“ReadMeNext”: A Book Recommendation System Through User-Data Insights

Ahmad Diab
ahmad.diab@pitt.edu
University of Pittsburgh

Mohammad Hajjaj
moh33@pitt.edu
University of Pittsburgh

ABSTRACT

Technological advancements have revolutionized human life, offering benefits such as internet connectivity, global access to diverse content, and unprecedented opportunities for engagement. However, this abundance of information, coupled with the fast pace of modern life, presents a pressing challenge: how to effectively select what to consume amidst the deluge of options available. In this project, we address the longstanding issue of book recommendation within the context of the contemporary era, characterized by the proliferation of Large Language Models (LLMs). While existing research has primarily focused on leveraging metadata such as book attributes, publisher information, and author profiles, we harness the power of LLMs —specifically, RoBERTa Large Model — and incorporate insights derived from user-generated content (i.e. review texts), to enhance the recommendation process. Our novel system, "ReadMeNext," employs a suite of machine learning algorithms to achieve commendable performance, as observed by an F1 score of 0.83, comparable to established models in the literature. Moreover, our approach holds promise for further refinement through the integration of additional features and methodologies.

1 INTRODUCTION

Book recommendation systems present a perennial challenge with multifaceted implications. While they hold the potential to elevate the popularity and success of certain titles, they also wield the power to either amplify or stifle the visibility and reception of others, thereby shaping literary landscapes and influencing reading behaviors. Extensive literature has scrutinized this conundrum from diverse angles, some even attempted to question the validity of the user-generated attributes in the prediction of books' quality and whether they are fake or reflect real opinions from people who read the book[2]. For instance, Wijaya et al.[7] delved into the feasibility of leveraging authors' attributes, such as their identity and publication history, while Maity et al.[6] delved into publishers' data to gauge a book's potential popularity and merit. However, these approaches are not without their biases; they often perpetuate a 'rich-get-richer' dynamic, wherein established authors and publishers enjoy preferential treatment due to past successes, while newcomers struggle to gain traction in the absence of historical data to buoy their works. Alternatively, some approaches have focused on intrinsic book attributes, as exemplified by Maghari et al.[5], which analyzed factors like title, page count, and abstract content to assess quality. Yet, such methods overlook the invaluable insights offered by literary critics and domain experts, whose nuanced evaluations could significantly augment prediction accuracy. Building upon this foundation, Kousha et al.[4] underscored the pivotal role of user-generated feedback— encompassing ratings, reviews, and commentary— in shaping individuals' book preferences. However, prior studies predominantly fixated on numerical ratings and review volumes, neglecting the rich semantic cues embedded within review texts. In this study, we embark on an exploration to ascertain whether the textual content of reviews can serve as a reliable indicator of a book's quality. The subsequent sections of this paper are structured as follows: Section 2 introduces the dataset and delineates its features, while Section 3 elucidates our feature selection methodology, with a particular focus on harnessing insights from user-generated content. Section 4 delineates our methodology and work pipeline, leading to the presentation of results in Section 5. We subsequently engage in a critical discourse on the limitations of our approach (Section 6), outline avenues for future research (Section 7), and conclude with a summary of key findings and implications (Section 8).

2 DATASET

Our dataset originates from GoodReads[1], a widely utilized platform renowned for fostering a vibrant community of readers, enthusiasts, and literary professionals who engage in sharing experiences, opinions, and recommendations. Specifically sourced from a dataset made publicly available on Kaggle[3], our dataset comprises 900,000 records, each meticulously documenting a single review authored by a user (denoted as 'ui') for a particular book (denoted as 'bj'). The data collection spans from 2008 until 2019. Each record encompasses a wealth of information, including unique identifiers for both the book and the user, the review text itself, metrics such as the number of comments and votes garnered by the review, the user's rating, and timestamps indicating the commencement and completion dates of the user's reading journey for the given book. This comprehensive dataset serves as a valuable resource for our investigation and is openly accessible via the Kaggle platform.

