How Much Do Reviews Really Contribute? A Study on Text-Enriched Matrix Factorization for Recommendations

Eduardo Ferreira da Silva[1], **Mayki dos Santos Oliveira**[1], **Joel Machado Pires**[1]
Denis Dantas Boaventura[1] **Frederico Araújo Durão**[1]

[1]Instituto de Computação – Universidade Federal da Bahia (UFBA)
Salvador – BA – Brazil

{eduardoferreira, maykioliveira, joelpires, denis.boaventura, fdurao}@ufba.br

Abstract. *Incorporating textual reviews into a Recommender System has become a prominent strategy for enriching collaborative signals with semantic information. However, the actual contribution of review-derived representations remains an open question, particularly when strong collaborative baselines are employed. In this work, we systematically investigate the impact of textual information on Matrix Factorization by introducing and comparing three enrichment strategies over a common collaborative backbone. First, we propose a learnable gating mechanism that adaptively balances collaborative and textual signals during training. This mechanism is applied to two distinct review representations: (i) aggregated topic profiles extracted from user and item histories, and (ii) full text embedding representations derived from reviews. Additionally, we explore a cross-attention mechanism that identifies and emphasizes the most informative dimensions of the textual representation before fusion with collaborative factors. We evaluate six variants: pure, enriched with topic profiles and text via gating; enriched with topics and text via gating; and enhanced with cross-attention over textual features. Experiments across multiple review-based datasets reveal that although adaptive fusion mechanisms improve representation flexibility, the marginal contribution of textual signals remains limited compared to the collaborative backbone. These findings suggest that, under typical rating-prediction settings, collaborative information continues to dominate performance, raising important considerations for the effective integration of semantic review signals into recommendation models.*

1. Introduction

Incorporating textual reviews into Recommender System (RS) has become a widely adopted strategy for enriching collaborative signals with semantic information [Silva et al. 2025b]. Reviews are often assumed to provide deeper insights into users' preferences, revealing latent aspects and subjective opinions that numerical ratings alone cannot capture [Hasan et al. 2024]. Motivated by this premise, a growing body of research has proposed review-aware recommendation models that integrate textual representations via regularization, topic models, neural architectures, and attention mechanisms [Lima et al. 2024]. The underlying assumption is straightforward: if reviews explain why users assign ratings, then leveraging them should improve rating prediction and ranking performance.

However, recent investigations have questioned the consistency and robustness of these claims. In particular, [Sachdeva and McAuley 2020] highlights discrepancies in

reported improvements across review-based RS and emphasizes the need for careful empirical validation under controlled settings. Their findings suggest that textual signals do not uniformly outperform strong collaborative baselines, especially when experimental conditions deviate from narrowly defined scenarios. This reveals a broader and still unresolved issue regarding the effective contribution of textual reviews when integrated into well-optimized collaborative models.

Beyond inconsistencies in reported gains, the existing literature exhibits two closely related limitations. First, review-aware models are predominantly optimized and evaluated under regression objectives, relying on error-based metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). While suitable for rating prediction, these metrics do not necessarily reflect ranking quality, the primary objective in many practical recommendation scenarios [da Silva et al. 2025]. Consequently, improvements in rating reconstruction may not translate into meaningful changes in top-N recommendation lists, leaving the impact of textual signals on ranking performance largely underexplored.

Second, despite the rich semantic content embedded in reviews, including aspects, sentiments, and contextual explanations, this content is typically integrated only to enhance rating prediction. The prevailing assumption is that better numerical reconstruction implies effective use of textual information. However, this perspective may be reductive, as it overlooks the broader potential of reviews to shape representation learning and influence interaction modeling beyond regression error. As a result, the true role and practical contribution of review text in collaborative recommendation remain insufficiently understood.

In this work, we revisit this issue from a controlled and systematic perspective. Rather than proposing an entirely new architecture, we adopt a strong Matrix Factorization (MF) backbone and investigate multiple strategies for incorporating review information[1]. Matrix factorization is chosen for its well-established effectiveness and interpretability in modeling latent user–item interactions. We then enrich this backbone using three alternative mechanisms designed to fuse collaborative and textual signals: (i) a learnable gating mechanism applied to topic-based user and item profiles, (ii) a gating mechanism applied to full text embedding representations, and (iii) a cross-attention mechanism that identifies and emphasizes the most informative dimensions of textual representations before fusion.

2. Related Works

The integration of review text into RS has evolved from simple regularization strategies to complex neural architectures that treat textual signals as core sources of representation. Early approaches extended matrix factorization by incorporating topic models or textual regularizers to align latent factors with semantic dimensions extracted from reviews. Later, deep learning models began to encode textual content directly using convolutional or attention-based mechanisms.

DeepCoNN [Zheng et al. 2017] is a foundational neural Review-based Recommender System (RBRS) that uses parallel convolutional networks to encode user and

[1]Available at: `https://github.com/ferreira-eduardo/review_analysis.git`

item reviews separately. It learns latent representations to predict ratings by extracting semantic features from text and combining them with collaborative signals. This approach shows that reviews can serve as primary inputs rather than just auxiliary constraints. Later developments incorporated attention mechanisms to improve performance and interpretability. For instance, DAML [Liu et al. 2019] applies dual attention to selectively focus on informative review components and align them with rating features. Similarly, KANN [Liu and Miyazaki 2023] integrates knowledge entities extracted from reviews into a knowledge-aware attention framework, enabling explainable interactions between users and items at the semantic level. More recently, graph-based approaches such as LETTER [Son et al. 2025] model user–user and item–item relations while preserving richer textual signals during propagation, thereby better capturing relational semantics.

