Leveraging Semantic Embeddings of User Reviews with Off-the-Shelf LLMs for Recommender Systems

Nicola Cecere[1,*], Andrea Pisani[1,2,*], Maurizio Ferrari Dacrema[1] and Paolo Cremonesi[1]

[1]*Politecnico di Milano, Italy*
[2]*Politecnico di Torino, Italy*

Abstract

Enhancing Recommender Systems (RS) with plain-text reviews has been a challenging effort despite significant efforts in the past. Recently, Large Language Models (LLMs) have demonstrated exceptional capabilities in understanding natural language semantics, leading to promising applications across various fields. Nonetheless, applying these models to recommendation tasks introduces several challenges, including high computational demands and the potential for generating inaccurate or fabricated content ("hallucinations"). Consequently, instead of directly employing LLMs as generative models for recommendations, our research explores whether embeddings derived from plain-text reviews can enrich traditional recommendation algorithms and analyze the recommendation impact of different LLM embeddings with high effectiveness in NLP tasks. We conduct our experimental analysis using two Amazon Review Datasets, and three pre-trained LLM embedding models.

Keywords

Large Language Models, User Reviews, Recommendation, Text embedding

1. Introduction

Recommender systems (RS) are essential tools in navigating the extensive digital catalogs available today, where users frequently contribute reviews of their chosen items. Historically, these textual reviews have been underutilized in RS, despite their potential to enhance system accuracy and user satisfaction [1]. Recent advancements in Natural Language Understanding, driven by the development of Large Language Models (LLMs), offer new opportunities to leverage these textual reviews effectively. In this study, we investigate the integration of semantic-rich embeddings generated from LLMs into traditional content-based (CBF) and collaborative filtering (CF) RS. Our research is driven by two primary objectives: firstly, to show how traditional RS work when augmented with review embeddings as side information, and secondly, to conduct a comparative analysis of various LLMs regarding their effectiveness.

The 14th Italian Information Retrieval Workshop, 5th - 6th September 2024, Udine, Italy
*Corresponding author.
nicola.cecere@mail.polimi.it (N. Cecere); andrea.pisani@polito.it (A. Pisani); maurizio.ferrari@polimi.it (M. Ferrari Dacrema); paolo.cremonesi@polimi.it (P. Cremonesi)
0009-0004-8486-6844 (N. Cecere); 0009-0001-9736-522X (A. Pisani); 0000-0001-7103-2788 (M. Ferrari Dacrema); 0000-0002-1253-8081 (P. Cremonesi)

2. Background

Numerous methods have been developed to extract information from reviews, however, most of these techniques are over a decade old and have fallen out of common use. This decline is largely because review-based RS rarely achieve the recommendation accuracy of CF models [1]. LLM technology has significantly influenced various fields, including RS. These models are increasingly utilized to provide text-based recommendations, in the form of generative RS. Employing methods such as pre-training, fine-tuning, and prompt engineering [2], LLMs improve conversational RS where recommendations are generated through natural language interactions. Additionally, LLM embeddings serve as auxiliary data to enhance traditional CF models, using approaches that are similar to this study but applied over different textual data, such as product descriptions.

Our interest is in models that incorporate embeddings as side information, enhancing contextual relevance. UniSRec [3] and Text-based CF (TCF) [4] explore the use of LLM embeddings in RS. UniSRec integrates these embeddings from various sub-categories of the Amazon Review Dataset for both pre-training and Parameter-Efficient Fine-Tuning (PEFT) [5], thereby enhancing its behavior encoder to improve recommendation effectiveness. TCF assesses the impact of LLM scale on the quality of recommendations. Finally, a recent study on review-based RS demonstrated promising preliminary results, further supporting the potential of these approaches [6].

3. Models and Methodology

To evaluate the effectiveness of using review embeddings generated by LLMs as side information in simple RS models, the following purely CF are used as a baseline:

ItemKNN-CF [7], **UserKNN-CF** [8], **MF** [9], **SLIM EN** [10], **GF-CF** [11], **RP$^3\beta$** [12].

Additionally, to explore the integration of LLM embeddings into RS, the following models have been adapted:

- **ItemKNN-CBF-E:** a neighbourhood CBF approach. The item features consist of a pooling (either the mean, the sum or the element-wise maximum) of all the review embeddings available for it.
- **ItemKNN-CFCBF-E - UserKNN-CFCBF-E:** a hybrid approach combining CF and CBF. The item features concatenate interaction data and pooled review embeddings, weighted by a hyperparameter.
- **RP$^3\beta$-E:** a graph-based approach derived from the RP$^3\beta$ baseline. The review embeddings are considered as a third class of nodes in the graph. They are pooled by item, so that each item i will have a corresponding embedding vector $e_i \in \mathbb{R}^F$.

