Beyond the Cover: Leveraging Text and Features for Advanced Book Recommendations

Francisco Wulf T.
Pontificia Universidad Católica
Santiago, Chile
francisco.wulf@uc.cl

José Tomás Valdivia C.
Pontificia Universidad Católica
Santiago, Chile
jtvaldivia@uc.cl

Vicente Thomas L.
Pontificia Universidad Católica
Santiago, Chile
vicente.thomas@uc.cl

Abstract

Recommender systems have a great potential to increase the number of readers by helping individuals discover books that match their preferences and interests. Given the vast popularity of books worldwide and the significant amounts of data available in this domain, we study the effectiveness of item recommendation in the context of book selection. Utilizing data from kaggle, we analyzed the results of our models, with the aim of improving the performance of predictions obtained by them. Based on the DeepFM recommender system and BERT for processing textual information, we sought to predict the next book a user might prefer, based on their past reading behaviors. Multiple data sources with different attributes, ranging from metadata about books to user ratings and descriptions, were processed. The textual attributes were transformed into embeddings using BERT, while additional metadata attributes were directly integrated into the DeepFM framework. These embeddings and features were then combined to generate a robust representation of the data.Additionally, a voting and ensemble system was implemented to enhance the prediction results by combining the outputs of the individual models. This approach allowed us to account for varying strengths of each model, improving the overall recommendation accuracy. Finally, the experimental processes, including the ensemble methodology and the variations in data handling, are described in detail, along with the obtained results.

Keywords

Book recommendation, Personalized recommendation, Recommender systems, DeepFM, BERT, Ensemble models, Machine learning, Natural language processing, Textual similarity, Metadata-based recommendation.

1 Introduction

In the current digital age, although we have access to millions of books, this abundance can become a disadvantage when readers face an overwhelming number of choices. Beyond bestsellers or classics, finding books that match specific tastes can be daunting, often leading many to abandon the search before discovering something that truly captivates them. This issue is exacerbated by the declining reading habits among younger generations, who prefer quick audiovisual content, such as social media and streaming platforms. According to the National Survey on Reading and Writing [5], only 33.4% of Chileans read in their free time, and in the U.S., 38% of young adults did not read a book in the past year [10].

This trend negatively impacts skills such as concentration and deep reflection, which are essential in daily life [8]. In the face of these challenges, there is a growing need for tools that facilitate personalized book searches. A suitable recommendation system could reduce the overload of options and enhance the reading experience by relying on users' previous preferences and comparing them with similar profiles. Moreover, integrating demographic factors, such as age, could help capture the attention of younger generations, reintroducing them to reading in an engaging way.

This project aims to simplify and personalize the reading experience, helping those who want to reconnect with their reading habit as well as those who haven't developed it yet.

2 State of Art

This section provides an overview of existing methods relevant to this study. In the wild, there exist various datasets with book ratings information. The most famous consists of bookcorpus [11], with more than seven thousand books, goodbooks-10k [12], with more than ten thousand books, and goodreads dataset [6], with more than 50 thousand books and more than two million interactions. All these datasets only consist of rating information. In contrast with the Book-crossing dataset [1] which contains more information, as we will see in the next section.

In terms of methods that can be used in the task of book recommendations, there exists systems that implement Funk's SVD and BPR. We propose a multimodal system composed of a Deep FM module and a BERT module, that performs better than both Funk's SVD and BPR.

2.1 FunkSVD

Funk's SVD method utilizes matrix factorization for rating prediction, and thus, can be used in the task of book recommendations. This method consists of performing the Single Value Decomposition of the interaction matrix. By doing Stocastic Gradient Descent, this method can minimize the Sum of Squared Error.

RecSys'24, December 2024, Santiago, CHILE

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2.2 BPR

Bayesian Personalized Ranking, although it is a collaborative filtering method, it relies on a different paradigm. BPR is a learning-to-rank method that uses the BPR loss function and implicit feedback for performing the recommendation task.