Despite reported improvements in review-aware architectures, the empirical robustness of these gains has been questioned. [Sachdeva and McAuley 2020] conduct a comprehensive reassessment of review-based recommendation models and identify inconsistencies in experimental protocols and reported results. Their analysis shows that state-of-the-art review-aware methods often fail to outperform strong collaborative baselines in settings beyond narrowly defined ones. They emphasize the importance of rigorous evaluation and caution against overestimating the universal benefit of textual augmentation. Complementarily, [Jurdi et al. 2021] analyzes the impact of natural noise in recommender datasets, highlighting how inconsistent user behavior can distort evaluation outcomes. They argue that variability in user profiles and rating inconsistencies significantly affect model performance and call for improved evaluation methodologies that account for data quality and behavioral coherence.

More recently, the emergence of Large Language Models (LLMs) has reshaped the discussion. [Tan et al. 2025] question whether specialized review-aware architectures remain necessary and introduce RAREval, a benchmark to systematically evaluate review contributions under zero-shot, few-shot, and fine-tuning settings. Their findings suggest that pretrained LLMs can implicitly capture semantic alignment and often outperform traditional review-based architectures, particularly in sparse and cold-start scenarios.

Unlike prior works that design increasingly complex neural or graph-based architectures to maximize the impact of review text, this research adopts a controlled perspective centered on matrix factorization. Rather than replacing collaborative filtering with deep textual encoders, we investigate how different mechanisms, such as gated fusion, cross-attention, and regularization, enrich a strong collaborative backbone. Additionally, motivated by the concerns raised in [Sachdeva and McAuley 2020] and [Jurdi et al. 2021], our work emphasizes rigorous multi-metric evaluation and explicitly distinguishes between regression and ranking behavior. This allows us to analyze not only whether reviews improve prediction accuracy, but also whether they preserve collaborative transitivity and ranking effectiveness, providing a more balanced and systematic assessment of review integration.

3. Matrix Factorization Backbone

Despite the complexity of review-aware architectures, MF remains a key element in many recommendation models. Often embedded in hybrid and neural systems, MF efficiently

captures latent user-item interactions through low-dimensional representations. In this work, MF serves as the foundation for integrating textual enrichment mechanisms, allowing us to analyze how review information impacts regression error and ranking performance.

Formally, the MF model is defined as follows. Given a user u and an item i, the model learns latent factor vectors $\mathbf{p}_u, \mathbf{q}_i \in R^d$, where $\mathbf{p}_u$ and $\mathbf{q}_i$ denote the collaborative embeddings of user u and item i, respectively and d denotes the dimensionality of the latent space. To capture baseline popularity and bias effects, we include scalar bias terms b_u (user bias), b_i (item bias), and b (global offset). The predicted interaction score $\hat{s}_{ui}$ is computed as:

$$\hat{s}_{ui} = \mathbf{p}_u^\top \mathbf{q}_i + b_u + b_i + b. \tag{1}$$

To ensure the outputs are bounded and comparable to the ground truth, we apply a logistic sigmoid function $\sigma(\cdot)$ to map the scores into the specific rating range $[r\min, r\max]$:

$$\hat{r}_{ui} = \sigma(\hat{s}_{ui})(r_{\max} - r_{\min}) + r_{\min}. \tag{2}$$

The model is optimized using mean squared error (MSE) between $\hat{r}_{ui}$ and the observed rating r_{ui}.

3.1. Gating Mechanism

To integrate collaborative and textual representations, we employ a learnable gating mechanism that adaptively balances both information sources in the latent factor space [Silva et al. 2025a]. Let $\mathbf{h}_u$ and $\mathbf{h}_i$ represent the raw topic-derived profiles extracted from review text. To ensure compatibility, these representations are aligned to the collaborative space through linear projections, yielding the textual embeddings $\mathbf{p}_u^t, \mathbf{q}_i^t \in R^d$:

$$\mathbf{p}_u^t = \mathbf{W}_u\, \mathbf{h}_u, \qquad \mathbf{q}_i^t = \mathbf{W}_i\, \mathbf{h}_i, \tag{3}$$

where $\mathbf{W}_u$, $\mathbf{W}_i$ are learnable projection matrices that map the topic features into the d-dimensional latent space.

Additionally, fusion is performed directly in the collaborative latent space using a gating network. For a generic entity $e \in \{u, i\}$ (user or item), the learned gate vector is computed as:

$$\mathbf{g}_e = \sigma\left(\mathbf{W}_g \left[\mathbf{z}_e^{cf} \,\|\, \mathbf{z}_e^t\right] + \mathbf{b}_g\right), \tag{4}$$

where $\mathbf{z}_e^{cf}$ denotes the collaborative representation ($\mathbf{p}_u^{cf}$ or $\mathbf{q}_i^{cf}$), $\mathbf{z}_e^t$ represents the textual/topic embedding ($\mathbf{p}_u^t$ or $\mathbf{q}_i^t$), $\|$ denotes vector concatenation, $\sigma(\cdot)$ is the sigmoid activation ensuring $\mathbf{g}_e \in (0,1)^d$, and $\mathbf{W}_g$, $\mathbf{b}_g$ are learnable parameters.

