Quando Comentários e Notas Discordam: Implicações para Sistemas de Recomendação Baseados em Reviews

Yan Andrade[1], Naan Vasconcelos[2], Adriano Pereira[1], Leonardo Rocha[2]

[1]Universidade Federal de Minas Gerais (UFMG)
[2]Universidade Federal de São João del-Rei (UFSJ)

{yan.ribeiro, adrianoc}@dcc.ufmg.br
naan.vasconcelos@aluno.ufsj.edu.br, lcrocha@ufsj.edu.br

Resumo. *Este trabalho investiga a relação entre notas e comentários no contexto de recomendação, avaliando inconsistências entre essas duas formas de interação, o potencial de notas inferidas a partir do texto e o impacto efetivo dos comentários em modelos Review-Aware Recommender Systems (RARs). Para isso, avaliamos três modelos estado-da-arte em três coleções amplamente utilizadas e empregamos a versão zero-shot do BERT para converter comentários em valores numéricos na mesma escala das notas originais. Os resultados mostram que há discrepâncias relevantes entre nota e comentário, que notas inferidas a partir do texto podem recuperar mais itens relevantes e que os RARs avaliados permanecem fortemente dependentes da nota original.*

Abstract. *This work investigates the relationship between ratings and reviews in the recommendation setting by examining inconsistencies between these two forms of interaction, the potential of review-inferred ratings, and the effective impact of reviews on Review-Aware Recommender Systems (RARs). To this end, we evaluate three state-of-the-art models on three widely used datasets and employ the zero-shot version of BERT to convert reviews into numerical values on the same scale as the original ratings. The results show relevant discrepancies between ratings and reviews, indicate that review-inferred ratings can recover more relevant items, and reveal that the evaluated RARs remain strongly dependent on the original rating.*

1. Introdução

Os Sistemas de Recomendação (SsR) passaram a desempenhar um papel estratégico na mediação do acesso de usuários a conteúdos e serviços. De modo geral, esses sistemas identificam preferências a partir de interações explícitas, como avaliações numéricas e comentários, e de sinais implícitos derivados do comportamento dos usuários, como cliques e tempo de visualização [Ricci et al. 2015]. A efetividade das recomendações depende, em grande medida, da capacidade dos modelos de representar e utilizar adequadamente essas preferências [Werneck et al. 2020].

Nesse contexto, os *Review-Aware Recommender Systems* (RARs) têm se destacado por incorporar comentários textuais ao processo de recomendação, sob a premissa de que o texto expressa percepções mais ricas sobre um item do que avaliações numéricas isoladas [Srifi et al. 2020]. Apesar dos avanços trazidos pelas arquiteturas neurais de aprendizado profundo nessa área [Bittencourt et al. 2026, Liu et al. 2020, Li et al. 2019,

Li et al. 2021], uma limitação central ainda persiste: esses modelos utilizam o comentário apenas como informação auxiliar, enquanto os *ratings* atribuídos pelos usuários permanecem como principal base para o aprendizado das preferências [Liu et al. 2025]. Essa limitação torna-se ainda mais relevante, pois nota e comentário podem não ser consistentes entre si. Usuários frequentemente atribuem avaliações que não refletem o teor do texto associado, seja por ambivalência, seja pela dificuldade de condensar múltiplos aspectos em uma única escala numérica [Ganu et al. 2009, Wu et al. 2024].

Essa possível inconsistência é relevante porque notas e comentários representam formas distintas de interação e, idealmente, deveriam refletir percepções compatíveis sobre o item consumido. Quando isso não ocorre, torna-se menos claro como representar adequadamente as preferências dos usuários. O problema é especialmente importante para os RARs, que buscam justamente incorporar comentários textuais ao processo de recomendação. Por isso, investigar a relação entre notas e comentários é um passo importante para compreender o valor do comentário textual nesse contexto.

Neste trabalho, realizamos uma investigação quantitativa da inconsistência entre as notas atribuídas pelos usuários e os comentários associados, do potencial informativo desses comentários no cenário de recomendação e de seu impacto na modelagem adotada pelos RARs. De forma objetiva, buscamos responder às seguintes perguntas de pesquisa:

- **PP1**: Em que medida há inconsistências entre as notas atribuídas pelos usuários e o conteúdo semântico de seus comentários?
- **PP2**: Qual é o potencial das notas inferidas a partir dos comentários para representar as preferências dos usuários?
- **PP3**: Qual é o impacto do uso de comentários de usuários na modelagem de preferências em modelos RAR?