2.3 DeepFM

DeepFM is a method which uses a Factorization Machines based Neural Network aimed for the task of maximizing the Click-Through Rate [3] on recommender systems. Utilizing a special FM layer within the hidden layers, Deep FM is capable of processing dense embeddings and extracting one-hot encoded sparse features.

2.4 Bert

BERT is a bidirectional transformer pretrained using a combination of masked language modeling objective and next sentence prediction on a large corpus comprising the Toronto Book Corpus and Wikipedia [4]. This model tokenizes the sentences into absolute position embeddings, state in which the pairwise distances or other similarity method can be applied. BERT has already been leveraged in conversational recommendation systems for books [9] but has not been used for tokenization of the book's summary.

3 Objectives

The main objective of this project is to develop a recommendation system that suggests books based on the user's preferences. In doing so, we aim to attract new readers and encourage prioritizing reading over other forms of entertainment that do not provide the same personal value.

A key focus of the project is to identify a recommendation approach that surpasses the Most-Popular baseline in at least one evaluation metric. To achieve this, the performance of individual models such as DeepFM and BERT will be assessed in isolation, analyzing their ability to deliver relevant book recommendations. Furthermore, the project will explore the effectiveness of combining these models into an ensemble, comparing their individual results with the performance improvements achieved through integration.

4 Dataset

The dataset used in this study corresponds to the Book-Crossing Dataset (BX) [1]. This dataset consists of 278858 anonymized users with geographical (country, city and state) and age information, 271379 publications or books, and almost 1150000 interactions or ratings. BX also contains a summary for each book, three images of the cover of the book, in different sizes, language information and categories. Figure 1 shows the age distribution of the users. Worth noting than 300 thousand users have 35 years. This is probably the default age when there is no information.

As Figure 2 shows, the BX dataset has both implicit and explicit information: the ratings table has ratings both from 1 through 10, as well as ratings marked as zero. In our investigation, we worked with only explicit information, this is, only the ratings going from one through ten. Figure 3 shows the distribution of the explicit ratings.

Finally, the books were published in a certain year, and this information is included in the BX dataset. Figure 4 shows the top 25 years with the most books published. It appears that 1998 through 2003 were the top years for publishers.