3 FEATURE SELECTION

Our research focuses on the utilization of user-generated data as a cornerstone of our methodology. To this end, we incorporated the review text, represented by its contextual embedding derived from the RoBERTa model (further elaborated in Section 4). Additionally, we included several key features to enrich our analysis. These

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/ToBeDefined

features encompassed the average rating assigned to each book, the count of user-generated reviews for each book on the platform, and metrics reflecting user engagement, such as the total number of comments and votes on a book and its corresponding average. These latter features serve as proxies for the level of engagement and resonance a particular book elicits within the reading community. By encompassing both textual content and quantitative measures of reader interaction, our feature selection strategy aims to capture the multifaceted dimensions of user engagement and opinion in the recommended reading materials.

4 METHODOLOGY

The raw dataset underwent a series of transformations to adapt its format to the requirements of our research objectives. In this section, we elaborate on the systematic process of data transformation and cleansing undertaken to render the dataset suitable for analysis. Additionally, we present a comprehensive overview of the models employed to address the task at hand, elucidating their respective methodologies and functionalities.

Data Format: The original dataset comprised individual records representing single reviews for each book. However, the primary objective of our research necessitated an aggregation of book-related information to evaluate the quality of each book holistically. To achieve this, the dataset underwent a transformation wherein information pertaining to each book was consolidated into a single record, organized based on the unique book ID attribute. For instance, if a book has n reviews in the original dataset, its corresponding record in the transformed dataset encompassed aggregated metrics such as the total number of reviews, cumulative comments received across all reviews, and a concatenated compilation of review texts separated by a newline character. The overall rating assigned to each book was computed as the average rating derived from all individual reviews. Attributes associated with users (e.g., user ID) were subsequently excluded from the transformed dataset, as they did not contribute to the final system's objectives. Consequently, the final dataset comprised 25,000 records, each representing a distinct book entity. The distribution of the rating in the transformed dataset can be observed in figure 1.

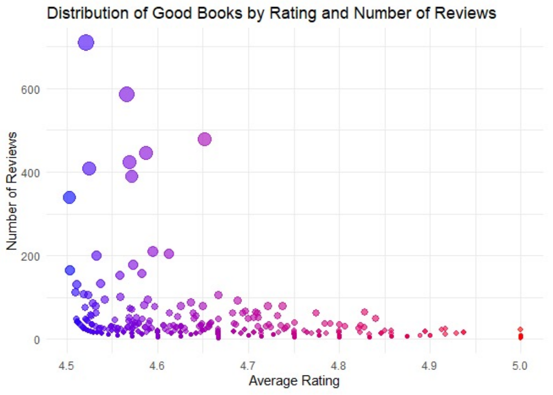

Figure 1: The distribution of the final rating of books aggregated from the reviews in the original dataset

Text Representation: Central to our methodology is the utilization of textual content extracted from reviews, which serves as the foundation of our analysis. To represent the concatenated text reviews for each book, we employ RoBERTa-large model. Renowned for its performance in capturing the nuanced semantics of natural language, RoBERTa has demonstrated efficacy across diverse applications, making it an ideal choice for our study. Leveraging this model, we obtain vector representations of the text, each comprising 1024 dimensions. This computation-intensive task was executed on an Nvidia GeForce GTX Titan GPU card equipped with 12GB of memory. The task took approximately 12 hours to reach completion.

Filtration: Records containing missing information were filtered out from the original dataset to ensure data integrity. Subsequently, following the aggregation process, any book with fewer than 5 reviews, comments, or votes was deemed insufficient for meaningful analysis and consequently excluded from the study. Additionally, to maintain the quality and relevance of textual reviews, any review containing less than 10 characters was discarded from the dataset, as such brevity typically precludes substantive insights into book quality. Cumulatively, this filtration process resulted in the removal of 3.5% of the dataset, while preserving the integrity and comprehensiveness of the remaining data for subsequent analysis.

