Integrating Embeddings, Sentiment, and Generative AI for Enhanced Product Recommendations

Yashas Acharya, Thomas Knickerbocker, Owen Ratgen
Backpropagation Nation
ratge006@umn.edu, knick073@umn.edu, achar100@umn.edu

Abstract

This study investigates integrating modern natural language processing (NLP) techniques into recommendation systems to address the cold start problem and improve user satisfaction. We evaluate four models, finding that approaches incorporating sentiment analysis and semantic reasoning achieved average star ratings of 4.44 compared to 2.50 for the baseline. Additionally, an exploratory generative model using LLMs and DALLE demonstrates potential in producing novel recommendations but faces challenges in diversity and scalability. Our findings highlight the value of hybrid approaches that combine retrieval-based and generative methods, offering practical and scalable solutions for businesses while identifying key areas for future improvement.

1 Introduction

1.1 Introduction

In the rapidly evolving landscape of e-commerce, delivering personalized product recommendations is crucial for enhancing user experience and driving sales. Traditional recommendation systems often grapple with the "cold start" problem, where limited user interaction data hampers the ability to provide relevant suggestions for new users or items (RoX818, 2024). This challenge is particularly pronounced in platforms with sparse review data, leading to suboptimal recommendations and reduced user engagement.

1.2 Background

Recommendation systems have traditionally relied on collaborative filtering and content-based methods to predict user preferences. Collaborative filtering leverages user behavior patterns, while content-based approaches utilize item attributes to make recommendations. However, both methods encounter significant challenges in cold start scenarios. Collaborative filtering struggles with new users or items due to insufficient interaction data, which is particularly relevant in our subset of the reviews dataset, where the vast majority of users have left very few reviews, and have often only reviewed a very small subset of all available products (Grčar et al., 2006). Content-based methods may lack the contextual richness needed for accurate predictions (Javed et al., 2021). However, augmenting recommender systems with RAG has been observed to be an efficient new tactic for powerful recommendations (Hou et al., 2024a).

To address these limitations, recent advancements have explored the integration of embeddings, sentiment analysis, and generative AI into recommendation systems. Embeddings transform items and user interactions into dense vector representations, capturing semantic relationships and enabling more nuanced recommendations (Mikolov et al., 2013). Sentiment analysis provides insights into user opinions, allowing systems to factor in qualitative aspects of user feedback (Dang et al., 2021). Generative AI models, such as Large Language Models (LLMs) and image generators like DALL·E, offer the potential to create personalized content and recommendations by understanding and generating human-like text and images (OpenAI, 2023).

1.3 Motivation

This project aims to enhance recommendation quality by integrating embeddings, sentiment analysis, and generative AI to address the cold start problem and improve user satisfaction (Geng et al., 2023). By leveraging embeddings, we can capture the semantic relationships between products, enabling more accurate similarity assessments. Incorporating sentiment analysis allows the system to consider the qualitative aspects of user reviews, aligning recommendations with user sentiments. Utilizing generative AI models facilitates the creation of personalized product suggestions and content, enhancing the relevance and appeal of recommendations.

The motivation behind this approach is to develop a recommendation system that not only overcomes the limitations of traditional methods in cold start scenarios but also provides a richer, more engaging user experience through personalized and contextually relevant recommendations. In a business context, this system can drive higher user engagement, increase conversion rates, and build customer loyalty, giving companies a competitive edge in the fast-paced e-commerce landscape (Ping et al., 2024). Moreover, such systems can enable businesses to better understand user preferences, paving the way for more informed product and marketing strategies.

2 Approach

2.1 Data Preparation

To build an effective recommendation system, we began with a dataset sourced from the 2023 Amazon Reviews repository (Hou et al., 2024b). This dataset was filtered to focus exclusively on the Toys and Games category. Several preprocessing steps were undertaken to ensure the dataset's quality and relevance:

- **Filtering**: Removed unused columns, discontinued products, and items with excessively long descriptions. Verified purchases were prioritized, and products containing profanity were filtered out using the `better_profanity` library and additional custom keyword searches.

