Role of Readability and Sentiment in Good Reads Reviews

Samip Devkota, Brian Wolf

September 2024

1 Abstract

Does the Reading Level of book reviews on Goodreads operate as an appropriate measure for recommending books to users and does it relate to the types of books users read? Can the analysis of Reading Level be used to find users with similar reading preferences? In this paper, we aim to use various text analysis methods to investigate the Reading Level and word choice of book reviews on Goodreads to answer these questions. For our research, we look at two datasets to compare genre specific reviews with all genre reviews that both contain Goodreads book review text tied to a user ID, book ID, and other characteristics. Our approach combines algorithmic computations and statistical methods, including sentiment analysis, clustering, and sampling methods to uncover similarities between the review-text and book preferences.

Initial results from this study show little success when exploring the correlation between Reading Level and book reading preferences. In our extended research, we include the Sentiment Level of the text in our analysis to attempt to improve performance. When adding in Sentiment Level, there are improved results but nothing of significance. We can conclude from this research that the Reading Level of book reviews is not enough to find users with similar reading preferences and more features are required. With further development, this comparison approach could potentially inform the design of a more personalized and Reading Level-sensitive recommendation system if carefully integrated with other reader characteristics.

2 Introduction

2.1 Background

Online reviews play a pivotal role in shaping consumer preferences and behaviors. Online platforms like Goodreads provide users with a repository and space to share opinions and perspectives. In doing so, these reviews and opinions help shape consumer perspectives. These reviews not only serve as tools for recommendation but also shape the perspective of the books overall. Some reviews offer concise summaries, while others provide in-depth analysis. This dynamic raises an intriguing question: how do reviewers' linguistic choices while writing online reviews relate to the type of books they read, and what insights can be drawn about the reviewer's reading preferences?

2.2 Related Work

Currently, several published works analyze various characteristics of Goodreads reviews. In Wang et al. 2019 [8], the content and context of the evaluation are examined. More specifically, the book's popularity, the reviewer's role (ex. librarian, author, normal user), and the review's sentimentality. By looking at these variables, Wang et al. investigate the impact a book review has on other readers and concludes that sentiment and an understanding of the reviewer's role are needed to understand such impact fully. Within Fang et al. 2014 [1], the Goodreads reviews are tagged and analyzed for correlation and clustering based on text data. This investigation uses parallel coordinate views to visualize tag clustering on a flat scale and 3D correlative cluster techniques to represent reader and book clusters. The result of this study is a greater understanding of the different types of book reviews, stratified into distinct groups based on their similarities and differences. Lastly, a 2019 study,[2] investigates the linguistics and emotion present in Goodreads reviews to evaluate the role and value of user-generated reviews. Through the textual analysis of 450,000+ reviews, findings include the stability of user language and a trend of opinionated language used to sway other readers' opinions.

While each of these three articles investigates various aspects of book reviews, none focus on the Reading Level of the book review and what it might reveal about the user's preferences. In a study done on recommending books for children, [4], Pera et al. use a multi-dimensional approach to examine both the preferences and reading ability of the user. However, they emphasize the analysis of book illustrations and link the Reading Level of the user to that of the book, not other users. Pera has another article [5] that introduces a model to recommend books based on known books the user enjoys, their Reading Level, and are popular with those who have similar reading patterns. This article is closer to what we are trying to investigate but it does not analyze book reviews specifically and again doesn't compare review Reading Levels between users. Both of these articles address the Reading Level of the user but in both cases, they lack the comparison of Reading Levels between users as a core function of the recommender system.

2.3 Purpose

After a review of the current research, we believe there is a gap in understanding the Reading Level and sentiment of the review text and the extent to which it can be used to predict the reading choices of users. While there is some research on the language used in a Goodreads review, there is little that categorizes reviews singularly by the Reading Level of the text. As a result of this analysis, we hope to gain insight into these factors and develop a predictive analysis based on these variables. By doing this, we can explore the possibility of building a more accurate book recommendation model customized to a user's preferences. This could then be integrated into Goodreads' existing recommender system to offer more catered recommendations to users.

3 Methods

To address these questions, this study leverages computational techniques, including natural language processing (NLP) to uncover the relationships between review text features like Reading Level and Sentiment Level and how they relate to the books users are reading.

We will apply clustering and collaborative filtering techniques to assess the effectiveness of using the Reading Level and Sentiment Level from book reviews to identify users with shared book preferences. These methods will allow us to find users with similar book review Reading and Sentiment Levels and compare their book lists. For each method, we will first look at only the Reading Level and then transition into incorporating the Sentiment Level for added complexity. If the users have books in common, we will know our methods are successful.