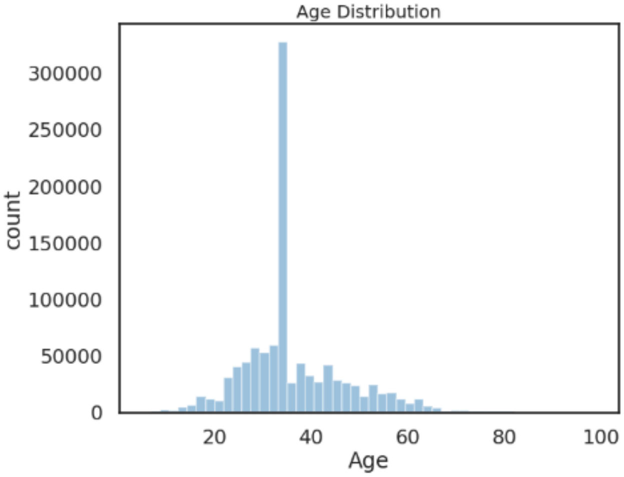

Figure 1: Age distribution of the BX dataset

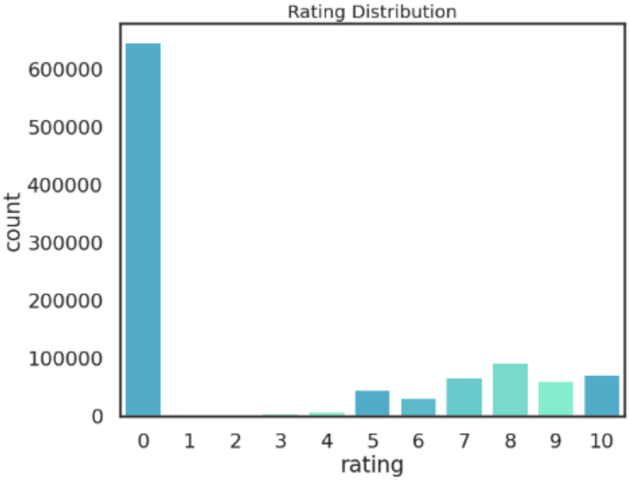

Figure 2: Rating distribution of the BX dataset

5 Experimentation and methodology

In this section, we describe the experimental setup and methodologies used to evaluate the performance of our book recommendation system. The primary goal of the experimentation is to test the individual components (BERT, DeepFM, and other baseline models) along with their integration into an ensemble framework. We also compare the proposed models against established baselines to assess their effectiveness in terms of MAP, Recall, and other key metrics. The experimentation is divided into several stages, including data preprocessing, model training, ensemble strategies, and evaluation.

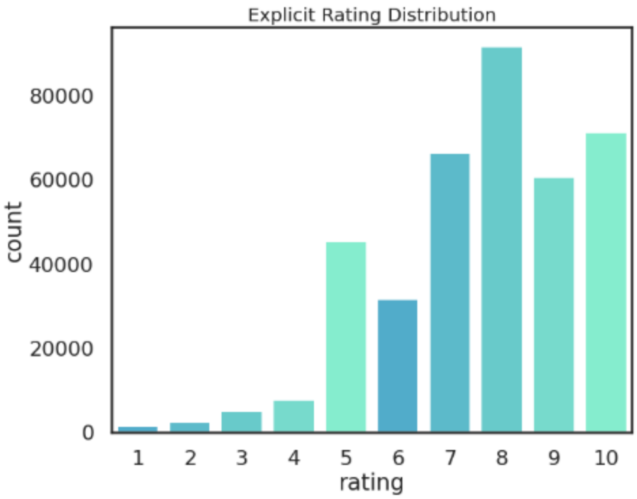

Figure 3: Explicit rating distribution of the BX dataset

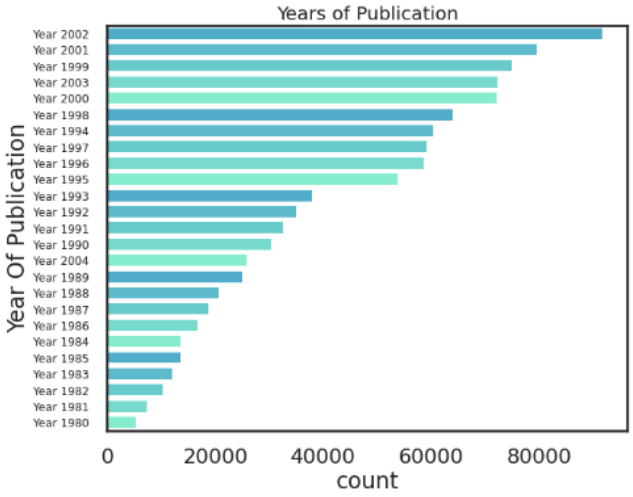

Figure 4: Top 25 years with the most publications in the BX dataset

5.1 Data preprocessing

The dataset used for the experimentation comprises information about books, including metadata such as title, genre, author, and publication year, alongside user-specific data like demographics and previous interactions. Before training, we applied preprocessing techniques such as removing duplicate entries of books, handling missing values by imputing and discarding records with incomplete data, using only a 5% of the total dataset and finally splitting the dataset into 80% training and 20% testing.

5.2 Model Training

Each component of the recommendation system was trained separately using the preprocessed data:

- **BERT for textual similarity of the summary:** BERT embeddings were obtained using the HuggingFace Transformers library. Then the cosine similarity between these embeddings was used to measure the semantic closeness between books, allowing recommendations based on textual similarity [2].
- **DeepFM for metadata:** this algorithm was implemented using the LibRecommender library [7], a Python library designed for building state of the art recommendation systems. The DeepFM model was trained using metadata features such as user age, user country and book publication year. The model's ability to capture both linear and non-linear interactions between features ensured personalized recommendations.
- **Baseline Models:** several baseline approaches were implemented, such as MostPopular, RandomBaseRecommender, FunkSVD, and Bayesian Personalized Ranking (BPR). These models served as benchmarks for evaluating the performance of our proposed methods.