The final fused representations are obtained via element-wise interpolation. For a generic entity $e \in \{u, i\}$ (user or item), we define:

$$\tilde{\mathbf{z}}_e = \mathbf{g}_e \odot \mathbf{z}_e^{cf} + (1 - \mathbf{g}_e) \odot \mathbf{z}_e^t, \tag{5}$$

where $\odot$ denotes element-wise multiplication. By performing fusion in the aligned latent space, the model maintains compatibility with the matrix factorization objective while enabling adaptive integration of semantic information derived from reviews.

3.2. Cross Attention

In addition to gated interpolation, we investigate a cross-attention mechanism proposed in [Vaswani et al. 2017] to selectively integrate textual representations into the collaborative latent space. Unlike gating, which performs element-wise interpolation between two aligned vectors, cross-attention enables the collaborative embedding to dynamically attend to multiple textual components, emphasizing the most informative dimensions conditioned on the interaction.

Let $\mathbf{p}_u^{cf} \in \mathbf{R}^d$ denote the collaborative embedding of a user (analogously for items), and let $\mathbf{T}_u \in \mathbf{R}^{K \times d_t}$ represents a sequence of K textual vectors derived from review representations (e.g., topic vectors or token/segment embeddings), where d_t denotes the dimensionality of the textual space.

To enable interaction between both modalities, we project them into a shared latent space of dimension d:

$$\mathbf{Q}_u = \mathbf{W}_Q \mathbf{p}_u^{cf}, \qquad \mathbf{K}_u = \mathbf{T}_u \mathbf{W}_K, \qquad \mathbf{V}_u = \mathbf{T}_u \mathbf{W}_V, \tag{6}$$

where $\mathbf{W}_Q \in \mathbf{R}^{d \times d}$ and $\mathbf{W}_K, \mathbf{W}_V \in \mathbf{R}^{d_t \times d}$ are learnable projection matrices. In this formulation, $\mathbf{Q}_u$ acts as the query, while $\mathbf{K}_u$ and $\mathbf{V}_u$ correspond to the key and value matrices derived from the textual sequence. Attention scores are computed using scaled dot-product attention:

$$\mathbf{a}_u = \frac{\mathbf{Q}_u \mathbf{K}_u^\top}{\sqrt{d}}, \tag{7}$$

where the scaling factor $\sqrt{d}$ stabilizes gradients. When a validity mask $\mathbf{m}_u \in \{0,1\}^K$ is available (e.g., to handle padded positions), invalid entries are assigned large negative values before normalization. The attention weights, context vector, and fused representation are then computed as:

$$\boldsymbol{\alpha}_u = \text{softmax}(\mathbf{a}_u), \qquad \mathbf{c}_u = \boldsymbol{\alpha}_u \mathbf{V}_u, \qquad \tilde{\mathbf{p}}_u = \mathbf{p}_u^{cf} + \mathbf{c}_u. \tag{8}$$

The vector $\mathbf{c}_u$ represents a weighted aggregation of textual information, where weights are determined by the relevance of each textual component to the collaborative query. The final representation $\tilde{\mathbf{p}}_u$ is obtained by adding residuals, preserving the original collaborative signal while allowing selective enhancement from review-derived features.

3.3. Text Regularizer

As a complementary strategy to adaptive fusion, we incorporate review information into matrix factorization via regularization. This approach preserves the pure collaborative interaction for rating estimation while encouraging alignment between collaborative and textual representations during training. In this context, textual information, either topic-based profiles or dense embedding representations, is integrated through an auxiliary regularization term. Let $\mathbf{p}_u^t$ and $\mathbf{q}_i^t$ denote the projected textual representations aligned to the collaborative latent space. The overall objective becomes:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda_u \|\mathbf{p}_u^{cf} - \mathbf{p}_u^t\|_2^2 + \lambda_i \|\mathbf{q}_i^{cf} - \mathbf{q}_i^t\|_2^2, \tag{9}$$

where $\mathcal{L}_{\text{MSE}}$ is the rating prediction loss and λ_u, λ_i control the influence of textual alignment. This formulation treats textual information as a structural prior rather than a direct predictive component. The collaborative embeddings remain solely responsible for rating

estimation, while textual representations guide their geometry in the latent space. As a result, the model allows us to analyze whether textual signals indirectly improve generalization, without explicitly altering the interaction function.

By contrasting this regularization-based approach with gated interpolation and cross-attention fusion, we obtain multiple perspectives on how review-derived information influences collaborative recommendation. This comparative framework enables a controlled investigation into whether textual signals meaningfully reshape latent representations or whether their contribution remains marginal when strong collaborative baselines are employed.

