Book Recommendation Using Hybrid Embeddings

Musa Tora

Spring 2025

Abstract

This project presents a method for recommending books based on the identification of similar characters across different literary works. A machine learning model generates personalized recommendations by analyzing character similarities within a user-provided list of previously read books. To accomplish this, contextual embeddings are generated from book titles using BERT and are concatenated with character embeddings derived from the Project Gutenberg dataset.

These combined embeddings form a unified semantic representation for each book. Cosine similarity is then used to compare these representations and rank relevant recommendations. The system outputs books with character profiles that align closely with those in the user's reading history, offering a deeper level of personalization. This work contributes to the development of semantic recommendation systems by demonstrating the effectiveness of integrating character-level and contextual embeddings to enhance recommendation accuracy and narrative relevance.

BERT is used to generate contextual embeddings from book titles, while character embeddings are sourced from Project Gutenberg. These are concatenated into a unified representation for each book.Recommendations are made using cosine similarity. The system will return books with similar character from different books.This project contributes to semantic recommendation systems and showcases the value of combining multiple embedding sources.

1 Introduction

In today's world, most book recommendation systems rely heavily on genre classification or collaborative filtering. While these methods are effective to some extent, they tend to trap readers in echo chambers—offering the same type of stories over and over. As a result, users often miss out on discovering books that are different in setting or theme, yet still resonate with them on a deeper level.

This project is driven by the idea that a reader's attachment to a book often stems not just from its genre, but from the characters themselves—their personalities, values, and emotional journeys. By focusing on character-level traits rather than surface-level metadata, the system seeks to offer users a way to explore more diverse and unexpected books while still maintaining a sense of emotional continuity.

The goal of this project is to develop a recommendation engine that identifies books with characters similar to those the user already enjoys, regardless of genre. By combining BERT-generated title embeddings with precomputed character embeddings, the system forms a hybrid semantic representation of each book. Using cosine similarity, it then recommends new books that feature characters with comparable traits, helping users branch out into new stories without losing the emotional connection they value most.

2 Related Works

2.1 Simple Embeddings and Similarity Measures

Subramanian highlights the effectiveness of basic models like Word2Vec combined with cosine similarity for generating recommendations. His work shows that even lightweight embeddings can capture meaningful semantic relationships between texts, although they may lack contextual depth.

2.2 Hybrid and Multi-Embedding Models

Javaji and Sarode propose a hybrid method that combines Sentence-BERT (SBERT) and RoBERTa to produce high-quality document embeddings. Their findings demonstrate that blending multiple embedding sources enhances recommendation accuracy. This directly supports my model, which merges BERT-generated title embeddings with character embeddings for deeper personalization.

2.3 Contextual Embedding and Semantic Understanding

BERT's layered architecture has proven valuable in various NLP tasks. Jawahar et al. show that its lower layers capture syntactic structures, while higher layers model complex semantic relationships—validating BERT's use in extracting nuanced meaning from book titles.

Zhang et al. further compare embedding methods in deep learning models and conclude that BERT outperforms traditional alternatives such as Word2Vec and GloVe, particularly in semantic tasks. These findings reinforce my decision to use contextual embeddings from BERT.

2.4 Character Embedding and Narrative Modeling

Inoue et al. introduce the Charembench dataset, a benchmark for evaluating character embeddings using Project Gutenberg novels. Although centered on Japanese fiction, the dataset underscores the importance of modeling character relationships in a vector space. My project draws from this idea to build character-aware recommendations.

Lima introduces a multi-layered embedding-based semantic retrieval system for legal documents. This architecture mirrors my design, which fuses title and character embeddings to build a rich semantic representation of books.

2.5 Applications, Gaps, and Extensions

Mansur and Hasan demonstrate the utility of semantic title embeddings derived from Wikipedia-based data. While their system focuses on structural link analysis, it affirms the effectiveness of semantic similarity models. My work extends theirs by incorporating character-level information.

Gundecha et al. combine BERT embeddings with TF-IDF for content-based recommendation using book descriptions. Their system lacks character awareness, which is the core innovation of my project.

Grootendorst presents BERTopic, a technique for clustering documents using BERT and TF-IDF. Although developed for topic modeling, it offers inspiration for future features like clustering books based on character types or shared themes.

Wang and Xu apply a fine-tuned BERT model to sentiment analysis. Their findings support my goal of fine-tuning BERT on literary corpora to enhance its ability to capture character traits and emotional nuance.

3 Results and Conclusions

Across the surveyed studies, embedding-based models consistently demonstrate strong performance in text similarity, recommendation, and semantic analysis. Subramanian confirms the value of Word2Vec for simple content-based filtering, while Javaji and Sarode show that hybrid embeddings lead to improved accuracy—reinforcing my system's multi-source embedding approach.

Jawahar et al. and Zhang et al. validate the use of BERT for capturing syntactic and semantic richness, especially in textual domains like book titles. Inoue et al.'s benchmark for character embeddings adds rigor to my use of character vectors, while Lima's multi-layered model confirms the viability of layered semantic fusion.

The use of title embeddings by Mansur and Hasan, and the BERT-TF-IDF approach of Gundecha et al. affirm the effectiveness of semantic modeling but lack the narrative depth achieved through character embedding. Grootendorst and Wang Xu suggest possible enhancements, thematic clustering and fine-tuning, that can expand the impact of this system.

Despite these advancements, a gap remains in combining title and character embeddings for book recommendation.This project directly addresses this underexplored area by integrating character-level insight with contextual title representation.