5.3 Ensemble Strategies

To enhance the system's overall performance, we experimented with two ensemble strategies:

- **Weighted Average Ensemble:** it combines the predictions from BERT, DeepFM and MostPopular by assigning predefined weights to each model based on their individual precision during testing. The weights were fine-tunned using Grid Search Cross-Validation to find the optimal weights that showed best precision.
- **Voting Mechanism:** implemented a voting-based approach where each model "voted" for a set of recommended books. Votes were weighted according to the relative importance of each model, and the book with the highest cumulative score was selected as final recommendation. This approach aimed to leverage the strengths of all models while mitigating individual weaknesses. There was also fine-tunning for the weights of this model, using Grid Search Cross-Validation.

5.4 Evaluation

The performance of each individual model and ensemble strategies was evaluated using MAP, Recall, NDCG, Diversity and Novelty. The results of Bert, DeepFM, and the ensemble strategies were compared against MostPopular, RandomBaseRecommender, FunkSVD and BPR.

Table 1: Experimentation Results

Model	MAP	NDCG@10	Recall	Diversity	Novelty	Time (s)
MostPopular	0.01068	0.0138	0.007	0.3	6.735	2.8
Funksvd	0	0	0	0.431	9.376	1.36
Random	0.0002	0.0005	0	0.466	9.653	1.89
BPR	0.001	0.0005	0.0004	0.473	9.644	2.01
DeepFM	0.0079	0.0104	0.0106	N/A	N/A	5.23
BERT	0.0048	0.0074	0.0055	0.258	9.618	13.43
Ensemble	0.0077	0.0116	0.0071	0.3647	7.292	20.27
Voting	0.0108	0.0142	0.0088	0.3625	7.419	21.16

6 Parameter analysis

A fine-tuning process was conducted to determine the optimal weights for the ensemble and voting mechanisms. This was achieved by performing a Grid Search Cross Validation. Specifically, tests were conducted using combinations of BERT, MostPopular, and DeepFM, denoted as [BERT, MostPopular, DeepFM]. In addition, various weight values were tested, ranging from 1 to 8, ensuring that the sum of the weights always equaled 10. This approach aimed to identify the optimal configuration of these models to maximize overall recommendation performance.

The weights obtained for the simple ensemble model and the voting model were 1, 7, 2 and 2, 7, 1 for BERT, MostPopular, and DeepFM, respectively. This highlights the dominant influence of the MostPopular model in both approaches, as it consistently received the highest weight. This dominance aligns with its strong baseline performance across precision-related metrics. However, the inclusion of BERT and DeepFM ensures a more balanced contribution by capturing nuanced patterns and generating less conventional recommendations.

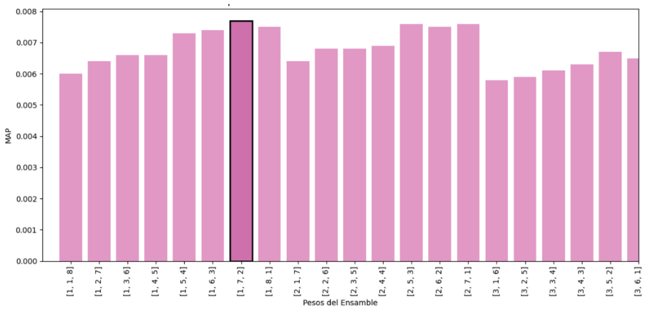

Figure 5: MAP Comparison with different weights in Ensemble

7 Results

All experiments utilized the metrics MAP, RECALL@10, NDCG@10, Diversity, and Novelty. The outcomes are summarized in Table 1.

With these results, we can see that the voting model slightly outperformed the MostPopular model in terms of MAP, while offering significantly higher novelty and diversity. These improvements suggest that the Voting model is more effective at providing personalized recommendations that not only cover a wider variety of items but also present less common, potentially more interesting choices to the user.