Book Label: Defining the criteria for assessing a book's quality poses a significant challenge in this research domain, given the inherent subjectivity and diverse perspectives within the literature. Drawing inspiration from existing works, we adopt a pragmatic approach by evaluating a book's quality primarily based on its overall rating. To establish a benchmark, we first identified the 10 most popular genres on the GoodReads website, subsequently manually scrutinizing the list of recommended books for each genre, each comprising 100 titles. Notably, all books featured in these lists boasted ratings of 3.7 or higher, serving as our threshold for defining a "good" book. Accordingly, we categorize a book as "good" if its cumulative rating in our dataset exceeds this threshold. The resulting distribution of labels, as depicted in Figure 2, reflects the efficacy of our defined criteria. It's noteworthy that the labeling of books exhibits a balanced distribution to some extent.

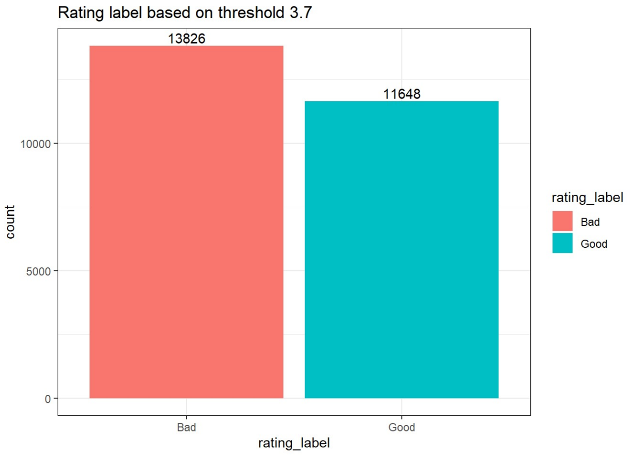

Figure 2: Label distribution of books in the final aggregated dataset, based on rating threshold of 3.7

Models and Training: In this task, we employed a variety of common machine learning models, including Generalized Linear Model (glmnet), Naive Bayes, K nearest neighbor (KNN), Decision Tree, Random Forest, and Support Vector Machine (SVM). To optimize model performance, each underwent a rigorous grid search to identify the optimal hyperparameters. Furthermore, we conducted each training iteration using 5-fold cross-validation to ensure the robustness and reliability of results. Given the substantial size of the dataset, we present the outcomes of these models in the subsequent section, wherein all models were trained and evaluated on a representative 20% subset of the data. Performance evaluation was conducted using standard error metrics, encompassing Accuracy, Precision, Recall, F1 score, and Area Under the ROC Curve (AUC), providing comprehensive insights into the efficacy and comparative performance of each model.

5 RESULTS

The culmination of our experiments is encapsulated in Table 1, presenting a comprehensive overview of model performance. Notably, glmnet emerged as the top performer across all measured metrics, boasting an F1 score of 0.8303. Following closely behind, SVM exhibited strong performance with an F1 score of 0.8225, aligning with established literature highlighting the efficacy of these models in similar problem domains. Random Forest demonstrated superior performance compared to Decision Tree, showcasing a marginal 0.1 increase in F1 score. Conversely, KNN exhibited the poorest performance across all metrics, suggesting that the variance in textual reviews, compounded by other features selected in our work, may contribute to its sub-optimal performance. Furthermore, consistent patterns were observed across Accuracy, Precision, and Recall metrics, mirroring the trends observed in F1 scores. A comparison of F1 performances can be observed in Figure 3.

To test the performance improvement gained from the textual semantics of the reviews, we ran the models under the same conditions without the text embeddings. The results can be seen in Table ??. A significant drop can be observed across all models. The final results of our experiments can be seen in Table 1. The glmnet outperformed other models in all measured performance metrics with F1 score of 0.8303. Followed closely by SVM with F1 score of 0.8225, these models were achieved notably higher results than the rest, which is similar to previous works finding that these models were the best in this type of problem. Random Forest outperformed Decision Tree by 0.1 in F1 score. KNN, on the other hand, was the worst performance in all metrics, we hypothesize that the variance in text reviews accompanied by other features are throwing off the performance of this model. Accuracy, Precision, and Recall have consistent patterns similar to F1.