- **Embedding Generation**: Generated dense vector representations for product descriptions using `jinaai/jina-embeddings-v3`, chosen for its speed and high performance on semantic text similarity tasks (Muennighoff et al., 2022).

- **Sentiment Analysis**: Applied `cardiffnlp/twitter-roberta-base-sentiment-latest` to calculate sentiment scores for product reviews (Camacho-Collados et al., 2022). Sentiment scores were normalized and weighted using the number of helpful votes per review.

- **OYT Score Calculation**: Computed the "One You Trust" (OYT) score for each product based on weighted sentiment scores and review helpfulness.

Figure: Total Pipeline Diagram

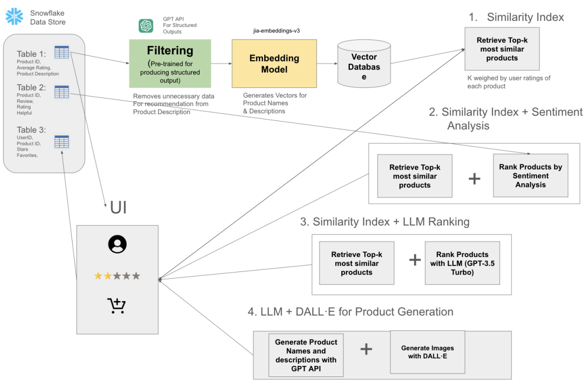
Snowflake Data Store
Table 1: Product ID, Average Rating, Product Description
Table 2: Product ID, Review, Rating Helpful
Table 3: UserID, Product ID, Stars Favorites,
GPT API For Structured Outputs
Filtering (Pre-trained for producing structured output)
Removes unnecessary data For recommendation from Product Description
jia-embeddings-v3
Embedding Model
Generates Vectors for Product Names & Descriptions
Vector Databas e
1. Similarity Index
Retrieve Top-k most similar products
K weighed by user ratings of each product
2. Similarity Index + Sentiment Analysis
Retrieve Top-k most similar products
+
Rank Products by Sentiment Analysis
UI
3. Similarity Index + LLM Ranking
Retrieve Top-k most similar products
+
Rank Products with LLM (GPT-3.5 Turbo)
4. LLM + DALL·E for Product Generation
Generate Product Names and descriptions with GPT API
+
Generate Images with DALL·E

Figure 1: Overview of Data Flow and Recommendation Models. The pipeline illustrates the integration of various models and data sources to generate personalized recommendations.

2.2 Recommendation Models

2.2.1 Model 1: Similarity Index

Model 1 forms the baseline by leveraging cosine similarity to rank products based on their semantic closeness to user-rated items. For any two product embeddings E_1 and E_2, the cosine similarity is defined as:

$$\text{Cosine Similarity}(E_1, E_2) = \frac{E_1 \cdot E_2}{|E_1||E_2|} \tag{1}$$

Products with the highest similarity scores are recommended, and user ratings influence subsequent iterations by weighting higher-rated products more heavily in the recommendation pool.

2.2.2 Model 2: Similarity Index + Sentiment Analysis

Model 2 improves upon the baseline by integrating sentiment scores into the ranking process. The ranking score is calculated as:

$$\text{Rank} = \frac{\text{Cosine Similarity} + \text{OYT Score}}{2} \tag{2}$$

This adjustment ensures that products with high-quality reviews, as indicated by sentiment analysis, are prioritized alongside semantically similar items.

2.2.3 Model 3: Similarity Index + LLM

Model 3 incorporates a Large Language Model (LLM) as an intermediary ranking layer. Products with high cosine similarity are passed to a GPT-based model, which evaluates them based on their titles and metadata to generate a refined ranking. The gpt-4o-mini model was selected for its high speed, accuracy, and low usage costs (OpenAI, 2024). This model leverages the LLM's reasoning capabilities to capture nuances that similarity scores alone might miss.