3.1 Data

3.1.1 Data Pre-Processing

We will use two datasets for this research to compare the results of working with reviews from all genres versus the reviews from one genre. The dataset we are using for reviews from all genres consists of 15 million book reviews, 2 million books, and 465,000 users. Our dataset for genre specific reviews consists of 3.4 million book reviews and 258,000 books all from the genre: Fantasy and Paranormal. To manage the large size of the data we break it down into chunks which consist of a varying number of rows. By doing this, we have a small dataset (one chunk) and a large dataset (the whole dataset). When doing simple tests, we operate on the small dataset in chunks typically consisting of the first 10,000 rows. When doing more serious explorations of algorithmic performance, chunks will consist of 100,000 rows or more. Final results and evaluation of the algorithm are performed on 500,000 rows. Since our algorithms were so computationally expensive, this was the maximum number of rows we could operate on at once. Since users have several reviews on different books, 100,000 rows do not mean 100,000 users. On average, a user will have roughly 2 reviews so 100,000 rows equates to roughly 50,000 users. The pandas library was used to read the JSON file using the pandas.read.json() method.

3.2 Text-based Features

3.2.1 Reading Level

When approaching the calculation of a text's Reading Level, we found two options, Flesch Reading Ease and Flesch-Kincaid [6]. The United States Department of Defense uses Flesch Reading Ease to standardize its documents. In contrast, Flesch-Kincaid Grade Level is primarily used in U.S. school systems by teachers and parents to determine the grade level of books and texts. Within the code, we use the "textstat" library which has the Flesch Reading Ease and Flesch Reading Grade functions that return the appropriate scores.

The Flesch Reading Ease measure has an output ranging from 0 to 100. A higher score indicates higher readability, while a lower score indicates lower readability and that the text is more complex. Texts with a score of 60-70 are considered easily understandable by 13- to 15-year-old students and is considered the standardized range for public use texts. Unlike the Flesch-Kincaid Grade Level, which is tied to education grade levels, this metric provides an overall "ease of reading" assessment.

The Flesch Reading Ease formula is given by:

$$FleschReadingEase = 206.835 - 1.015 \times \left(\frac{totalwords}{totalsentences}\right) - 84.6 \times \left(\frac{totalsyllables}{totalwords}\right)$$

The Flesch-Kincaid Grade metric is specifically designed to assess the readability of a text by providing an estimate of the education level (in terms of U.S. school grades) required to understand the material. This metric lies on a scale from 0-18 with 18 being academic papers and 8 being 8th-grade student material. Using the Flesch-Kincaid Grade Level helps to evaluate the complexity of the reviews in terms of the educational background required to fully understand them. In the context of Goodreads reviews, it can offer insights into the kind of readers that a book attracts and whether it's more suitable for younger or older readers.

The Flesh-Kincaid Grade is given by:

$$Flesch - KincaidGradeLevel = 0.39 \times \left(\frac{totalwords}{totalsentences}\right) + 11.8 \times \left(\frac{totalsyllables}{totalwords}\right) - 15.59$$

While both metrics measure readability, they offer different benefits. The Flesch-Kincaid Grade Level directly correlates to the U.S. school grade level, which makes it easier to interpret in educational terms. On the other hand, the Flesch Reading Ease score provides a more neutral and unaffiliated measure of complexity that is not tied to a school system.

For our methods and analysis, we will use the Flesch Reading Ease score as the primary metric for calculating Reading Level. This choice is motivated by the metrics' ability to capture a wide range of text complexities on a standardized scale from 0 to 100, making it applicable to diverse types of texts, including reviews. Its straightforward interpretation—where higher scores indicate greater readability and lower scores suggest more complexity—provides a universal measure not tied to specific educational systems or grade levels. Additionally, it would be difficult to combine the scores into a composite because they are on different scales. By focusing on the Reading Ease score, we aim to maintain consistency and ensure that the metric remains accessible and comparable across different datasets. From this point forward, references to "Reading Level" will specifically pertain to the Flesch Reading Ease score, unless otherwise noted.

3.2.2 Sentiment Level

TextBlob

One popular natural language analysis tool used to gauge the sentiment of text is the TextBlob library, which allows for a calculation of polarity and subjectivity [7]. The polarity score is given within the range [-1.0, 1.0] where -1 is strongly negative and 1 is strongly positive. The subjectivity score is given within the range [0, 1] where 0 is very objective and 1 is very subjective. For our analysis, we only used the polarity score.

Example:

```
testimonial = TextBlob("Textblob is amazingly simple to use. What great fun!")
testimonial.sentiment
Sentiment(polarity=0.39166666666666666, subjectivity=0.4357142857142857)
testimonial.sentiment.polarity
0.39166666666666666
```

NLTK VADER

Initially, we used the TextBlob library for sentiment analysis, which provided a straightforward calculation of polarity (ranging from -1 to 1). However, we transitioned to using the VADER sentiment analyzer from NLTK because the text under analysis consisted of review content which often includes nuanced expressions such as slang, intensifiers, and contextual cues like punctuation or capitalization [3]. VADER is specifically designed to handle short, opinionated texts with these nuances making it better suited and more accurate for analyzing our book reviews. Its ability to break down sentiment into positive, negative, neutral, and compound scores also provides a more comprehensive understanding of the text's sentiment compared to the simpler approach offered by TextBlob. This shift allowed for a more robust and context-aware sentiment analysis of the review texts.