Table 1: Models’ performance. The best results in each column is in bold

Model	Accuracy	Precision	Recall	F1
glmnet	**0.8197**	**0.8512**	**0.8103**	**0.8303**
KNN	0.5444	0.5856	0.5573	0.5711
naive Bayes	0.7168	0.7441	0.7313	0.7376
SVM	0.8102	0.8374	0.8082	0.8225
Decision Tree	0.6655	0.6971	0.6816	0.6892
Random Forest	0.7648	0.7809	0.7891	0.7851

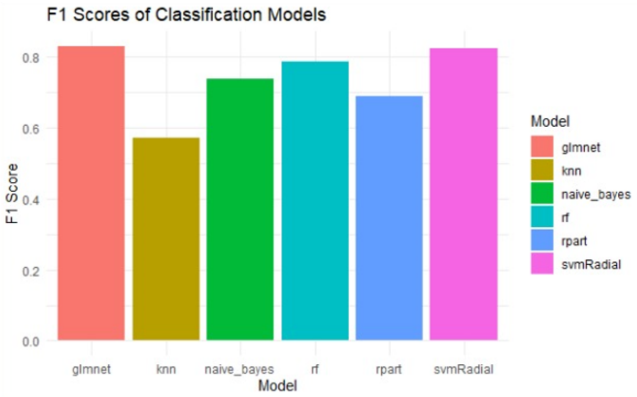

Figure 3: Performance of models on the dataset represented by F1 score

Table 2: Models’ performance without text representation. A significant drop is observed across all models

Model	Accuracy	Precision	Recall	F1
glmnet	**0.5916**	0.6005	**0.7575**	**0.6699**
KNN	0.5447	0.5899	0.5509	0.5697
naive Bayes	0.5914	0.6012	0.7522	0.6682
SVM	0.5837	**0.6175**	0.6288	0.6231
Decision Tree	0.5851	0.6132	0.6546	0.6332
Random Forest	0.5499	0.5855	0.6079	0.5969

6 LIMITATIONS

While our study contributes valuable insights, it is essential to acknowledge its limitations. Firstly, the determination of a book’s quality inherently invites subjectivity and ambiguity. While prior research often relied on established best-seller lists, we opted to utilize recommendation lists sourced from GoodReads, potentially introducing variability in labeling decisions for the same book. Moreover, assessing the quality of literature is inherently subjective, varying significantly among individuals. While we endeavored to mitigate this challenge by leveraging the collective opinion of the community, biases, and diverse preferences may still influence outcomes.

Secondly, our dataset was sourced from Kaggle, a reputable platform hosting a diverse array of datasets across different domains. However, the lack of comprehensive documentation regarding data collection methodologies and filtration procedures poses a notable limitation. The absence of such information raises concerns regarding the transparency and reliability of the dataset, thereby potentially compromising the integrity of our analyses.

Finally, comparing results across various studies in this domain proves challenging due to discrepancies in dataset composition and methodological approaches. As a result, drawing direct comparisons and generalizing findings becomes problematic. To address this limitation, we intend to extend our research by incorporating additional features suggested by prior studies, such as book and author information. By adopting a more comprehensive feature set, we aim to offer a better-nuanced understanding of the factors influencing book recommendations and their relative significance.

7 FUTURE WORK

In our current investigation, we have emphasized the significance of user-generated data, particularly reviews, in refining the efficacy of our classifiers. Through an ablation study, we demonstrated the tangible impact of incorporating text embeddings on classifier performance. However, it's imperative to acknowledge that prior research has highlighted the potential utility of additional features, including book attributes and sentiment analysis of reviews, in augmenting classification accuracy. Our model, "ReadMeNext," possesses the flexibility to expand its feature set and encompass these valuable dimensions. Thus, a logical progression for our future work involves the incorporation of supplementary features to enrich the recommendation process further.