4. Topics Extraction

The topic extraction stage aims to identify fine-grained aspects of user experience expressed in reviews, such as emotional tone, immersion, and technical quality, that are not captured by structured metadata. We evaluated two sentence embedding models: `thenlper/gte-small` [Li et al. 2023], a lightweight transformer designed for efficiency and scalability; and `all-mpnet-base-v2`[2], a larger model that provides richer semantic representations at a higher computational cost. The overall process follows the BERTopic framework [Grootendorst 2022]. High-dimensional document embeddings are first reduced using UMAP and then clustered with K-Means to group semantically related reviews, prioritizing stable and coherent clusters. Topic representation is derived using a KeyBERT-inspired keyword-extraction method aligned with document embeddings. To improve semantic clarity, domain-specific stopwords are removed, and n-grams up to trigrams are considered. The resulting pipeline produces interpretable topic labels, per-review topic distributions, and low-dimensional visualizations to support subsequent analysis and model integration.

5. Experimental Settings

5.1. Evaluation Methodology

To ensure unbiased evaluation, we use a 5-fold cross-validation strategy with a leave-one-out (LOO) protocol. For each user, one interaction is held out for testing, while the remaining interactions are used for training. This approach simulates realistic next-item recommendations, particularly in sparse datasets, and is repeated across five folds to reduce variance.

For ranking-based metrics, we follow the protocol of sampling 99 negative items per user. To identify the relevant items for each user, we normalize ratings using z-scores. Items with significantly negative normalized scores are treated as negative. This personalized normalization accounts for individual user rating bias and helps reduce label noise during ranking evaluation. These items correspond to unseen or non-relevant interactions and are combined with the held-out positive item to form a candidate set of 100 items [Ma et al. 2025]. The model's ranking quality is evaluated using Hit Rate (HR), Normalized Discounted Cumulative Gain (nDCG), and Mean Reciprocal Rank (MRR). These metrics emphasize the position of the relevant item within the top-K recommendations.

[2]Available at: https://huggingface.co/sentence-transformers/all-mpnet-base-v2

To extract semantic information from reviews, we employ BERTopic, which leverages transformer-based embeddings, clustering, and class-based TF-IDF to generate topic distributions. Each review is represented as a topic vector, and user/item textual profiles are constructed by aggregating the topic distributions of their associated reviews. These topic-aware representations are used as structured semantic signals within the recommendation architecture.

5.2. Datasets

The integration of Amazon Movies and TV, IMDb, and Rotten Tomatoes enables the analysis of RSs from multiple evaluative perspectives. Table 1 summarizes the statistical information of each dataset.

Table 1. Comparative statistics of the datasets before and after preprocessing.

	Amazon Movies and TV		**IMDb**		**Rotten Tomatoes**	
Metric	**Raw**	**Processed**	**Raw**	**Processed**	**Raw**	**Processed**
Reviews	17.3M	1.4M	932.4K	264.9K	1.1M	543.5K
Users	6.5M	76K	427K	7k	11K	2.6k
Items	747.7K	27.3K	1.1K	1.1K	17.7K	17.4K
Reviews/User	2.664	19.12	3.28	37.07	100.06	208.17
Reviews/Item	23.173	53.18	12.28	230.40	63.77	31.13
Sparsity	99.99%	99.93%	99.81%	96.77%	99.42%	98.80%

5.3. Metrics

To assess predictive accuracy in rating estimation, we use standard regression metrics from RS literature [Ricci et al. 2022]. We report MAE and RMSE, which measure the deviation between predicted and actual ratings in the test set. MAE offers a linear view of prediction errors, while RMSE applies a quadratic penalty, emphasizing larger errors and being more sensitive to outliers.

$$_{S}\text{MAE} = \frac{1}{|\mathcal{T}|}\sum_{(u,i)\in\mathcal{T}} |r_{ui} - \hat{r}_{ui}|\,, \qquad \text{RMSE} = \sqrt{\frac{1}{|\mathcal{T}|}\sum_{(u,i)\in\mathcal{T}} (r_{ui} - \hat{r}_{ui})^2}, \qquad MRR = \frac{1}{|U|}\sum_{u\in U}\frac{1}{rank_u}. \quad (10)$$

To evaluate ranking performance under the Leave-One-Out (LOO) protocol, we use HR, nDCG, and MRR, which assess the ordering quality of the top-K recommendations. HR verifies whether the relevant item appears within the top-K. nDCG incorporates the item's rank using a logarithmic discount, emphasizing higher positions.

$$HR@K = \frac{1}{|U|}\sum_{u\in U}\mathbf{I}(rank_u \le K), \qquad nDCG@K = \frac{1}{|U|}\sum_{u\in U}\begin{cases}\dfrac{1}{\log_2(rank_u+1)}, & \text{if } rank_u \le K,\\ 0, & \text{otherwise,}\end{cases} \quad (11)$$

5.4. Hyperparameters analysis

We performed a hyperparameter sensitivity study using Optuna[3], running 20 trials per dataset on a user-stratified 20% subset to reduce computational cost while preserving the

[3]Available at: https://optuna.org/

interaction structure. Each candidate configuration was trained for 15 epochs using the leave-one-out protocol (fold 0), with optimization guided by the validation RMSE loss. The search space covered dropout rates $\{0.2, 0.5\}$, learning rates $\{1e^{-4}, 1e^{-3}\}$, and four embeddings sizes $\{256, 128, 64, 32\}$, to assess the influence of model depth and capacity. The topic latent dimension was kept fixed to ensure fair comparison across trials. For each dataset, the final configuration was chosen based on validation RMSE. The optimal configuration identified across datasets utilized an embedding dimension of 64, a dropout rate of 0.2, and a learning rate of 0.001.