Two different categories of the Amazon Reviews Dataset [13, 14] are used: the 2014 version of the Digital Music category, and the 2012 version of the Fine Foods category. For both datasets, the preprocessing consisted in extracting the 5-core subgraph, to reduce the datasets size on account of the computational cost of embedding the reviews using multiple LLMs. Both datasets

are split in train, validation and tests sets. The train sets contain 80% of the corresponding dataset's interactions, while the validation and test sets contain 10% each.

The plain-text reviews contained in both datasets were embedded using multiple LLMs, selected from the MTEB Leaderboard [15]: OpenAI's `text-embedding-ada-002`[1], Wang's `e5-mistral-7b-instruct` [16], and VoyageAI's `voyage-lite-02-instruct`[2].

The hyperparameters of the models presented above are optimized using Bayesian Optimization [17, 18]. We evaluate the models on the top-K recommendation task with NDCG at cutoff 10. We also report two beyond-accuracy metrics to measure how the recommendations are distributed: Item Coverage, which represents the quota of items in the catalogue that were recommended at least once, and Item Coverage Hit, which represents the quota of items in the catalogue that were recommended *correctly* at least once.

We investigate the possible presence of consistent trends in recommendation accuracy by evaluating all models that employ LLM embeddings with three different sets of embeddings presented above. No LLM fine-tuning nor prompt engineering is employed.

4. Results and Future Investigations

Table 1 shows evaluation results for the Amazon Music and Fine Foods Dataset. In both cases, the models that exploit reviews embedded by LLMs as side information fail to outperform the best performing pure CF baselines in recommendation effectiveness, measured by NDCG. This is evidence that LLM embeddings do not pair well with simple recommenders, and likely need more expressive architectures to be interpreted effectively and bring value.

Enriching the side information by pooling the available embeddings did not produce the expected results. We hypothesize that the method of combination is the core issue, as the pooling technique described in Section 3 may not adequately preserve the nuances of individual reviews [19]. To address this, we are investigating a novel approach that uses embeddings from user-preferred item reviews to construct detailed user profiles. These profiles are then used to compute similarity scores with the single embeddings of other item reviews.

On average, OpenAI's embedder achieves the highest accuracy on the Amazon Music Dataset with a mean NDCG of 0.118, surpassing VoyageAI's embedder by 2.59%. Wang's embedder performs the worst, with a mean NDCG of 0.096. Across all models, Wang's embeddings yield the lowest recommendation accuracy.

In the Amazon Fine Foods Dataset, embedders are more balanced. OpenAI's embeddings achieve the highest average accuracy (NDCG of 0.647), slightly surpassing Wang's (0.645) and VoyageAI's (0.643).

The ItemKNN-CBF-E model, relying solely on embeddings, can be useful to establish the inherent quality of the embeddings for recommendation. It shows the best NDCG for OpenAI on the Music Dataset (0.128), followed by VoyageAI (0.123) and Wang (0.055). On the Fine Foods Dataset, OpenAI again leads (NDCG of 0.799), with Wang (0.756) outperforming VoyageAI (0.747). Notably, the model's top-K parameter for Wang's embeddings on the Fine Foods Dataset

[1] https://openai.com/blog/new-and-improved-embedding-model
[2] https://docs.voyageai.com/docs/embeddings

Table 1
Experimental results for models over the Amazon Music and Fine Foods Dataset. Baseline models have an empty 'LLM' column.

