Pre-Trained LLM Embeddings of Product Reviews for Recommendation

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

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

Abstract

A significant amount of past literature has shown that it is difficult to leverage plain-text reviews to improve recommendation effectiveness. Since then, Large Language Models (LLMs) have shown unprecedented ability to capture natural language semantics, which has been applied to multiple domains with good results. However, re-purposing them for recommendation is not straightforward, due to their high computational cost and the risk of hallucinations. For these reasons, rather than using LLMs as models to directly generate recommendations, we investigate if LLM embeddings of plain-text reviews can be a useful input to improve the quality of traditional review-based recommendation algorithms, by adapting their architecture to process said embeddings rather than word-level ones. We structure an empirical analysis using two Amazon Review Datasets and three LLMs to produce embeddings: OpenAI, Wang's Mistral and VoyageAI. The results show that LLM embeddings can be effectively used in review-based models developed for word-level embeddings, yet one baseline model still achieves greater accuracy.

Keywords

Large Language Models, User Reviews, Recommendation, Text embedding

1. Introduction

Recommender systems (RS) are widely adopted to help people navigate the vast and expanding catalogues of digital platforms, most of which encourage users to also leave reviews for the items they select. Despite their abundance, said reviews are underutilised by current RS [1], even though numerous efforts have been made to extract and leverage valuable information from reviews to enhance recommendation quality. Matching the benchmark set by state-of-the-art Collaborative Filtering (CF) models remains a challenging goal for review-based RS.

Nonetheless, textual reviews hold untapped potential for RS, since many users rely on them in their browsing. It can be argued that the shortcomings of previous integration attempts are related to Natural Language Understanding, a field in which recent advances, particularly through Large Language Models (LLMs), have been noteworthy. LLMs are transformer-based neural architectures with billions of trainable parameters.

IIR2024: The 14th Italian Information Retrieval Workshop, 5th-6th September 2024, Udine, Italy
*Corresponding author.
andrea.pisani@polito.it (A. Pisani); nicola.cecere@mail.polimi.it (N. Cecere); maurizio.ferrari@polimi.it (M. Ferrari Dacrema); paolo.cremonesi@polimi.it (P. Cremonesi)
0009-0001-9736-522X (A. Pisani); 0009-0004-8486-6844 (N. Cecere); 0000-0001-7103-2788 (M. Ferrari Dacrema); 0000-0002-1253-8081 (P. Cremonesi)

There have been several attempts to use LLMs as RS, primarily exploiting them as generative RS. However, such attempts show a number of limitations, mostly tied to the need for extensive refactoring of interaction data, the high computational cost of training and fine-tuning LLMs, and model reliability problems, such as hallucinations [2].

In this study, leveraging LLM-generated embeddings from plain-text product reviews, we investigate integrating the semantic capabilities of LLMs with traditional review-based RS. We then compare the effectiveness of existing algorithms using word-level embeddings to those using LLM-generated embeddings.

2. Background

Chen et al. [3] classify review-based RS into four categories, based on their way of exploiting reviews: *text-mining* models, *sentiment analysis* models, *rating weighting* models and *item profile enrichment* models. Among review-based RS, we focus on three among the most effective algorithms: Hidden Factors and Hidden Topics (HFT) [4], Neural Attentional Regression with Review-level Explanations (NARRE) [5] and Hybrid neural Recommendation with joint Deep Representation learning of ratings and reviews (HRDR) [6].

HFT is a *rating weighting* RS based on the combination of two techniques: a classic global effects-aware matrix factorisation (MF) model [7], which predicts ratings according to latent representations of users and items, and Latent Dirichlet Allocation (LDA) [8], which is used to extract topics from review texts. A likelihood function over the extracted topics is used as a regularization term in its squared loss function.

NARRE [5] is a two-tower *profile enrichment* neural model. The user tower is fed with all the reviews posted by a given user, while the item tower is fed with the reviews regarding a given item. In both towers, the reviews are decomposed in word-level embeddings [9] and processed by a Convolutional Neural Network. The resulting feature vectors are weighted through an attention layer, whose output is remapped through a fully connected neural layer. Representations are merged to the ratings-based user and item profiles, and the rating prediction is computed as a dot product of the feature vectors.