6. Results

Table 2 reveals a clear distinction between prediction accuracy and ranking effectiveness across datasets. For the error metrics (RMSE and MAE), the hybrid architectures consistently outperform the Pure collaborative baseline. In particular, Gated Text achieves the best overall prediction performance on Amazon and IMDb, while Cross Attention attains the lowest error on Rotten Tomatoes. The margin of improvement over the Pure model is especially noticeable on Amazon (RMSE reduction of 0.085) and modest but consistent on Rotten Tomatoes (0.016). These results indicate that incorporating textual representations, especially when directly gated at the embedding level, effectively refines rating estimation. The small standard deviations across folds also suggest that these models perform stably on the regression task.

However, this improvement in rating prediction does not translate to superior ranking quality. Across all three datasets, the Pure model consistently achieves the best HR@10, nDCG@10, and MRR@10. The gap is particularly large on Amazon, where the Pure model substantially outperforms the hybrids in ranking metrics. This pattern suggests that while textual signals help calibrate rating magnitudes, they may dilute the latent collaborative structure that drives effective item ordering. In contrast, the Pure model appears better at capturing global interaction patterns, interpreted as collaborative transitivity, thereby producing stronger top-K recommendations.

Table 2. Performance comparison across datasets. Bold values indicate the best result for each metric within the corresponding dataset. The ± symbol denotes the standard deviation computed across folds.

Model	Dataset	RMSE	MAE	HR@10	NDCG@10	MRR@10
Cross Attention	Amazon	1.026 (± 0.031)	0.738 (± 0.020)	0.104 (± 0.008)	0.046 (± 0.005)	0.051 (± 0.005)
Gated Text	Amazon	**0.985 (± 0.032)**	**0.699 (± 0.018)**	0.135 (± 0.009)	0.062 (± 0.004)	0.064 (± 0.003)
Gated Topic	Amazon	1.043 (± 0.037)	0.770 (± 0.029)	0.205 (± 0.015)	0.114 (± 0.012)	0.107 (± 0.011)
Pure	Amazon	1.070 (± 0.028)	0.828 (± 0.019)	**0.311 (± 0.028)**	**0.187 (± 0.021)**	**0.169 (± 0.018)**
Text Regularizer	Amazon	1.068 (± 0.029)	0.824 (± 0.020)	0.111 (± 0.009)	0.050 (± 0.007)	0.054 (± 0.006)
Topic Regularizer	Amazon	1.069 (± 0.028)	0.825 (± 0.020)	0.112 (± 0.007)	0.050 (± 0.006)	0.055 (± 0.006)
Cross Attention	IMDb	3.311 (± 0.012)	2.948 (± 0.013)	0.200 (± 0.052)	0.095 (± 0.010)	0.085 (± 0.009)
Gated Text	IMDb	**1.976 (± 0.031)**	**1.448 (± 0.030)**	0.172 (± 0.018)	0.083 (± 0.009)	0.082 (± 0.006)
Gated Topic	IMDb	1.986 (± 0.036)	1.463 (± 0.030)	0.175 (± 0.015)	0.085 (± 0.007)	0.083 (± 0.005)
Pure	IMDb	1.988 (± 0.028)	1.467 (± 0.027)	**0.210 (± 0.015)**	**0.105 (± 0.006)**	**0.099 (± 0.004)**
Text Regularizer	IMDb	1.988 (± 0.031)	1.468 (± 0.031)	0.114 (± 0.025)	0.048 (± 0.013)	0.051 (± 0.014)
Topic Regularizer	IMDb	1.990 (± 0.032)	1.468 (± 0.032)	0.114 (± 0.039)	0.048 (± 0.019)	0.050 (± 0.013)
Cross Attention	Rotten Tomatoes	**0.761 (± 0.008)**	**0.598 (± 0.007)**	0.104 (± 0.078)	0.048 (± 0.037)	0.054 (± 0.030)
Gated Text	Rotten Tomatoes	0.777 (± 0.007)	0.610 (± 0.004)	0.256 (± 0.016)	0.128 (± 0.009)	0.116 (± 0.008)
Gated Topic	Rotten Tomatoes	0.775 (± 0.005)	0.611 (± 0.003)	0.272 (± 0.023)	0.143 (± 0.015)	0.130 (± 0.012)
Pure	Rotten Tomatoes	0.772 (± 0.006)	0.610 (± 0.005)	**0.293 (± 0.016)**	**0.149 (± 0.009)**	**0.131 (± 0.007)**
Text Regularizer	Rotten Tomatoes	0.772 (± 0.004)	0.610 (± 0.003)	0.088 (± 0.052)	0.037 (± 0.026)	0.043 (± 0.019)
Topic Regularizer	Rotten Tomatoes	0.772 (± 0.004)	0.610 (± 0.003)	0.088 (± 0.052)	0.037 (± 0.026)	0.043 (± 0.019)

Dataset characteristics also influence model behavior. The IMDb dataset exhibits a markedly different pattern: although Gated Text achieves the best error metrics, the Cross Attention model performs dramatically worse in RMSE and MAE (3.311), diverging from its behavior on the other datasets. This instability may be associated with IMDb's substantially longer average review length. Longer and denser textual sequences increase model complexity and may introduce noise or overfitting in attention-based fusion mechanisms, particularly when textual variance is high. In contrast, shorter, more homogeneous reviews (as on Amazon and Rotten Tomatoes) appear more compatible with attention-based integration.