When using VADER, there are four different scores that are returned: positive sentiment, neutral sentiment, negative sentiment, and compound score. The positive and negative sentiment scores both range from [0, 1] and evaluate the extent to which a text expresses positive or negative sentiment, respectively. The neutral score also ranges from [0, 1] but measures the amount to which a text is objective or fact based. The compound score is a normalized score that combines the positive, negative, and neutral sentiment scores, providing an overall sentiment score. The two primary scores we will be utilizing in our analysis are the neutral and compound scores to examine the objectivity and polarity, respectively. For the rest of the paper, ”Sentiment Level” will refer to our strategic usage of these scores.

For further context, here are some example texts and the output from the function:

```
# Example texts
texts = {
    "text1": "I love this Book! It’s absolutely amazing.",
    "text2": "I hate this Book! It’s absolutely terrible.",
    "text3": "This book was published in 2003"
}
```

The results of running these example texts through the `NLTK` sentiment analysis function:

- **Text1**:
  - `Compound Score - Polarity` = 0.826
  - `Neutral Score - Objectivity` = 0.316
- **Text2**:
  - `Compound Score - Polarity` = -0.811
  - `Neutral Score - Objectivity` = 0.351
- **Text3**:
  - `Compound Score - Polarity` = 0.0
  - `Neutral Score - Objectivity` = 1.0

The results demonstrate VADER’s ability to effectively and accurately analyze nuanced, opinionated texts like book reviews. In Text1 (*”I love this Book! It’s absolutely amazing.”*), the high positive polarity (0.826) and low objectivity (0.316) reflect the strong subjective sentiment, driven by emotional words, intensifiers, and punctuation. Similarly, Text2 (*”I hate this Book! It’s absolutely terrible.”*) shows a highly negative polarity (-0.811) and low objectivity (0.351), capturing the text’s strong negative tone. In contrast, Text3 (*”This book was published in 2003.”*) has a neutral polarity (0.0) and high objectivity (1.0), accurately reflecting its factual, emotionless nature. Compared to TextBlob, which uses a simpler approach, VADER accounts for contextual cues like exclamation marks, capitalization, and intensifiers, making it better suited for the nuanced sentiment found in review data. These results highlight VADER's capacity to distinguish subjective opinions from objective statements while providing detailed sentiment analysis. In our results section, we will include the results from using TextBlob as a comparison of performance against VADER.

3.2.3 Grouping and Aggregation

Within the dataset, each user has several reviews. In grouping by user ID, we can operate on all the reviews of one user at a time and find their average Reading Level and Sentiment Level. We can then use this average score to establish an accurate representation of each user to compare with others. To do this, we used the pandas "df.groupby()" function which returns a DataFrameGroupBy object. This DataFrameGroupBy object is a collection of data that is grouped by certain criteria and can be iterated over. By iterating over each user ID group, we calculated the average Reading Level and Sentiment Level score for each user.

3.3 Preliminary Analysis

As an initial exploration of the data, we visualized the distribution of ratings and the relationship between average Reading Level and average ratings of the reviews given by the users. This was done as to figure out if there is any initial co-relation between the average Reading Level of the review and the average rating provided by the user. The following figures summarize key findings:

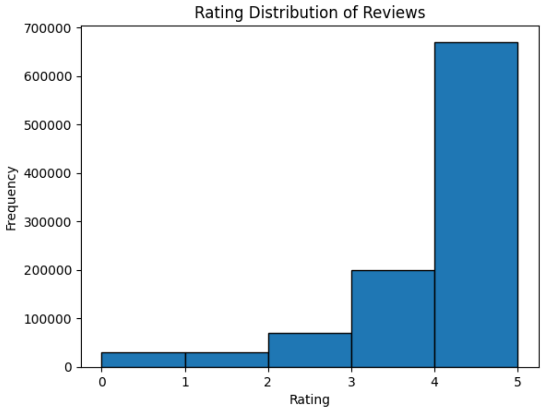


Figure 1: Rating Distribution of One Million Goodreads Reviews

Here the histogram demonstrated that the reviews are skewed to the right with a heavy majority of users rating the books higher. The average rating for books in the dataset was predominantly high, with most reviews falling within the 3 - 5 range with an average rating of 3.7 and a median of 4.0. This suggests that books receive a generally good review on Goodreads.

The scatter plot below shows the relationship between the average Reading Level of a user's reviews and the average rating they gave. Each point represents a user, with the x-axis representing the average rating and the y-axis representing the average reading ease score. This plot serves as an exploratory step to see if there is any visible correlation between review Reading Level and user ratings.