Moreover, the predominance of textual representation in our current framework—comprising 1024 features compared to a mere 6 for other dimensions—prompts us to explore feature reduction techniques, such as Principle Component Analysis (PCA). By effectively reducing the dimensionality of our feature space while preserving relevant information, PCA holds the potential to streamline model computation and enhance efficiency without sacrificing predictive performance. This avenue represents a promising avenue for optimizing our model's computational efficiency while retaining its predictive accuracy.

Finally, to evaluate the robustness and generalizability of our approach, we intend to subject our best-performing model to external validation by applying it to datasets curated by other researchers in the field. By benchmarking our model against diverse datasets, we can gauge its performance across varied contexts and ascertain its capacity for generalization beyond our specific experimental conditions.

8 CONCLUSION

In this study, we introduce a novel approach to the book recommendation challenge, leveraging user-generated data, including comment and review statistics, alongside textual content from reviews. Our findings reveal that the glmnet model emerges as the top performer, achieving an F1 score of 0.83, closely followed by SVM with an F1 score of 0.823. These results align with existing literature, reinforcing the efficacy of these models within the realm of book recommendation systems. Furthermore, our ablation study underscores the pivotal role of textual context in enhancing model performance, demonstrating a notable 23.9% improvement in F1 score results when incorporating textual embeddings. This underscores the potential of textual reviews to enrich feature sets and bolster model performance, particularly when integrated with other attributes identified in prior research. However, it's crucial to approach the deployment of such models with caution, recognizing their capacity to influence the popularity dynamics of specific books. Therefore, future analyses of model outcomes should diligently scrutinize potential biases that may emerge, ensuring that these systems operate equitably and ethically. Conclusively, our study opens new horizons in the realm of book recommendation systems, illuminating the transformative potential inherent in user-generated data and textual insights. As we persist in refining our methodologies and integrating a diverse array of features, we lay the groundwork for the emergence of more personalized and effective recommendation mechanisms. Through these efforts, we aspire to elevate and enrich the reading experiences of individuals on a global scale, fostering a deeper connection between readers and the literary works that resonate with them.

ACKNOWLEDGMENTS

We wish to express our gratitude to Professor Yuru Lin for her invaluable support and guidance throughout our Data Mining course. Her expertise and encouragement were instrumental in shaping this project. Thank you, Professor Lin, for your unwavering commitment to our academic journey.

REFERENCES

[1] GoodReads 2015. *Goodreads, a social cataloging website for books.* Retrieved April 22, 2024 from https://www.goodreads.com/

[2] Lala Hajibayova. 2019. Investigation of Goodreads' reviews: Kakutanied, deceived or simply honest? *Journal of Documentation* 75, 3 (2019), 612–626.

[3] kaggle 2017. *Goodreads Books Reviews.* Retrieved April 22, 2024 from https://www.kaggle.com/competitions/goodreads-books-reviews-290312/data

[4] Kayvan Kousha, Mike Thelwall, and Mahshid Abdoli. 2017. Goodreads reviews to assess the wider impacts of books. *Journal of the Association for Information Science and Technology* 68, 8 (2017), 2004–2016.

[5] Alaa Mazen Maghari, Iman Ali Al-Najjar, Said Jamil Al-Laqtah, and Samy S Abu-Naser. 2020. Books' rating prediction using just neural network. (2020).

[6] Suman Kalyan Maity, Abhishek Panigrahi, and Animesh Mukherjee. 2017. Book reading behavior on goodreads can predict the amazon best sellers. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017.* 451–454.

[7] Rachel Anastasia Wijaya, Stephanie Staniswinata, Maria Clarin, Nunung Nurul Qomariyah, and Ida Bagus Kerthyayana Manuaba. 2023. Prediction Model of Book Popularity from Goodreads "To Read" and "Worst" Books. In *2023 10th International Conference on ICT for Smart Society (ICISS).* IEEE, 1–7.