6.1. Statistical Analysis

To analyze the models' behavior across five evaluation metrics and three datasets, we moved beyond isolated performance values and examined the consistency of their relative rankings. Rather than relying solely on absolute scores, the comparison emphasizes each method's stability across evaluation dimensions and data contexts, providing a more comprehensive assessment of overall robustness and competitiveness. Following the non-parametric statistical procedure proposed by [Demšar 2006], we adopted a global ranking framework combined with hypothesis testing to assess algorithmic superiority.

To consolidate the metrics (RMSE, MAE, HR, nDCG, and MRR) into a single comparative measure, we calculated the indicator Hypervolume (HV). Our evaluation protocol separates error and classification metrics to examine model behavior from different performance perspectives. For statistical comparison, we applied the HV indicator in three configurations: (i) error metrics only, (ii) ranking metrics only, and (iii) all metrics together. This design allows us to assess whether improvements in ranking prediction translate into competitive ranking performance and to evaluate the model's overall balance. This strategy transforms the multi-objective ranking vector into a single scalar value while preserving Pareto dominance relationships, enabling consistent, unbiased comparisons across models within the ranking scenario without conflating them with regression-based performance.

For each dataset–metric context, algorithms were ranked from best to worst to avoid distortions arising from scale differences across domains. We then applied the Friedman test to verify whether ranking differences were statistically significant. Finally, global average ranks were computed, and Critical Difference (CD) diagrams based on the Nemenyi post hoc test were used to identify statistically significant pairwise differences. This integrated procedure ensures that the selected model exhibits consistent, statistically supported performance across all evaluation dimensions.

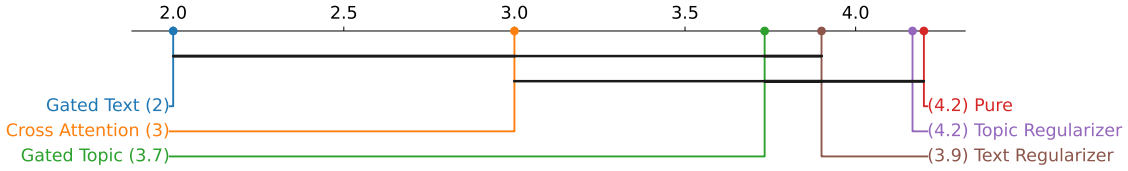

Figure 1. CD diagram based on the Nemenyi post-hoc test, illustrating the average error metrics (RMSE, MAE) of the models across datasets. The models are ordered from left to right. Horizontal bars connect algorithms whose performance differences are not statistically significant at α = 0.05.

Figure 1 presents the CD diagram based on the average rankings across all datasets and folds, considering only the error metrics (RMSE, MAE). The Gated Text (2.0) method achieves the best average ranking (leftmost values), followed by Cross Attention (3.0) and Gated Topic (3.7), forming the main cluster. Although Gated Text achieves the best average ranking, its advantage over Cross Attention and Gated Topic is not consistently statistically significant, suggesting that the leading methods are competitive rather than exhibiting absolute dominance.

Figure 2 presents the calculated CD diagram for the ranking metrics (HR@10, nDCG@10, and MRR@10). In contrast to the behavior observed for the error metrics, the results are now reversed. The Pure model achieves the best (leftmost values) average ranking (1.1). The Gated Topic (2.5) and Gated Text (2.9) models form an intermediate group, while Cross Attention (4.3), Text Regularizer (5.0), and Topic Regularizer (5.1) occupy the weakest positions. The diagram indicates that Pure is statistically competitive with variants under tighter content constraints, but consistently ranks in the leftmost position, reinforcing its dominance in ranking performance. Models based on regularizers form a statistically indistinguishable cluster at the right end, reflecting their consistently inferior ranking behavior.

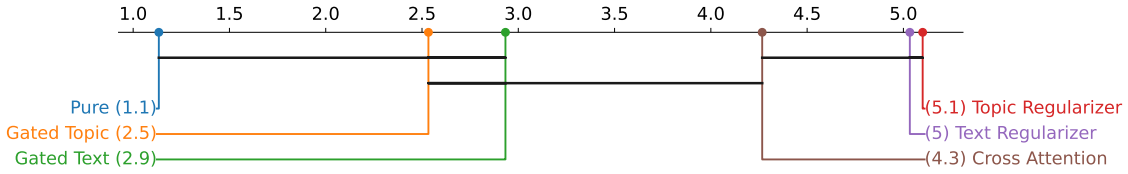

Figure 2. CD diagram for the ranking metrics (HR@10, nDCG@10, and MRR@10). Models are ordered by average rank (lower is better). Horizontal bars connect models whose performance differences are not statistically significant according to the Nemenyi post-hoc test.

6.2. Discussion

Incorporating review information into a matrix factorization framework effectively enhances recommendation quality. The results reveal a consistent pattern: textual augmentation improves rating prediction accuracy, but does not necessarily translate into superior ranking performance. Hybrid models, particularly the Gated Text variant, achieve lower RMSE and MAE across datasets, confirming that textual signals help refine regression estimates. This supports the idea that reviews provide complementary semantic information, thereby reducing prediction error.