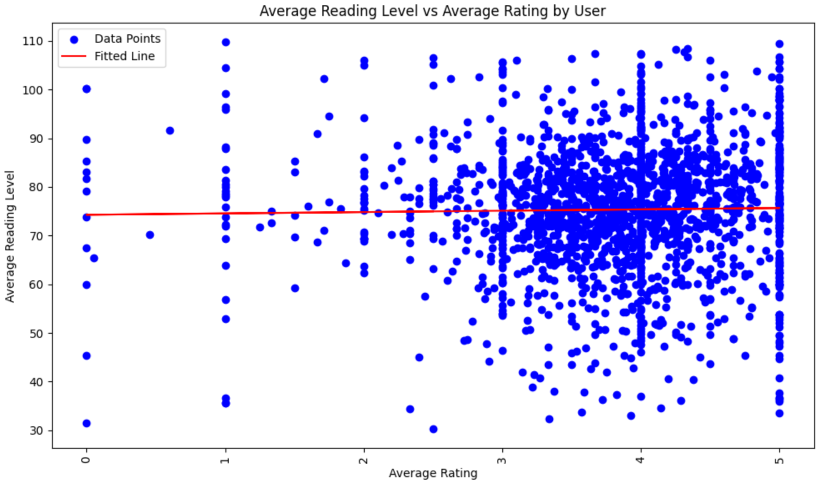


Figure 2: Average Reading Level of the user plotted against their average ratings provided

Key Observations:

- **Lack of Strong Correlation:** There is no obvious linear relationship between the average Reading Level and average rating. The points are widely scattered across the range of ratings (from 1 to 5), suggesting that the complexity of a review (in terms of Reading Level) does not strongly influence the rating a user assigns to a book. The R-squared value for the linear regression fitted for the Average Reading Level of the review and the Average Rating provided by the user is 0.

- **Cluster around Middle Ratings and Average Reading Levels:** A large concentration of users have average ratings between **3 and 5**. Similarly, the average Reading Levels for most users fall between **60 and 90** on the Flesch Reading Level scale, indicating that many users write moderately simple-to-understand reviews. The score range of 60 to 90 corresponds to texts that are generally considered fairly readable, with complexity akin to high school-level to grade school reading.

- **Outliers:** Some users consistently give very low ratings (close to 1) or very high ratings (close to 5) but their Reading Levels vary significantly, ranging from 40 (more complex texts) to 100 (very simple texts). This suggests that both highly critical and highly positive reviewers do not necessarily write reviews that are either overly simplistic or complex.

Interpretation: The plot suggests that the **readability** or complexity of a review does not have a direct influence on how users rate the books. Whether users write more sophisticated or simple reviews, the ratings they provide appear to be influenced by other factors, such as personal preferences or the book content, rather than the complexity of their writing.

3.4 Extended Analysis

For our extended analysis of the Goodreads reviews, we executed two different methods with two approaches for each. Our methods were K-Nearest Neighbor and Clustering. For each method, our first approach looked only at the Reading Level, and our second approach looked at both the Reading Level and Sentiment Level.

3.4.1 Method 1: Nearest Neighbor

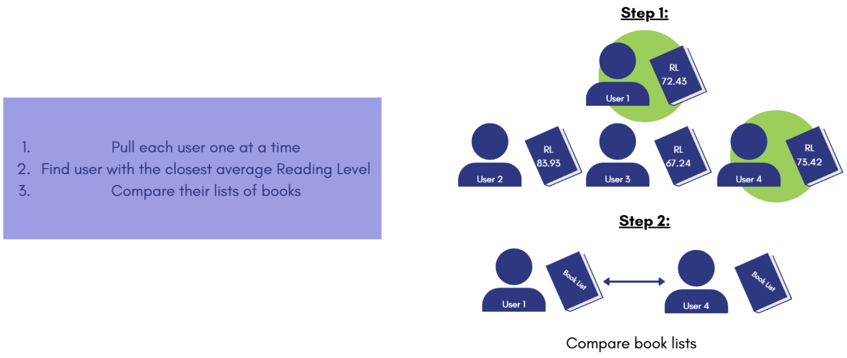
For our first method, we took each user (user 1) in our dataset and looked to find the closest other user (other user) based on average Reading Level and, eventually, average Sentiment Level. For this method, when calculating Sentiment Level, we are only taking into account the text's compound score. After finding the other user, we compared the book lists of user 1 and the other user to see if they had books in common. If they did, we considered that trial a success and moved on to the next user in the dataset.

3.4.2 Nearest Neighbor Approach 1: Reading Level

For our first approach, we only looked at the average Reading Level. To do this, we looked at all the reviews of one user simultaneously by grouping the user reviews by user ID. By then applying our Flesch Reading Ease function on each grouping of users, we found the average Reading Level for each user. With these average scores and their respective user ID aggregated in a list, we converted the list into a dataframe so it could be merged with our original dataset. We used a left-outer merge to add the average Reading Level column into our original dataset. With this new column of ”Average Reading Level”, we could start comparing users on a Reading Level basis.