Integration Details: The gpt-4o-mini model processes product titles using prompts designed to emphasize user-relevant features. Example prompt:

> *"Rank these products for recommendation based on their similarity, relevance, and potential user interest. The original product is:* ***<Product Title>****. Consider semantic similarity, user preferences, and product attributes."*

2.2.4 Model 4: LLM + DALL E for Product Generation

Model 4 is exploratory and focuses on generating entirely new product ideas. GPT generates product descriptions based on user ratings and relevant topics, which are then visualized using DALL E. This approach addresses the cold start problem by creating products that align with user preferences without relying on existing inventory.

Integration Details: The GPT-generated content includes:

- Summarized product name and description.
- Highlighted features and target audience.
- Keywords aligned with user preferences.

The generated descriptions are fed into DALL E to create corresponding visuals, providing users with a holistic recommendation experience.

2.3 Iterative Recommendation Workflow

Recommendations are iteratively refined based on user interactions. When a user rates or likes a product, the system adjusts the weights of similar items in subsequent iterations. For example, the recommendation score for a product i is updated as:

$$\text{Updated Score}_i = \text{Base Score}_i \times (1 + \alpha \times r_u) \quad (3)$$

where α is a scaling factor, and r_u is the user's rating.

3 Experimentation

3.1 Experimental Setup

To evaluate the performance of the proposed recommendation models, we designed a controlled experimentation framework using a testing playground. The setup involves logging user interactions, generating recommendations iteratively, and analyzing results across key metrics. The experimentation process for each model is outlined as follows:

1. **Initialization**: A user ID is assigned, and an initial set of 8 recommendations is provided to the user.
2. **Feedback Collection**: User interactions, including star ratings (1-5 scale) and likes (binary), are recorded and stored in a table.
3. **Algorithm Execution**: The selected recommendation algorithm processes the stored feedback to generate a new set of recommendations.
4. **Iteration**: The process is repeated for a maximum of 3 iterations (refreshes) per user.
5. **Metric Logging**: Metrics are computed and logged for each iteration to facilitate comparative analysis.

Workflow Diagram: Figure 1 illustrates the iterative workflow from user interaction to feedback integration and recommendation generation.

3.2 Evaluation Metrics

The performance of each model is evaluated based on the following metrics:

- **Aggregate Star Ratings**: The average star ratings of recommended products over iterations. Higher averages indicate better alignment with user preferences.
- **Like Ratio**: The percentage of recommended products that users liked. This metric provides qualitative insights into user satisfaction.
- **Diversity**: A measure of the variety of categories or attributes in the recommended products. Higher diversity ensures broader exposure to different items.

Table: Table 1 summarizes these metrics across iterations and models.

3.3 Experimentation Plan for Models

3.3.1 Model 1: Similarity Index

Model 1 serves as the baseline, relying solely on cosine similarity between product embeddings to generate recommendations. Performance was measured across all primary metrics, with a focus on diversity and aggregate star ratings.

Quantitative Analysis: Model 1 achieved moderate performance, with aggregate star ratings improving incrementally from 1.49 in Trial 0 to 2.50 in Trial 3. Diversity remained consistently high, with an average of 78 unique product IDs per trial.

3.3.2 Model 2: Similarity Index + Sentiment Analysis

Model 2 integrates pre-computed OYT scores with cosine similarity to enhance recommendation quality. Metrics such as like ratio and aggregate star ratings were emphasized in evaluating the model.

Quantitative Analysis: Model 2 showed significant improvements in aggregate star ratings, rising from 2.12 in Trial 0 to 4.44 in Trial 3. Like ratios peaked at 88%, indicating strong user satisfaction. Diversity remained high, with an average of 76 unique product IDs per trial.

3.3.3 Model 3: Similarity Index + LLM

Model 3 incorporates a GPT-based intermediary to refine rankings based on semantic understanding of product titles and metadata. This model's ability to capture nuanced user preferences was analyzed, with a particular focus on engagement.