Model Name	Amazon Music Dataset				Amazon Fine Foods Dataset		
	LLM	NDCG	Item Cov.	Item Cov. Hit	NDCG	Item Cov.	Item Cov. Hit
MF	-	0.137	0.806	0.169	0.822	0.922	0.464
ItemKNN-CF	-	0.170	0.766	0.188	0.845	0.949	0.486
UserKNN-CF	-	0.153	0.916	0.209	0.841	0.931	0.484
$RP^3\beta$	-	**0.171**	0.802	0.184	0.843	0.940	0.483
GF-CF	-	0.163	0.830	0.190	0.841	0.930	0.476
SLIM EN	-	**0.171**	0.790	0.202	**0.851**	0.923	0.488
ItemKNN-CBF-E	OpenAI	0.128	0.913	0.207	0.799	0.951	0.480
	Wang	0.055	0.160	0.048	0.756	0.597	0.446
	VoyageAI	0.123	0.929	0.193	0.747	0.951	0.444
ItemKNN-CFCBF-E	OpenAI	0.157	0.687	0.176	0.837	0.943	0.484
	Wang	0.133	0.420	0.117	0.830	0.928	0.471
	VoyageAI	0.152	0.767	0.174	0.824	0.881	0.462
UserKNN-CFCBF-E	OpenAI	0.146	0.529	0.145	0.766	0.802	0.430
	Wang	0.132	0.592	0.126	0.814	0.822	0.440
	VoyageAI	0.138	0.468	0.135	0.809	0.759	0.430
$RP^3\beta$-E	OpenAI	0.162	0.718	0.165	0.833	0.948	0.476
	Wang	0.161	0.881	0.193	0.832	0.950	0.477
	VoyageAI	0.165	0.846	0.184	0.833	0.944	0.476

is 996, contrasting with 22 for the Music Dataset and below 10 for OpenAI and VoyageAI on both datasets.

Embedding vector size seems to influence model performance. Shorter vectors, from OpenAI (1,536) and VoyageAI (1,024), yield better accuracy than Wang's larger vectors (4,096). The simpler architecture of the models likely limits their ability to process Wang's complex embeddings. The effectiveness of LLMs in top-K recommendation does not align with their MTEB leaderboard positions for NLP tasks. OpenAI's embedder, despite ranking below 71st (as of June 2024), is the most effective for recommendation. VoyageAI and Wang, ranked 10th and 12th, perform worse in comparison. Wang's embedder shows varying effectiveness, sometimes significantly better or worse than VoyageAI's despite their close leaderboard positions.

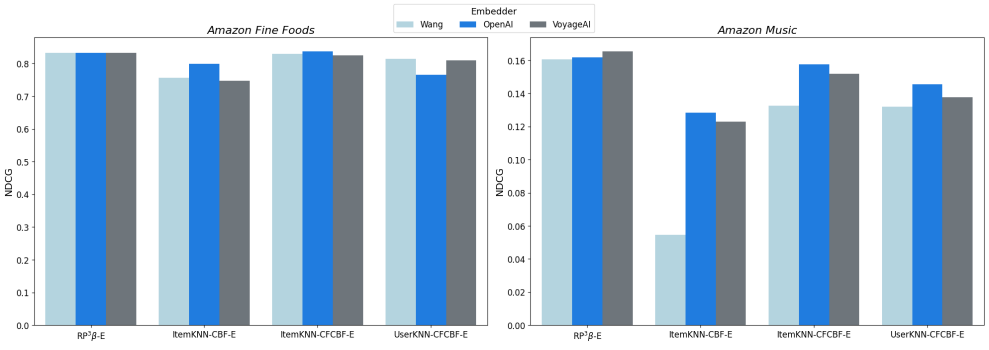

Figure 1: Evaluation results in terms of NDCG for all models that employ LLM embeddings

References

[1] N. Sachdeva, J. J. McAuley, How useful are reviews for recommendation? A critical review and potential improvements, in: J. X. Huang, Y. Chang, X. Cheng, J. Kamps, V. Murdock, J.-R. Wen, Y. Liu (Eds.), Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020, ACM, 2020, pp. 1845–1848. doi:10.1145/3397271.3401281.

[2] W. Fan, Z. Zhao, J. Li, Y. Liu, X. Mei, Y. Wang, J. Tang, Q. Li, Recommender systems in the era of large language models (LLMs), CoRR abs/2307.02046 (2023). doi:10.48550/ARXIV.2307.02046. arXiv:2307.02046.

[3] Y. Hou, S. Mu, W. X. Zhao, Y. Li, B. Ding, J.-R. Wen, Towards universal sequence representation learning for recommender systems, in: A. Zhang, H. Rangwala (Eds.), KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022, ACM, 2022, pp. 585–593. doi:10.1145/3534678.3539381.

[4] R. Li, W. Deng, Y. Cheng, Z. Yuan, J. Zhang, F. Yuan, Exploring the upper limits of text-based collaborative filtering using large language models: Discoveries and insights, CoRR abs/2305.11700 (2023). doi:10.48550/ARXIV.2305.11700. arXiv:2305.11700.

[5] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, volume 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 2790–2799.