Our next step was to find users with similar Reading Level and examine their book lists for overlap. Our initial user will be called User 1 and the other users will be call Other Users. We achieved this comparison we looped through each user one at a time to find their nearest neighbor in average Reading Level. We used an initial filter to trim down the number of Other Users we searched through by only looking at Other Users within a Reading Level of +- 0.5. After our initial filter, we found the difference between each Other User and User

1's average Reading Level. We then called `.idxmin()` on that list to find the minimum value and consequently, nearest neighbor. We pulled the user ID associated with that Other User and used it to pull the IDs of all the books they had read into a set. When comparing the sets of book lists of User 1 and Other User, we used a boolean to indicate if they had books in common or not. If they had books in common, we incremented a count and if they didn't, we started over with the next User 1.



1. Pull each user one at a time
2. Find user with the closest average Reading Level
3. Compare their lists of books

Figure 3: Illustration of Comparing User Book Lists: Reading Level

3.4.3 Nearest Neighbor Approach 2: Reading Level and Sentiment Level

In our second approach, we aimed to examine if adding features to our search would improve it's performance. To do this we added in the average Sentiment Level scores of user reviews in the same way we added in the average Reading Level scores. We used grouping to look at all the reviews of one user and applied our Sentiment Level function. With these scores and their related user ID in a list, we converted it to a dataframe and used a left outer merge to add a new column of "Average Sentiment Level" to our original dataset.

From this point on we used similar code to our first approach but used average Sentiment Level to act as the secondary method of selection after our initial filter using average Reading Level. After finding all the Other Users with an average Reading Level +- 0.5 of our User 1, we used the average Sentiment Level to select which user to use for comparison. Then after calculating the difference between each Other User and User 1's average Reading Level, we called `.idxmin()` to find the nearest neighbor. Similarly to Approach 1, we then pulled the book lists of each user and flagged if they had books in common or not.

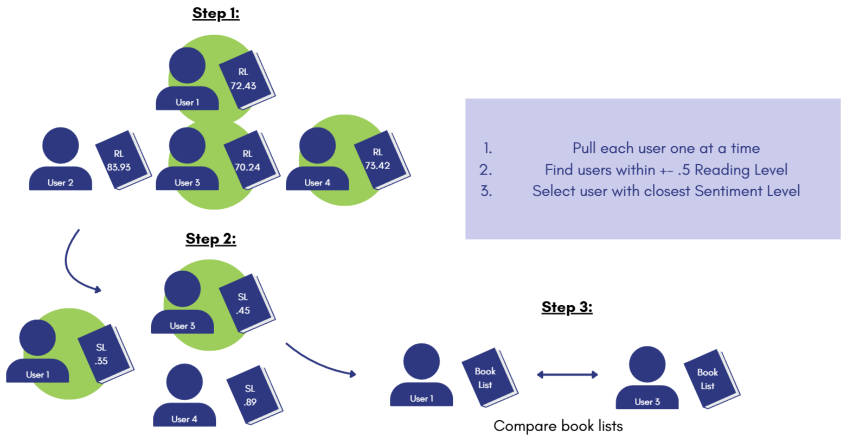


Figure 4: Illustration of Comparing User Book Lists: Reading Level and Sentiment Level

3.4.4 Method 2: Clustering

For our second method, we explored two clustering approaches to evaluate user similarity based on their Goodreads reviews. The primary goal was to determine whether users with similar Reading Levels and sentiments had a higher likelihood of sharing common books. Below, we describe the detailed steps of each approach, including data processing, clustering techniques, and similarity calculations.

3.4.5 Clustering Approach 1: Single Feature - Reading Level

Our first approach uses Reading Level as the sole feature to cluster users. This approach aimed to assess whether users with similar Reading Levels share common book preferences.

Reading Level Calculation: For each review, we calculated the Flesch Reading Ease score, which estimates the readability of the text.

User-Level Aggregation: We grouped reviews by `user_id` and calculated the average Reading Level score across all reviews for each user. This resulted in an average Reading Level score for each user.

Clustering: Using the user average Reading Level scores, we performed K-means clustering to group users with similar Reading Levels. The optimal number of clusters, k, was determined using the Elbow Method.

The scatter plot below demonstrates how users were clustered based on their average reading ease scores.