4 Opportunities for Future Work

Most current systems focus on metadata or descriptions, overlooking the role of characters in shaping reader preferences. Combining character and title embeddings offers a novel and more human-centered approach to recommendation. Several avenues exist for enhancing this work further. One promising direction involves fine-tuning BERT on literature-specific or character-rich datasets to better capture emotional tone and narrative style. Another enhancement could involve incorporating additional dimensions such as genre, plot arcs, or sentiment flow to support more nuanced similarity matching. Theme-based clustering techniques such as BERT can also be explored to group books or characters by common motifs or emotional journeys. Additionally, developing an interactive front-end—such as a Streamlit interface—could allow users to visualize the relationships between books and characters, further enriching the recommendation experience. These extensions aim to deepen personalization, improve recommendation quality, and expand the system's functionality beyond standard filtering approaches.

Preliminary Design

The proposed system will rely on two primary sources of data: BERT-generated title embeddings and character embeddings derived from the Project Gutenberg dataset. Users will begin by providing a list of book titles they have previously read. Using a pretrained BERT model accessed through the Sentence-Transformers library, the system will generate contextual sentence embeddings that capture the semantic meaning of each book title. Simultaneously, character embeddings sourced from the Project Gutenberg dataset will represent character-level features. These two types of embeddings—title and character—will be concatenated to form a unified vector for each book. The system will then compute cosine similarity between these unified vectors and those in a larger book database to identify and rank the most similar books. The final output will be a personalized list of recommended books with similar character profiles. Development will utilize tools and frameworks including Python, PyTorch or TensorFlow, Hugging Face Transformers, Sentence-Transformers, NumPy, pandas, and scikit-learn.

5 Evaluation Plan

To evaluate the effectiveness of the character-based book recommendation system, I will use both quantitative analysis and qualitative user feedback. The evaluation will begin with a curated subset of books from the Project Gutenberg dataset, which includes title and character information. I will create an evaluation set consisting of several books and simulate user input by selecting small reading histories of 3–5 titles. Quantitatively, the system will compute cosine similarity between unified embeddings (a combination of title and character vectors) and rank recommendations based on similarity scores.

For qualitative evaluation, I will conduct informal testing with a small group of users who will input a list of books they've enjoyed and rate the system's recommendations. They will also have the opportunity to provide written feedback on why they found a recommendation helpful or not.

For quantitative testing, cosine similarity scores will be computed between book vectors. I will compare the hybrid model (title + character embeddings) with a baseline (title embeddings only) to demonstrate the added value of character-level information.

6 IRB Considerations

Because this project involves collecting user feedback, I will submit an application to Earlham College's Institutional Review Board (IRB). All participation will be voluntary and anonymous, with minimal risk. Collected feedback will be used only to evaluate system effectiveness.

7 Anticipated Contributions

This project aims to deliver several key contributions to the field of book recommendation systems. At its core, it will introduce a character-driven recommendation engine that leverages vector similarity to match books based on character traits. The system will utilize cosine similarity to compare unified vectors composed of both character and title embeddings, showcasing a hybrid embedding strategy that enhances personalization by combining semantic and narrative features.

In addition to these primary contributions, the project may also explore several auxiliary features. If time allows, BERT will be fine-tuned on character descriptions to generate higher-quality embeddings tailored to literary analysis. Further, theme-based clustering methods may be implemented to group books according to shared narrative arcs or emotional tones. Lastly, an interactive interface—possibly built with Streamlit—may be developed to visualize the system's recommendations and provide an engaging user experience.

8 Analysis of Major Risks

A primary risk is that title embeddings alone may not provide accurate recommendations. In this case, I may incorporate genre or book descriptions to enhance results. Another risk is the unavailability or quality of character embeddings. As a contingency, I will fine-tune BERT on character descriptions to generate approximate vectors.

9 Special Resources

I will use the Project Gutenberg dataset, which includes character data, story content, and genre information. Required packages include: transformers, sentence-transformers, numpy, pandas, and scikit-learn.

10 References

1. Grootendorst, M. (2020). BERTopic: Neural topic modeling with class-based TF-IDF. *arXiv*. https://arxiv.org/pdf/2005.13012

2. Gundecha, S., Kumari, P., & Prasad, R. (2021). Content-based book recommendation using BERT embeddings. *arXiv*. https://arxiv.org/pdf/2103.11943

3. Inoue, N., Pethe, C., Kim, A., & Skiena, S. (2022). Learning and evaluating character representations in novels. *ACL 2022*. https://aclanthology.org/2022.findings-acl.81/

4. Javaji, S. R., & Sarode, K. (2023). Multi-BERT for embeddings in recommendation systems. *arXiv*. https://arxiv.org/abs/2308.13050

5. Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? *ACL*. https://arxiv.org/pdf/1906.04341

6. Lima, J. A. de O. (2024). Unlocking legal knowledge with multi-layered embedding-based retrieval. *arXiv*. https://arxiv.org/pdf/2411.07739

7. Mansur, F., & Hasan, A. (2023). Neural embedding for book recommendation. *Solid State Technology*. https://www.researchgate.net/publication/377748533

8. Subramanian, D. (2021). Recommendation system using word embeddings. *ResearchGate*. https://www.researchgate.net/publication/350353485

9. Wang, Y., & Xu, J. (2024). BERT model in sentiment analysis. *arXiv*. https://arxiv.org/pdf/2403.08217

10. Zhang, Y., Zhou, Y., & Zhang, S. (2021). Word embeddings for text classification. *arXiv*. https://arxiv.org/pdf/2103.11943