Quantitative Analysis: Model 3 delivered comparable results to Model 2, with star ratings rising from 2.05 in Trial 0 to 4.34 in Trial 3. Like ratios reached 90%, the highest among all models. Diversity was well-maintained, averaging 77 unique product IDs per trial.

3.3.4 Model 4: LLM + DALLE for Product Generation

Model 4 is exploratory, using GPT to generate product descriptions and DALLE to create corresponding visuals. This model aimed to address the cold start problem and was evaluated primarily on like ratio.

Quantitative Analysis: Model 4 showed moderate performance with star ratings improving from 1.92 in Trial 0 to 3.41 in Trial 3. Like ratios increased to 65%, while diversity was constrained to 8 unique product IDs per trial due to the generative nature.

Model	Final Avg Star Rating
Model 1	2.50
Model 2	4.44
Model 3	4.34
Model 4	3.41

Model	Final Like Ratio
Model 1	45%
Model 2	88%
Model 3	90%
Model 4	65%

Model	Diversity (Unique Products)
Model 1	78
Model 2	76
Model 3	77
Model 4	32

Table 1: Aggregate Metrics for Models

3.4 Metric Comparison

Table 1 presents the aggregate metrics logged per iteration for each model.

3.5 Graphical Comparisons

- Figure 2 compares average star ratings across models and trials.
- Figure 3 visualizes like ratios for each model over iterations.

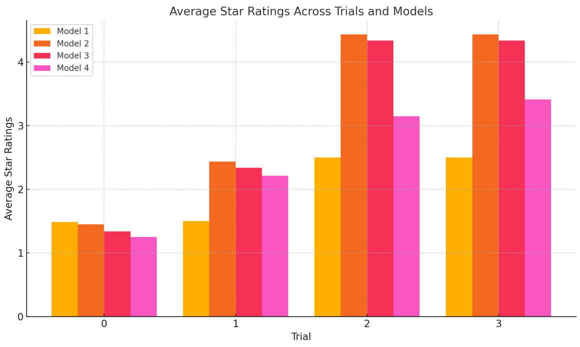

Figure 2: Average Star Ratings Across Trials and Models

4 Discussion and Limitations

4.1 Discussion

This work explores four distinct approaches to product recommendation, each leveraging state-of-the-art natural language processing (NLP) techniques. The quantitative results demonstrate the effectiveness of incorporating embeddings, sentiment anal-

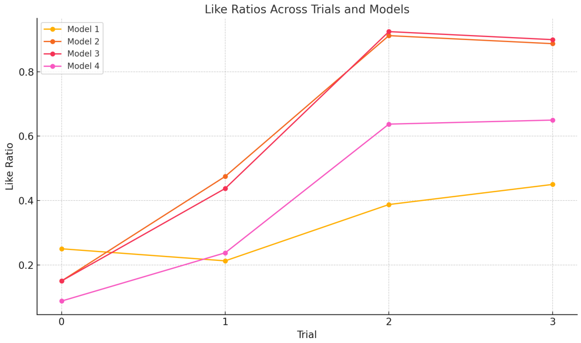

Figure 3: Like Ratios Across Trials and Models

ysis, and large language models (LLMs) in enhancing recommendation relevance and user satisfaction.

Model 2 and Model 3 emerged as the most promising approaches, with substantial improvements in average star ratings and user likes over iterations. This highlights the importance of integrating sentiment analysis and semantic reasoning into recommendation pipelines. Sentiment analysis provides critical insights into user preferences, while the semantic capabilities of LLMs offer nuanced understanding, resulting in recommendations that align closely with user expectations. The exploratory nature of Model 4 showcased the potential of generative approaches to address cold start problems, though its limited diversity underscores challenges in balancing novelty with practical utility.

Additionally, the results suggest that simple cosine similarity-based models (Model 1) are insufficient in complex recommendation scenarios, particularly when user data is sparse or unstructured. The incremental improvements in this model point to its limitations in addressing deeper contextual nuances, which more sophisticated methods like sentiment integration and LLMs successfully capture.