HRDR [6] is a two-tower *profile enrichment* neural model, very similar to NARRE in terms of general intuition and architecture, with two main differences. First of all, the URM is processed through Multi-Layer Perceptrons which output one latent representation per tower. These latent representations are multi-purposed: they are injected in the final merging of profiles right before computing the rating predictions through global effects-aware MF, and also into the attention layer that processes the reviews.

3. Methodology

We adapted NARRE and HRDR to use LLM embeddings of reviews as input to compare their effectiveness when using word-level embeddings versus LLM embeddings. Originally, their input would be a matrix of word embeddings representing a single review, which was aggregated into a single vector through the use of a CNN module. As LLM-based embeddings already represent the reviews in whole, such CNN module was eliminated in their adapted versions, which we

respectively named NARRE-LLM and HRDR-LLM. HFT was considered as a baseline model and not adapted in its architecture, since it does not incorporate review embeddings, but instead exploits LDA to do review topic modelling. We also slightly modified all the models, including HFT, in their training process, in order to evaluate them on top-K recommendation tasks instead of rating prediction. For NARRE, HRDR, NARRE-LLM and HRDR-LLM, we changed the loss function from squared loss to Bayesian Personalised Ranking (BPR). The HFT algorithm defines a custom loss function which we did not change; however, for the top-K recommendation task, the model needs to distinguish positive user-item interactions from negative ones, i.e., those that did not occur, associated to a rating value of 0. Thus, our version of HFT samples negative interactions during training, with uniform probability defined by a hyperparameter.

We use two datasets, consisting of different categories of the Amazon Reviews Dataset [10, 11]: the 2014 version of the Digital Music category, and the 2012 version of the Fine Foods category. The datasets were preprocessed by extracting their 5-core subgraph. This was done to reduce their size, considering the high computational cost of embedding reviews using multiple LLMs. The interactions of both datasets were split in 80% training, 10% validation and 10% test.

The plain-text reviews contained in both datasets were embedded using multiple LLMs, selected from the MTEB Leaderboard [12]: OpenAI's `text-embedding-ada-002`[1], Wang's `e5-mistral-7b-instruct` [13], and VoyageAI's `voyage-lite-02-instruct`[2]. While Wang's and VoyageAI's embedders were chosen for their high position within the leaderboard, OpenAI's embedder was selected for its widespread use.

The models were trained iteratively through stochastic gradient descent, using Bayesian Optimization [14, 15] to optimize hyperparameters with respect to the NDCG@10 metric over the validation set. To mitigate the risk of overfitting, we employ early-stopping, performing an evaluation over the validation set every 5 epochs and terminating the training if NDCG@10 has not improved throughout the 5 latest evaluations. We evaluated 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 measures the portion of items in the catalogue that were recommended at least once, and Item Coverage Hit, which represents the portion of items that were *correctly* recommended at least once.

4. Results

Evaluation results are shown in Table 1. For both datasets, using the LLM embeddings as input for HRDR and NARRE resulted in visible improvements in recommendation effectiveness. NARRE-LLM improved up to 47.2% in NDCG with respect to NARRE over the Amazon Music Dataset, and up to 100.6% over the Amazon Fine Foods Dataset, employing embeddings from OpenAI and Wang, respectively. Similarly, HRDR-LLM outperforms HRDR by up to 25.5% and 5.84% over the two datasets, employing embeddings from VoyageAI and OpenAI. All models also visibly improve with respect to Item Coverage and Item Coverage Hit, over both datasets. In particular, NARRE-LLM achieves the best Item Coverage when using Wang's embeddings, while HRDR-LLM's Item Coverage improves the most when using OpenAI's embeddings.

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

Table 1
Experimental results for review-based models. Baseline models have an empty 'LLM' column. The best NDCG value is highlighted in bold.