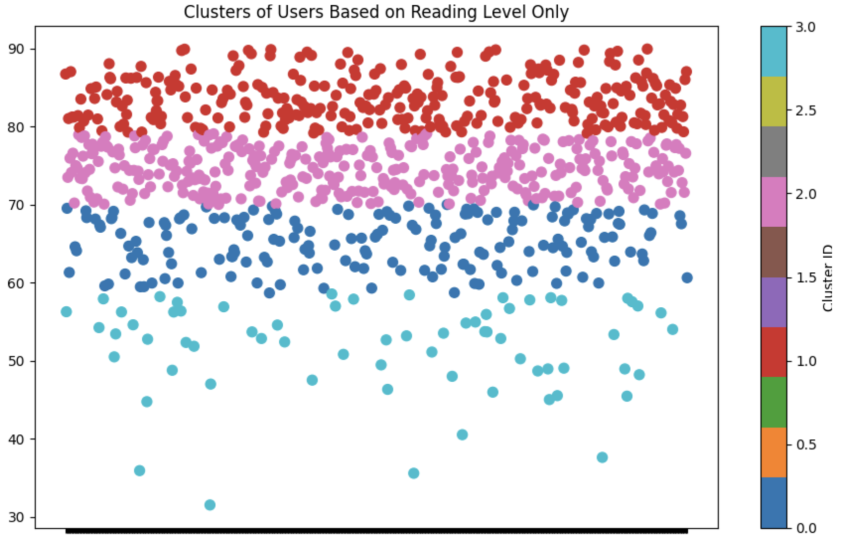


Figure 5: Users clustered according to their average Reading Levels.

Similarity Evaluation: One user was randomly sampled from a random cluster. We calculated the percentage of books they shared with other users in the same cluster. Sets of book IDs for each user were generated to facilitate comparison. This random sampling was repeated 1000 times to accurately capture what is the average percent of books a user might have in common with the rest of the cluster.

3.4.6 Clustering Approach 2: Multiple Features - Reading Level and Sentiment Level)

Given the limitations of a single-feature approach, we executed a more complex approach that incorporated the sentiment of the text in addition to reading Level. In contrast to our first method which only looked at the compound score, our multi-feature clustering considered both the compound score and the neutral score. This multi-feature approach allowed us to examine if adding sentiment of the reviews as a feature would improve the ability to find users with shared book preferences.

Sentiment Polarity Calculation: In addition to Reading Level, we calculated each review's polarity and objectivity using NLTK, which measures the degree of positive or negative sentiment within the text and how objective the statement is.

User-Level Aggregation: Similar to the first approach, we grouped reviews by `user_id` and computed both the average Reading Level, Polarity, and objectivity for each user

Multi-Feature Clustering: Using the average Reading Level and Sentiment Polarity scores, we applied K-means clustering to form groups based on both features. The optimal number of clusters was again determined using the Elbow Method, and each user was assigned to a cluster.

The scatter plot below demonstrates the plot of users plotted based on their average Reading Level, review text objectivity, and review text polarity.

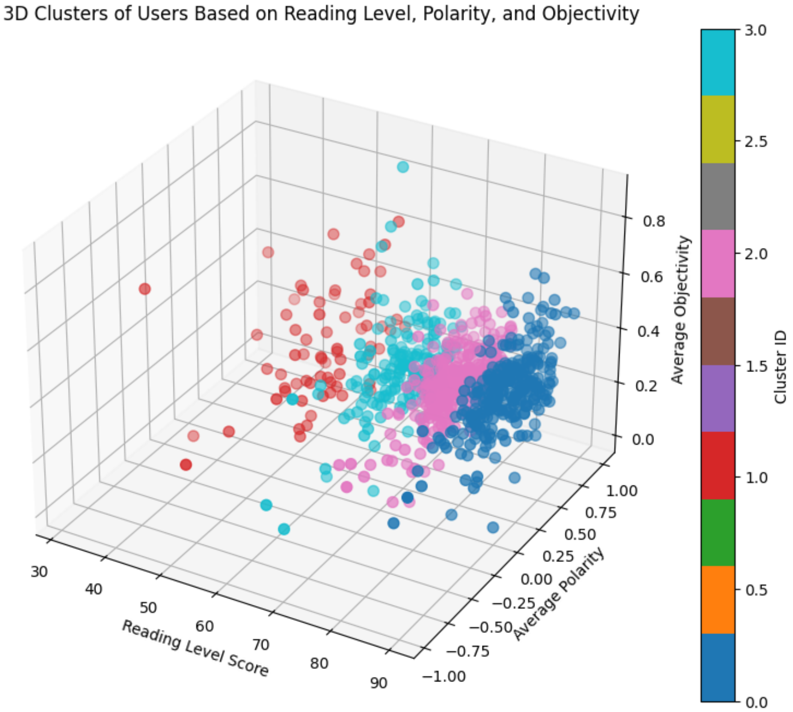


Figure 6: Clustering on Average Reading Level, Polarity, and Objectivity

Similarity Evaluation with Multiple Filters: Similar to the clusters based on average Reading Level. A cluster was chosen at random and a user within the cluster was chosen at random and compared with other users to find the set of similar books.

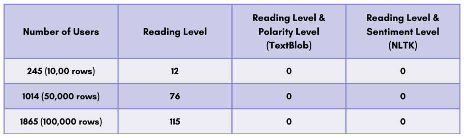
4 Results

4.1 Method 1: Nearest Neighbor Results

The main goal of this rather simplistic method was to determine with authority whether Reading Level could be used to find users with similar reading tastes. When applying our algorithm, we operated on our ”All Genre” dataset and our ”Genre Specific” dataset. We also used both TextBlob and NLTK to compare their results. Our method of evaluating the success of this method consisted of calculating the number of successful comparisons between the book lists of two users. A successful comparison was when there were identical books in each book list.