This study demonstrates that modern NLP methods can provide scalable, cost-efficient solutions for personalized recommendations, potentially enabling smaller e-commerce platforms to adopt sophisticated systems without extensive computational resources. By prioritizing low-cost implementations for Models 1 and 2, businesses with constrained budgets can achieve meaningful enhancements in recommendation quality.

4.2 Limitations

Despite the promising findings, this study has several limitations that warrant consideration.

Data Sparsity and Bias: The sparse nature of the dataset, particularly for reviews, posed challenges for models relying on sentiment analysis (Model 2). Products with few or no reviews were disadvantaged in their ranking, potentially biasing results towards well-reviewed items.

Generative Approaches: While Model 4 demonstrated the capability to generate novel recommendations, the lack of existing product metadata led to limited diversity. Additionally, generated products often lacked meaningful contextual alignment with user preferences compared to retrieval-based methods.

Evaluation Metrics: While metrics like aggregate star ratings and like ratios provided valuable insights, the absence of user engagement data beyond likes (e.g., click-through rates, time spent) limits a comprehensive understanding of user satisfaction.

Scalability and Computational Overheads: Although Models 1 and 2 are computationally inexpensive, Models 3 and 4 rely on LLMs, which require significant resources. This may limit their practicality for smaller businesses with constrained computational budgets.

Generalizability: The study focuses on a single dataset, which may not generalize to other domains or e-commerce verticals. Further experimentation with diverse datasets is needed to validate the robustness of the proposed approaches.

Ethical Considerations: The potential for biases in LLM-generated recommendations and the reliance on pre-computed scores highlights the importance of incorporating fairness and transparency mechanisms in future iterations.

5 Ethics and Broader Impacts

5.1 Ethical Considerations

Bias and Fairness: One of the primary ethical concerns in recommendation systems is the potential for bias in the data or algorithms. While no significant biases were observed during testing, the sentiment-driven approach in Model 2 relies heavily on pre-computed OYT scores, which can disadvantage products with sparse or no reviews. This issue could inadvertently limit the visibility of new or niche products, potentially affecting fairness in recommendations. Addressing this challenge could involve future work on synthetic review generation or adaptive weighting mechanisms to balance recommendations for underrepresented products.

Privacy: User interaction data, including star ratings and likes, is collected for experimentation purposes. While this project focuses on model architecture and performance, considerations for anonymizing or encrypting this data are critical to ensure user privacy. Future iterations of this system could incorporate privacy-preserving techniques, such as differential privacy or secure storage protocols, to align with ethical standards for data handling.

Generative Models: The use of GPT and DALL E for generating product descriptions and visuals introduces specific ethical challenges. Automated processes can occasionally produce misleading or culturally insensitive outputs, especially if the input data is ambiguous or lacks context. While no review mechanism is currently in place, future improvements could include human oversight or automated filters to ensure the quality and appropriateness of generated content. This would mitigate potential risks and enhance user trust in the system.

5.2 Broader Impacts

Positive Outcomes: This project demonstrates significant potential for improving user experiences in e-commerce platforms. By addressing the cold start problem and personalizing recommendations, the proposed system enhances user satisfaction and engagement. Lightweight models, such as Model 1 and Model 2, are particularly beneficial for smaller vendors, offering cost-effective and resource-efficient solutions. These models provide an accessible alternative for businesses with limited compute budgets or infrastructure, enabling them to compete more effectively in the e-commerce market. Additionally, generative models like GPT and DALLE offer creative opportunities for marketing and user engagement, such as personalized product descriptions or promotional visuals.

Potential Risks: Over-personalization remains a concern, as it can lead to filter bubbles, where users are exposed only to a narrow range of products. This limits diversity and reduces the opportunity for serendipitous discovery. Furthermore, reliance on fully automated generative models introduces risks of producing irrelevant or inappropriate content, which could negatively impact user trust. These challenges highlight the importance of ongoing monitoring and refinement of recommendation algorithms.