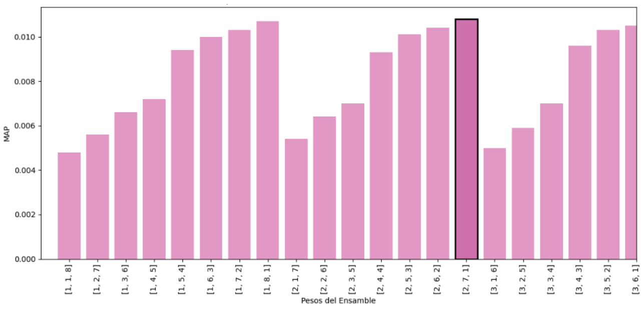

Figure 6: MAP Comparison with different weights in Votation

On the other hand, models like Funksvd, Random and BPR showed lower performance across the majority of the metrics, with particularly poor results in MAP and NDCG@10. However, Funksvd demonstrated a relatively high diversity score, which indicates that, while it may not suggest the most relevant items, it does provide a broader range of recommendations.

Moreover, the DeepFM and BERT models showed moderate performance in terms of MAP and recall. The DeepFM model, in particular, exhibited the best performance in recall, although it was less competitive in terms of novelty. BERT, while providing some level of novelty, required considerably more computational time, which might limit its practical application in real-time systems.

8 Conclusion and future work

The results indicate that the voting model slightly outperforms the MostPopular model. However, the difference is not statistically significant. These findings enable an effective comparison between the models, fulfilling the stated objectives.

To enhance performance, future work should focus on fine-tuning model parameters and conducting more exhaustive optimization. Additionally, we propose incorporating an ensemble approach that combines the top 100 most popular items with other models, potentially improving diversity and novelty.

Future investigations should also explore the impact of assigning zero weights during the fine-tuning process, as this was not considered in the current study. We believe this omission might have led to suboptimal results and addressing it could provide valuable insights into the ensemble's behavior.

References

[1] R. Bhatia. 2021. Book-Crossing dataset mined by Bhatia R. https://www.kaggle.com/datasets/ruchi798/bookcrossing-dataset

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*. https://arxiv.org/abs/1810.04805

[3] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction. *arXiv preprint arXiv:1703.04247* (2017). https://doi.org/10.48550/arXiv.1703.04247

[4] Huggingface. 2023. BERT: Pretrained model for natural language understanding. https://huggingface.co/docs/transformers/model_doc/bert

[5] Ipsos. 2022. Leer en Chile 2022. https://www.ipsos.com/sites/default/files/ct/publication/documents/2022-10/Leer Ipsos.

[6] Bahram Jannesarr. 2023. Goodreads book datasets 10M. https://www.kaggle.com/datasets/bahramjannesarr/goodreads-book-datasets-10m

[7] massquantity. 2023. LibRecommender: A comprehensive library for recommender systems (Version 1.5.1). https://github.com/massquantity/LibRecommender Computer software.

[8] Abdón Crisóstomo Paucar, Lidia Janeth Llacsa Puma, and Rosana A. Meleán Romero. 2024. HÁBITO DE LECTURA EN ESTUDIANTES DE EDUCACIÓN PRIMARIA. *Aula Virtual* 5, 11 (Feb. 2024), 29–43. https://doi.org/10.5281/zenodo.10464908

[9] Gustavo Penha and Claudia Hauff. 2020. What does BERT know about books, movies and music? Probing BERT for Conversational Recommendation. In *RecSys '20: Proceedings of the 14th ACM Conference on Recommender Systems*. 388–397. https://doi.org/10.1145/3383313.3412249

[10] A. Perrin. 2021. Who doesn't read books in America? https://www.pewresearch.org/short-reads/2021/09/21/who-doesnt-read-books-in-america/ Pew Research Center.

[11] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *The IEEE International Conference on Computer Vision (ICCV)*.

[12] zygmuntz. 2023. Goodbooks-10k dataset. https://github.com/zygmuntz/goodbooks-10k

Received 9 December 2024