4.1.1 Nearest Neighbor Results: All Genre

The results from this algorithm don’t show a lot of promise and it seems like a poor way to find similar users who may share books in common. Since this was a one-dimensional approach, we had low expectations for finding any significant number of users with books in common for a recommendation system.



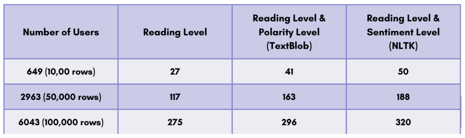
Number of Users	Reading Level	Reading Level & Polarity Level (TextBlob)	Reading Level & Sentiment Level (NLTK)
245 (10,00 rows)	12	0	0
1014 (50,000 rows)	76	0	0
1865 (100,000 rows)	115	0	0

Figure 7: All Genre Results

These numbers were a little surprising as we did not expect to see consistent zeroes. However, these results served to confirm our expectations of Reading Level being a poor method of finding users with similar reading preferences.

4.1.2 Nearest Neighbor Results: Genre Specific

These results were a lot better than our All Genre dataset and we even saw trends of improvement when incorporating and refining our features. While the numbers are still low, the improvement shows us that we were on the right track and adding even more layers and features might cause increased improvement. In addition, the improved performance in Genre Specific data shows us that Reading Level might only work on more narrow subsets of users and could potentially be improved on more refined data.



Number of Users	Reading Level	Reading Level & Polarity Level (TextBlob)	Reading Level & Sentiment Level (NLTK)
649 (10,00 rows)	27	41	50
2963 (50,000 rows)	117	163	188
6043 (100,000 rows)	275	296	320

Figure 8: Genre Specific Results

Future tests should be done on other genres to see if this improved performance is consistent across genre specific datasets.

4.1.3 Nearest Neighbor Results: Significance

The results of our Nearest Neighbor algorithm show that Reading Level alone is not enough to identify users with similar reading preferences. Our results also indicate that there is potential improvement to be made when using Reading Level in this way if the dataset is made more specific or if Reading Level is integrated with more features. However, too niche a dataset will require another method of validation as the users will all have the same book lists.

4.2 Method 2: Clustering Results

The construction and analysis of the single-feature and multi-feature clusters were as follows.

4.2.1 Clustering Results: Single Feature

Reading Level-Based Clustering (100,000 rows):

- **Average Percentage of Common Books**: 46.06%
- **Median Percentage of Common Books**: 43.81%

This measure indicates that, on average, 46.06% of the books reviewed by a random user in the cluster are also reviewed by other users within the same cluster. Similarly, the median percentage of shared books between users is 43.81%. This indicates the overlap in book reviews among users within a single-feature cluster.

4.2.2 Clustering Results: Multi-Feature

Multi-Feature Based Clustering (100,000 rows):

- **Average Percentage of Common Books**: 39.87%
- **Median Percentage of Common Books**: 37.08%

These metrics show that, on average, 39.87% of the books reviewed by a random user in the cluster are also reviewed by other users within the same cluster. The median percentage of overlap is slightly lower at 37.08%. This indicates a significant, though not overwhelming, level of shared interests in book reviews among users within the cluster.

4.2.3 Random Sampling

For further analysis and to test how well the clusters were performing an additional analysis of random sampling was added. Where a user was chosen at random from the dataset and the set of books the user reviewed was compared to the set of books for the whole dataset and an average percentage of common books was calculated. These were the results:

- **Average Percentage of Common Books**: 49.41%
- **Median Percentage of Common Books**: 50%

These results indicate that, on average, nearly half of the books reviewed by a randomly chosen user are also reviewed by others across the entire dataset, with the median percentage being exactly 50%. This provides a baseline for evaluating how well the clusters capture meaningful overlaps in book review patterns.

4.2.4 Clustering Results: Significance

Interestingly, the random sampling baseline showed higher percentages of common books compared to both clustering methods. This indicates that the clusters while providing some level of segmentation, did not outperform random groupings in terms of maximizing common book reviews. The higher baseline percentages suggest that users in the dataset inherently have substantial overlap in book preferences, which clustering methods could not significantly enhance.

The results suggest that adding sentiment and rating as features to the clustering process did not substantially impact the similarity in terms of common books among users. This could imply that **Reading Level** and **sentiment features** do not contribute significantly to identifying shared book preferences in this data set. Furthermore, understand bettervides a larger and more diverse sample, which increases the likelihood that users have books in common, leading to a higher baseline value compared to clustering methods.

5 Discussion

As we reflect on the results of our investigation, it is clear that there is room for improvement in both our methodology and analysis to better understand the impact of Reading Level and other features on reader preferences. Our research and findings are the beginnings of potential future research on using the Reading Level of book reviews to recommend books to users. There is plenty more investigation that can and should be done to further clarify when and how Reading Level should be used in book recommendations.