[6] A. Pisani, N. Cecere, M. Ferrari Dacrema, P. Cremonesi, Pre-trained llm embeddings of product reviews for recommendation, in: E. Maddalena, S. Mizzaro, K. Roitero, M. Viviani (Eds.), Proceedings of the 14th Italian Information Retrieval Workshop (IIR 2024), Udine, Italy, September 5-6, 2024, CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[7] J. Wang, A. P. de Vries, M. J. T. Reinders, Unifying user-based and item-based collaborative filtering approaches by similarity fusion, in: E. N. Efthimiadis, S. T. Dumais, D. Hawking, K. Järvelin (Eds.), SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, August 6-11, 2006, ACM, 2006, pp. 501–508. doi:10.1145/1148170.1148257.

[8] B. M. Sarwar, G. Karypis, J. A. Konstan, J. Riedl, Item-based collaborative filtering recommendation algorithms, in: V. Y. Shen, N. Saito, M. R. Lyu, M. E. Zurko (Eds.), Proceedings of the Tenth International World Wide Web Conference, WWW 10, Hong Kong, China, May 1-5, 2001, ACM, 2001, pp. 285–295. doi:10.1145/371920.372071.

[9] Y. Koren, Factorization meets the neighborhood: A multifaceted collaborative filtering model, in: Y. Li, B. Liu, S. Sarawagi (Eds.), Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008, ACM, 2008, pp. 426–434. doi:10.1145/1401890.1401944.

[10] X. Ning, G. Karypis, SLIM: sparse linear methods for top-n recommender systems, in: D. J. Cook, J. Pei, W. Wang, O. R. Zaïane, X. Wu (Eds.), 11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011, IEEE Computer Society, 2011, pp. 497–506. URL: https://doi.org/10.1109/ICDM.2011.134. doi:10.1109/ICDM.2011.134.

[11] Y. Shen, Y. Wu, Y. Zhang, C. Shan, J. Zhang, K. B. Letaief, D. Li, How powerful is graph convolution for recommendation?, in: G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, H. Tong (Eds.), CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021, ACM, 2021, pp. 1619–1629. URL: https://doi.org/10.1145/3459637.3482264. doi:10.1145/3459637.3482264.

[12] B. Paudel, F. Christoffel, C. Newell, A. Bernstein, Updatable, accurate, diverse, and scalable recommendations for interactive applications, ACM Trans. Interact. Intell. Syst. 7 (2017) 1:1–1:34. doi:10.1145/2955101.

[13] J. J. McAuley, J. Leskovec, From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews, in: D. Schwabe, V. A. F. Almeida, H. Glaser, R. Baeza-Yates, S. B. Moon (Eds.), 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, International World Wide Web Conferences Steering Committee / ACM, 2013, pp. 897–908. doi:10.1145/2488388.2488466.

[14] R. He, J. J. McAuley, Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering, in: J. Bourdeau, J. Hendler, R. Nkambou, I. Horrocks, B. Y. Zhao (Eds.), Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016, ACM, 2016, pp. 507–517. doi:10.1145/2872427.2883037.

[15] N. Muennighoff, N. Tazi, L. Magne, N. Reimers, MTEB: Massive text embedding benchmark, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023, Association for Computational Linguistics, 2023, pp. 2006–2029. doi:10.18653/V1/2023.EACL-MAIN.148.

[16] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, F. Wei, Improving text embeddings with large language models, CoRR abs/2401.00368 (2024). doi:10.48550/ARXIV.2401.00368. arXiv:2401.00368.

[17] P. I. Frazier, Bayesian Optimization, in: E. Gel, L. Ntaimo, D. Shier, H. J. Greenberg (Eds.), Recent Advances in Optimization and Modeling of Contemporary Problems, INFORMS, 2018, pp. 255–278. doi:10.1287/educ.2018.0188.

[18] M. Ferrari Dacrema, S. Boglio, P. Cremonesi, D. Jannach, A troubling analysis of reproducibility and progress in recommender systems research, ACM Transactions on Information Systems 39 (2021) 20:1–20:49. URL: https://doi.org/10.1145/3434185. doi:10.1145/3434185.

[19] M. M. Abdollah Pour, P. Farinneya, A. Toroghi, A. Korikov, A. Pesaranghader, T. Sajed, M. Bharadwaj, B. Mavrin, S. Sanner, Self-supervised contrastive bert fine-tuning for fusion-based reviewed-item retrieval, in: European Conference on Information Retrieval, Springer, 2023, pp. 3–17.