In particular, although review-enhanced models introduce additional signals, they tend to reduce stability in the ranking metrics. As a result, the Pure model attains the best overall ranking across datasets and folds. This behavior is confirmed by the statistical comparison presented in Figure 4, which highlights the significant differences among model groups, while Figure 3 provides a detailed view of the results across folds, illustrating the consistent superiority of the Pure model in the aggregated ranking analysis.

However, when the evaluation perspective shifts to ranking metrics, the Pure collaborative model consistently outperforms the others. This indicates that collaborative signals remain more effective for capturing the relational structure necessary to order

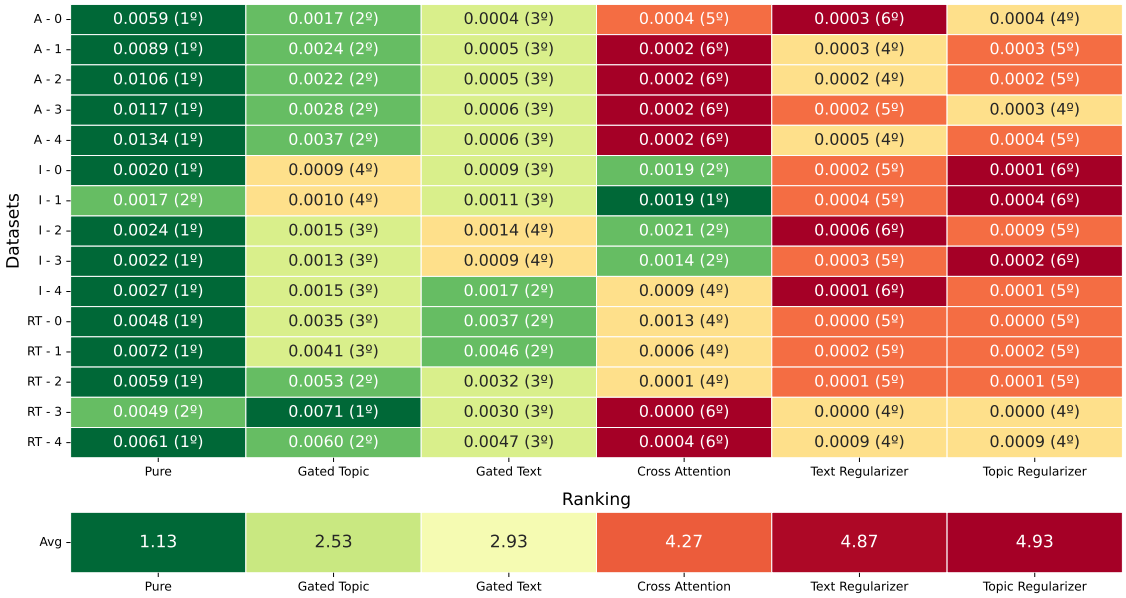

Figure 3. Ranking performance across datasets (A = Amazon, I = IMDb, and R = Rotten Tomatoes) and folds using the hypervolume (HV) indicator computed from all evaluation metrics. Each cell shows the HV value for a model in a dataset–fold combination, with its rank in parentheses (1º = best). The bottom row reports the average rank across all folds and datasets.

relevant items in top-K recommendation tasks. One possible explanation is that emphasizing regression optimization may alter the latent space in ways that improve numerical rating estimation but weaken the transitive structure of user–item interactions. In other words, strengthening pointwise prediction may partially reduce the collaborative propagation effects that benefit ranking.

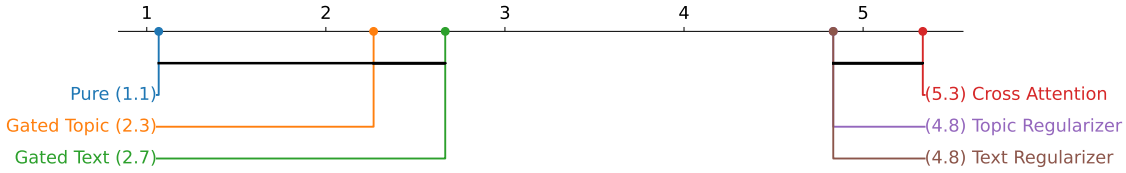

Figure 4. CD diagram summarizing the overall comparison of models across all metrics. Horizontal bars connect groups of models whose differences are not statistically significant according to the Nemenyi post-hoc test.

This behavior may also be interpreted through the lens of *natural noise*. As discussed by [Jurdi et al. 2021], datasets inherently contain variability in user profiles. When reviews are incorporated, particularly in regression-oriented settings, models may become more sensitive to inconsistencies in user ratings. If user ratings do not align perfectly with textual sentiment or behavioral coherence, including textual signals may amplify this variability, improving local prediction fit without strengthening it and potentially diluting the global relational patterns required for ranking.

Therefore, the findings reinforce two important insights. First, improvements in regression metrics do not necessarily imply improvements in ranking quality. Second, collaborative structure alone remains highly robust for capturing global preference tran-

sitivity in top-K recommendation scenarios. These results highlight the need to evaluate recommendation models across multiple metric families and suggest that future research should explicitly balance regression optimization, ranking effectiveness, and robustness to natural noise.