		Amazon Music Dataset			Amazon Fine Foods Dataset		
Model	LLM	NDCG	Item Cov.	Item Cov. Hit	NDCG	Item Cov.	Item Cov. Hit
HFT	-	**0.131**	0.694	0.167	**0.798**	0.885	0.458
NARRE	-	0.053	0.105	0.039	0.168	0.202	0.108
HRDR	-	0.094	0.625	0.138	0.719	0.583	0.369
NARRE-LLM	OpenAI	0.078	0.230	0.076	0.167	0.137	0.089
	Wang	0.066	0.295	0.072	0.337	0.376	0.245
	VoyageAI	0.069	0.151	0.057	0.206	0.218	0.134
HRDR-LLM	OpenAI	0.105	0.731	0.139	0.761	0.760	0.402
	Wang	0.101	0.543	0.126	0.750	0.665	0.378
	VoyageAI	0.118	0.499	0.137	0.754	0.729	0.392

Nevertheless, the HFT model, when trained with implicit interaction data and negative interaction sampling, achieved the best NDCG values. It outperforms HRDR-LLM by a margin of 11.02% over the Amazon Music Dataset and by 4.86% over the Amazon Fine Foods Dataset. NARRE-LLM shows the poorest recommendation accuracy, being outperformed by HFT by a margin of 67.95% when tested on the Amazon Music Dataset, and of 136.8% over the Amazon Fine Foods Dataset. HFT performs solidly also in terms of both Item Coverage and Item Coverage Hit. In terms of Item Coverage, HRDR-LLM with OpenAI embeddings is the best performer over the Music Dataset, improving HFT's baseline by 5.33%; over the Fine Foods Dataset, HFT performs the best, being 16.44% better than HRDR-LLM with OpenAI embeddings. In terms of Item Coverage Hit, HFT is on top over both datasets, while HRDR-LLM with OpenAI embeddings is second best in both cases.

5. Conclusions

Although LLM embeddings are indeed an improvement over the ones used by NARRE and HRDR, exploiting reviews as means for regularization like HFT still appears the better choice. Since HFT only exploits reviews through LDA-based topic modelling, thus not making use of review embeddings at all, its recommendation accuracy also comes at a lower computational cost. These findings on top-K accuracy are consistent with those of [1], which focused on rating prediction instead. We also tested the proposed approach on non-review based models in [16].

Possible continuations to our investigation might include the use of review embeddings for item feature extraction or feature weighting, and the development of RS specifically engineered to exploit LLM embeddings of textual information as input data, possibly going beyond reviews.

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] Z. Xu, S. Jain, M. Kankanhalli, Hallucination is Inevitable: An Innate Limitation of Large Language Models (2024). doi:10.48550/ARXIV.2401.11817.

[3] L. Chen, G. Chen, F. Wang, Recommender systems based on user reviews: The state of the art, User Model. User Adapt. Interact. 25 (2015) 99–154. doi:10.1007/S11257-015-9155-5.

[4] J. J. McAuley, J. Leskovec, Hidden factors and hidden topics: Understanding rating dimensions with review text, in: Q. Yang, I. King, Q. Li, P. Pu, G. Karypis (Eds.), Seventh ACM Conference on Recommender Systems, RecSys '13, Hong Kong, China, October 12-16, 2013, ACM, 2013, pp. 165–172. doi:10.1145/2507157.2507163.

[5] C. Chen, M. Zhang, Y. Liu, S. Ma, Neural attentional rating regression with review-level explanations, in: P.-A. Champin, F. Gandon, M. Lalmas, P. G. Ipeirotis (Eds.), Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018, ACM, 2018, pp. 1583–1592. doi:10.1145/3178876.3186070.

[6] H. Liu, Y. Wang, Q. Peng, F. Wu, L. Gan, L. Pan, P. Jiao, Hybrid neural recommendation with joint deep representation learning of ratings and reviews, Neurocomputing 374 (2020) 77–85. doi:10.1016/J.NEUCOM.2019.09.052.

[7] Y. Koren, R. M. Bell, Advances in collaborative filtering, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer, 2015, pp. 77–118. doi:10.1007/978-1-4899-7637-6_3.

[8] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, in: T. G. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada], MIT Press, 2001, pp. 601–608.

[9] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. J. C. Burges, L. Bottou, Z. Ghahramani, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a Meeting Held December 5-8, 2013, Lake Tahoe, Nevada, United States, 2013, pp. 3111–3119.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] N. Cecere, A. Pisani, M. Ferrari Dacrema, P. Cremonesi, Leveraging semantic embeddings of user reviews with off-the-shelf llms for traditional recommender systems, in: E. Maddalena, S. Mizzaro, K. Roitero, M. Viviani (Eds.), IIR2024: 14th Italian Information Retrieval Workshop, 5th - 6th September 2024, Udine, Italy, CEUR-WS.org, 2024.