 0.83	0.27 / 0.88	0.76	0.58
XGBoost	0.43 / 0.83	0.25 / 0.92	0.79	0.60
Neural Network	0.38 / 0.86	0.43 / 0.83	0.75	0.62

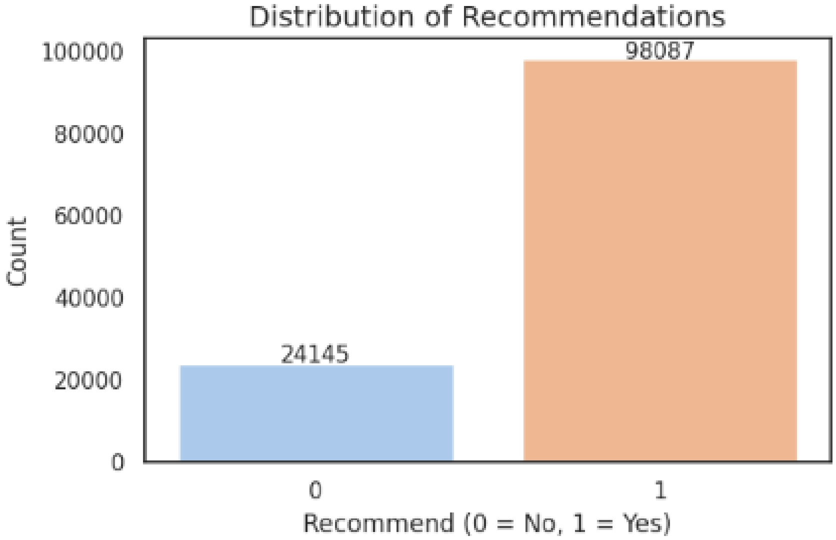

Figure 10: Recommendation distribution

7.2. Recommendation BPR Introduction

Recommendation systems play a significant role in shaping user experience across various platforms—from suggesting music during our daily commute to influencing purchase decisions on e-commerce websites like Amazon. In this project, we implement a recommendation model using Bayesian Personalized Ranking (BPR), which focuses on ranking items such that a user's preferred (positive) items are ranked higher than less relevant (negative) ones. Specifically, we aim to recommend books to users based on their previous interactions. The model is implemented using TensorFlow, and its performance is evaluated using three metrics: BPR loss, hit rate @10, and AUC (area under the ROC curve).

7.3. BPR Model

The construction of our BPR model involves several key components. First, we define latent variables: γ_u for user embeddings, γ_i for item (book) embeddings, and β_i for item bias terms. Using these, we calculate a preference score for each user-item pair (u, i). For training, we adopt a pairwise approach where, for each positive user-item interaction, we sample a negative item j that the user has not interacted with. Each training step involves sampling 5,000 (u, i, j) triplets, and we repeat this process for 1,000 iterations. This setup allows the model to learn to rank positive items higher than negative ones, ultimately reducing the loss and improving recommendation quality.

7.4. BPR model evaulation

7.4.1 Loss

To evaluate our model, we first examine the loss. The loss function with regularization is defined as $\mathcal{L}\text{BPR} = -\log \sigma(xui - x_{uj}) + \lambda|\mathbf{\Theta}|^2$, where we aim for the loss to decrease over training iterations. While the loss does initially drop, after around 1000 iterations the improvement becomes negligible, making further training less efficient relative to the run time. Our final recorded loss is **0.54**, which is acceptable but suggests limited gains from extended training.

7.4.2 Hit Rate @10

To compute the hit rate at 10, we evaluate whether the model can correctly identify a positive item among a set of negatives. For each user, we sample 99 negative items (items not interacted with) and 1 positive item (an item the user has interacted with). After scoring all 100 items for the user, we rank them. If the positive item appears in the top 10, it counts as a "hit." We repeat this process across 1000 users to obtain our hit rate. The final result is **0.491**.

7.4.3 Area Under the ROC Curve (AUC)

We also evaluate performance using the ROC AUC, which measures how often the model ranks a positive item higher than a negative one. For each user, we again sample a positive and multiple negative items. After scoring, we compute the percentage of negative items that have a lower score than the positive item. This gives us the AUC for that user, and we average across users for the final result. For our model, the final AUC is **0.75**, indicating reasonably good discrimination between positive and negative items.

.

8. Limitations

Our project faced several challenges stemming from the structure and quality of the Amazon book reviews dataset. A major issue was the significant class imbalance in review scores, as most ratings were overwhelmingly positive—typically 4 or 5 stars. This skewed distribution biased our models toward the majority class and reduced the effectiveness of standard classification techniques. In addition, metadata quality was inconsistent; genre information was either missing or poorly labeled, which hindered our ability to conduct genre-specific analysis. Topic modeling with LDA further posed challenges due to its bag-of-words assumption and the unrealistic expectation of topic independence, resulting in topics that were occasionally incoherent or redundant, especially given the short and informal nature of many reviews.

We also encountered computational limitations due to running all experiments on the free tier of Google Colab. While accessible, this environment restricted our available RAM, processing power, and session duration. These constraints severely impacted our ability to run more compute-intensive models such as SVMs, which took over 20 minutes on a reduced dataset. Hyperparameter tuning and cross-validation were especially problematic, often causing the Colab environment to crash or timeout, leading us to avoid it completely.

9. Conclusion

Text mining uncovers rich insight into reader sentiment and book preferences. Positive reviews reflect emotional attachment and perceived value, while negative ones often express dissatisfaction with quality or expectations. Book features such as genre and author strongly influence ratings, while price does not. Though classification and recommendation models perform reasonably well, further refinement like deep learning could enhance predictive power.

From a business perspective, these insights have clear applications. Publishers can leverage topic modeling results to identify which thematic elements resonate with audiences across genres, informing editorial decisions and marketing strategies. E-commerce platforms like Amazon can improve book discovery by integrating sentiment-driven recommendation systems, especially those that emphasize emotional language in five-star reviews. Additionally, understanding that price is not a significant driver of sentiment or score suggests that businesses should prioritize content quality and presentation—such as beautiful editions or engaging storytelling—over pricing strategies. By combining metadata with review analysis, companies can more accurately target customer preferences, personalize recommendations, and improve overall reader satisfaction.

References

1. Amazon Books Review Dataset. Kaggle. Retrieved from `https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews`

2. LDA Article from GeeksforGeeks. Retrieved from `https://www.geeksforgeeks.org/topic-modeling-using-latent-dirichlet-allocation-lda/`