6.3. Limitations and Improvement Points

Despite the promising results, this study presents some limitations that open avenues for further improvement. First, the experimental analysis focuses exclusively on matrix factorization as the collaborative backbone. Although this choice ensures controlled comparisons and methodological clarity, it restricts the generalizability of the findings. Second, the notion of sentiment incorporated into the model was not formally defined theoretically. While sentiment signals were operationalized through extracted review representations, a clearer conceptual and mathematical definition of how sentiment is measured, scaled, and integrated into the recommendation process would strengthen interpretability and reproducibility. Providing a formal definition and detailing its computation would enhance transparency and methodological rigor. Finally, the current approach does not leverage Convolutional Neural Network (CNN) to capture localized, high-impact textual patterns in reviews.

7. Conclusion and Future Works

This work investigated three enrichment strategies for a hybrid matrix factorization model integrating review-based information. The empirical findings consistently indicate that hybrid variants tend to improve regression-oriented metrics, such as RMSE and MAE, while degrading ranking-oriented metrics. This pattern was observed across the Amazon, IMDb, and Rotten Tomatoes datasets, suggesting that the phenomenon is not dataset-specific. Although these results do not constitute definitive evidence, they provide strong empirical indications of a structural trade-off between error minimization and ranking effectiveness in review-aware hybrid models. Statistical hypothesis tests further support this interpretation by revealing two statistically distinct performance groups when models are evaluated under error-based versus ranking-based criteria. These findings reinforce the hypothesis that optimizing for rating prediction accuracy does not necessarily translate into improved item ordering, highlighting a fundamental tension between pointwise and ranking objectives in hybrid recommendation settings.

Future research should extend the proposed enrichment mechanisms beyond matrix factorization to other collaborative paradigms, including graph-based architectures and neural autoencoder frameworks. Exploring whether the observed trade-off persists under alternative modeling assumptions may provide deeper insights into the interaction between textual signals and collaborative representations. Additionally, investigating ranking-aware optimization strategies and more principled integration of textual information may help reconcile regression and ranking performance within unified recommendation frameworks.

References

da Silva, D., Pires, J., and Durão, F. (2025). Exploiting surrogate submodular and cost-effective lazy forward algorithms for calibrated recommendations. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 98–111, Porto Alegre, RS, Brasil. SBC.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7(1):1–30.

Grootendorst, M. (2022). Bertopic: Neural topic modeling with a class-based tf-idf procedure.

Hasan, E., Rahman, M., Ding, C., Huang, J. X., and Raza, S. (2024). Review-based Recommender Systems: A Survey of Approaches, Challenges and Future Perspectives. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, XX(X).

Jurdi, W. A., Abdo, J. B., Demerjian, J., and Makhoul, A. (2021). Critique on natural noise in recommender systems. *ACM Trans. Knowl. Discov. Data*, 15(5).

Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., and Zhang, M. (2023). Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Lima, M., Silva, E., and da Silva, A. (2024). Um estudo sobre o uso de modelos de linguagem abertos na tarefa de recomendação de próximo item. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 510–522, Porto Alegre, RS, Brasil. SBC.

Liu, D., Li, J., Du, B., Chang, J., and Gao, R. (2019). Daml: Dual attention mutual learning between ratings and reviews for item recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 344–352, New York, NY, USA. Association for Computing Machinery.

Liu, Y. and Miyazaki, J. (2023). Knowledge-aware attentional neural network for review-based movie recommendation with explanations.

Ma, H., Xie, R., Meng, L., Feng, F., Du, X., Sun, X., Kang, Z., and Meng, X. (2025). Negative sampling in recommendation: A survey and future directions.

Ricci, F., Rokach, L., and Shapira, B., editors (2022). *Recommender Systems Handbook: Third Edition*. Springer US, United States, 3 edition. Publisher Copyright: © Springer Science+Business Media, LLC, part of Springer Nature 2011, 2015, 2022.

Sachdeva, N. and McAuley, J. (2020). How useful are reviews for recommendation? a critical review and potential improvements. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1845–1848, New York, NY, USA. Association for Computing Machinery.

Silva, E., Pires, J., Dantas, D., Oliveira, M., and Durão, F. (2025a). A gated review attention framework for topics in graph-based recommenders. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web*, pages 19–27, Porto Alegre, RS, Brasil. SBC.

Silva, F., Filho, S. A., Paradeda, R., and Silva, L. (2025b). Análise de sentimentos em avaliações de livros utilizando a api gemini para recomendação personalizada. In *Anais do XL Simpósio Brasileiro de Bancos de Dados*, pages 671–684, Porto Alegre, RS, Brasil. SBC.

Son, J., Kim, H., and Kim, S. W. (2025). Rating-Aware Homogeneous Review Graphs and User Likes/Dislikes Differentiation for Effective Recommendations. *SIGIR 2025 - Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2070–2080.

Tan, C. H., Zheng, H., Wang, J., Lin, Z., Feng, S., Zhan, H., Li, X., and Senthilnath, J. (2025). Do reviews matter for recommendations in the era of large language models? *ArXiv*, abs/2512.12978.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Zheng, L., Noroozi, V., and Yu, P. S. (2017). Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 425–434, New York, NY, USA. Association for Computing Machinery.