Scalability for Smaller Vendors: The cost-effectiveness of Models 1 and 2 makes them particularly suitable for smaller businesses, as they require minimal computational resources. By providing scalable solutions, this system aligns with sustainability goals, reducing the need for high-capacity infrastructure while supporting vendor inclusivity and diversity in the marketplace.

6 Conclusion

This study explored the integration of modern NLP techniques into recommendation systems, focusing on four distinct models designed to improve recommendation quality and address the challenges posed by the cold start problem. Through quantitative and qualitative analyses, we demonstrated the potential of embeddings, sentiment analysis, and large language models (LLMs) to enhance user satisfaction and system effectiveness.

Key findings highlight the importance of combining similarity-based retrieval with sentiment and semantic reasoning. Models 2 and 3 consistently outperformed the baseline approach, achieving higher user satisfaction as indicated by average star ratings and like ratios. These results underscore the role of sentiment analysis and LLM-driven semantic understanding in tailoring recommendations to user preferences. The exploratory generative approach (Model 4) further showcased the utility of generative AI in creating novel, user-aligned recommendations, albeit with limitations in diversity and scalability.

Despite its success, this work also identified critical areas for improvement, including addressing data sparsity, enhancing the diversity of generative outputs, and refining evaluation metrics to capture nuanced aspects of user engagement. These insights lay the groundwork for future advancements in recommendation systems, emphasizing the need for hybrid approaches that combine retrieval and generative methods for balanced performance (Evidently AI, 2024).

The implications of this study extend beyond academic exploration, offering practical solutions for businesses aiming to deliver personalized user experiences. By leveraging computationally efficient models, even smaller e-commerce platforms can adopt sophisticated recommendation strategies, bridging the gap between state-of-the-art technology and real-world applicability.

In conclusion, this work demonstrates that integrating modern NLP methods into recommendation systems can lead to significant improvements in user satisfaction and recommendation quality. Future research should continue to explore hybrid models, diverse datasets, and fairness considerations to advance the field and ensure its responsible application.

References

Jose Camacho-Collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, Eugenio Martínez Cámara, et al. 2022. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, UAE. Association for Computational Linguistics.

Cach N. Dang, María N. Moreno-García, and Fernando De la Prieta. 2021. An approach to integrating sentiment analysis into recommender systems. *Sensors*, 21(16):5666. Submission received: 20 July 2021 / Revised: 13 August 2021 / Accepted: 19 August 2021 / Published: 23 August 2021.

Evidently AI. 2024. Ranking and recommendation metrics guide: 10 metrics to evaluate recommender and ranking systems. Last updated: September 5, 2024, accessed December 12, 2024.

Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2023. Recommendation as language processing (rlp): A unified pretrain, personalized prompt predict paradigm (p5).

M. Grčar, D. Mladenič, B. Fortuna, and M. Grobelnik. 2006. Data sparsity issues in the collaborative filtering framework. In O. Nasraoui, O. Zaïane, M. Spiliopoulou, B. Mobasher, B. Masand, and P.S. Yu, editors, *Advances in Web Mining and Web Usage Analysis*, volume 4198 of *Lecture Notes in Computer Science*, pages 61–72. Springer, Berlin, Heidelberg.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024a. Bridging language and items for retrieval and recommendation.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024b. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

U. Javed, K. Shaukat, A. Hameed, F. Iqbal, T. Mahboob Alam, and S. Luo. 2021. A review of content-based and context-based recommendation systems. *International Journal of Emerging Technologies in Learning (iJET)*, 16(3):274–306. Retrieved December 12, 2024.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

OpenAI. 2023. Dall·e 3: A new era of image generation. Accessed: 2024-12-12.

OpenAI. 2024. Gpt-4o-mini: A smaller, efficient version of gpt-4.

Yang Ping, Yi Li, and Jie Zhu. 2024. Beyond accuracy measures: the effect of diversity, novelty and serendipity in recommender systems on user engagement. *Electronic Commerce Research*. Accepted: 18 January 2024 / Published: 18 February 2024.

RoX818. 2024. Solving the cold start problem in recommendation systems. An AI blog, accessed December 12, 2024.