The findings in this paper allude to the strengths and weakness of using Reading Levels in different ways. When simply using it to tie two users together, as seen in our nearest neighbor method, it is largely unsuccessful and likely worse than selecting two users at random. However, as seen in our clustering method, when using Reading Level to build out a profile of a user, it is more successful. Based on these findings, we can confidently say that Reading Level should be used alongside other features to build a user profile, which can then be compared to other profiles.

However, when additional variables are considered, as in multi-feature clustering, the clustering performance slightly declines, highlighting the complexity of user preferences and the potential for feature selection to influence results. While the combination of variables did not outperform the baseline, the noticeable difference in results across methods underscores the possibility that an optimal combination of features could yield stronger predictive power for identifying shared book preferences among users.

For example, in our earlier analysis using Reading Level combined with sentiment polarity, the inclusion of an additional variable positively influenced the percentage of common books. This indicates that certain feature combinations may enhance the alignment between cluster members' preferences. These findings suggest that further research is needed to identify and evaluate alternative features or feature sets that could better capture the nuances of user preferences and improve clustering outcomes.

In summary, while our initial exploration highlights some limitations, it also reveals opportunities for refining our methods and deepening our analysis. Future investigations should prioritize exploring advanced clustering techniques and testing additional combinations of features to better predict book preferences and enhance the utility of user clusters.

5.1 Impact

We hoped this research would be a step towards learning how book review Reading Level might be incorporated into existing book recommendation systems such as Goodreads. We can confidently say that Reading Level by itself is not enough to accurately recommend books and should be carefully combined with other variables to build out a user profile. Lastly, we hope this research inspires further investigation into the Reading Level of book reviews.

5.2 Limitations

The limitations of our research include the ways in which we measured the Reading Level and Sentiment Level of the text. For a more comprehensive understanding of the text, better algorithms should be used and tested to improve the accuracy of the average scores for users. For example, we used the NLTK VADER package which may not be the best algorithm designed specifically for Goodreads reviews.

Another limitation is how poorly book reviews act as a true indicator of a reader's Reading Level. For example, a book review is something that users may not take seriously and write in a manner that is not reflective of their ability to write and read. The degree to which this varies from user to user is another variable that we are unable to control.

When operating on our datasets, further improvement could possibly be made upon the implementation of a test, train, validate split of the dataset. We did not include this as a part of our data pre-processing and also only operated on a maximum of 500,000 rows which may have also hampered our results.

Lastly, we make a large assumption that since users have books in common, they will also have similar reading preferences. For example, even though two users have both read popular books such as 'Harry Potter', they might have different preferences when it comes to more niche books or topics. This was helped when we looked at the Genre Specific dataset but was still an assumption.

References

[1] Shiaofen Fang, Lanfang Miao, and Eric Lin. "Visualization and clustering of online book reviews". In: *2014 International Conference on Information Visualization Theory and Applications (IVAPP)*. IEEE. 2014, pp. 187–194.

[2] Lala Hajibayova. "Investigation of Goodreads' reviews: Kakutanied, deceived or simply honest?" In: *Journal of Documentation* 75.3 (2019), pp. 612–626.

[3] Natural Language Toolkit (NLTK). *NLTK :: nltk.sentiment.vader*. `https://www.nltk.org/_modules/nltk/sentiment/vader.html`. Retrieved December 8, 2024. n.d.

[4] Maria Soledad Pera and Yiu-Kai Ng. "Analyzing Book-Related Features to Recommend Books for Emergent Readers". In: *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. HT '15. Guzelyurt, Northern Cyprus: Association for Computing Machinery, 2015, pp. 221–230. ISBN: 9781450333955. DOI: `10.1145/2700171.2791037`. URL: `https://doi.org/10.1145/2700171.2791037`.

[5] Maria Soledad Pera and Yiu-Kai Ng. "What to read next? making personalized book recommendations for K-12 users". In: *Proceedings of the 7th ACM Conference on Recommender Systems*. RecSys '13. Hong Kong, China: Association for Computing Machinery, 2013, pp. 113–120. ISBN: 9781450324090. DOI: `10.1145/2507157.2507181`. URL: `https://doi.org/10.1145/2507157.2507181`.

[6] Readable. *Flesch Reading Ease and the Flesch Kincaid Grade Level*. `https://readable.com/readability/flesch-reading-ease-flesch-kincaid-grade-level/`. Retrieved December 8, 2024. n.d.

[7] TextBlob. *TextBlob: Simplified Text Processing—TextBlob 0.18.0.post0 documentation*. `https://textblob.readthedocs.io/en/dev/`. Retrieved December 8, 2024. n.d.

[8] Kai Wang, Xiaojuan Liu, and Yutong Han. "Exploring Goodreads reviews for book impact assessment". In: *Journal of Informetrics* 13.3 (2019), pp. 874–886.