Para responder a essas perguntas, realizamos uma avaliação experimental detalhada com três modelos RARs considerados estado da arte [Bittencourt et al. 2026] (CARP [Li et al. 2019], CARM [Li et al. 2021] e HRDR [Liu et al. 2020]) e três coleções amplamente utilizadas na área (*Musical Instruments*, *Digital Music* e *Office Products*). Para cada coleção, aplicamos análise de sentimento, mais especificamente a versão zero-shot do BERT [Town 2019], para converter os comentários em valores numéricos na mesma escala das notas atribuídas, de 1 a 5. Os resultados mostraram, em relação à **PP1**, que uma parcela considerável das avaliações apresenta inconsistências entre a nota e o comentário. Em relação à **PP2**, observamos que o uso de notas inferidas a partir dos comentários permitiu recuperar uma proporção maior de itens relevantes. Já para a **PP3**, verificamos que os RARs avaliados permanecem fortemente dependentes da nota original.

2. Trabalhos Relacionados

Os *Review-Aware Recommender Systems* (RARs) incorporam comentários textuais ao processo de recomendação com o objetivo de complementar a informação fornecida pelos *ratings* explícitos [Srifi et al. 2020, Jannach et al. 2010]. Nessa linha, o HRDR [Liu et al. 2020] combina *ratings* e *reviews* por meio de codificadores paralelos, o CARP [Li et al. 2019] explora representações textuais para capturar preferências e gerar explicações, e o CARM [Li et al. 2021] utiliza mecanismos de atenção para integrar informações textuais e colaborativas. Em geral, esses modelos utilizam o texto para enriquecer a representação de usuários e itens dentro da formulação supervisionada tradicional da recomendação [Liu et al. 2025].

Em paralelo, outros trabalhos têm explorado o conteúdo textual dos *reviews* de forma mais direta. Em [Al-Ghuribi et al. 2023], por exemplo, os autores utilizam análise de sentimento para derivar *ratings* implícitos a partir dos comentários e integrá-los a métodos de filtragem colaborativa [Dang et al. 2021]. Nesse contexto, modelos baseados em BERT [Devlin et al. 2019] ampliaram a capacidade de capturar informações contextuais em textos opinativos [Sayeed et al. 2023], favorecendo seu uso em tarefas relacionadas à recomendação [Bittencourt et al. 2026].

Há também trabalhos que discutem a relação entre notas e comentários. Em [Ganu et al. 2009], os autores mostram que a avaliação numérica nem sempre resume o conteúdo expresso no texto. De forma mais recente, [Wu et al. 2024] investigam essa relação por meio de um modelo supervisionado contrastivo. Diferentemente desses trabalhos, nosso objetivo é realizar uma análise mais ampla das notas atribuídas e dos comentários. Empregando análise de sentimento em regime zero-shot para obter, a partir dos comentários, uma avaliação na mesma escala dos *ratings* originais, investigamos três aspectos centrais: as discrepâncias entre nota e comentário, o potencial dos comentários para representar preferências e o impacto efetivo do texto em modelos RARs.

3. Metodologia Experimental

Esta seção descreve os experimentos realizados para avaliar às perguntas de pesquisa.

3.1. Bases de Dados

Neste trabalho, utilizamos três coleções da plataforma Amazon amplamente exploradas na literatura sobre SsR: *Musical Instruments*, *Digital Music* e *Office Products* [Ni et al. 2019]. A seleção dessas bases buscou reunir cenários com propriedades distintas, incluindo diferenças de domínio, variações na densidade das interações e na extensão média dos textos produzidos pelos usuários. Como cada interação contém tanto o *rating* quanto o comentário associado, essas coleções são adequadas para a investigação proposta. A Tabela 1 resume as principais características dos conjuntos utilizados.

Base de Dados	# Usuários	# Itens	# Revisões	Esparsidade
Musical Instruments	27.528	10.620	231.344	99,92%
Digital Music	16.561	11.797	223.445	99,91%
Office Products	101.498	27.965	940.156	99,98%

Tabela 1. Visão geral das coleções utilizadas na avaliação.

3.2. Modelos e Configurações

Para os experimentos com RARs, foram considerados três modelos: CARP [Li et al. 2019], CARM [Li et al. 2021] e HRDR [Liu et al. 2020]. Todos foram treinados do zero com seus hiperparâmetros padrões, conforme descrito nos respectivos trabalhos originais. A escolha desses modelos se deve ao seu desempenho competitivo em avaliações recentes [Bittencourt et al. 2026]. Para inferir notas a partir dos comentários, utilizamos um modelo BERT multilíngue treinado para predição de avaliações de 1 a 5 estrelas [Town 2019][1]. O modelo foi aplicado diretamente aos comentários das coleções, sem *fine-tuning* adicional. A nota associada a cada comentário foi calculada como o valor esperado da distribuição de probabilidades produzida pelo modelo.

[1]https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

3.3. Processo de Avaliação

A avaliação foi organizada em três etapas, uma para cada pergunta de pesquisa.

Na primeira etapa, buscamos quantificar possíveis desalinhamentos entre as notas atribuídas pelos usuários e as inferidas pelo BERT a partir dos comentários. Consideramos inconsistente toda avaliação em que a nota original é positiva (maior ou igual a 3) e a nota inferida é negativa (inferior a 3), ou vice-versa. Essa análise é apresentada de forma quantitativa, por meio da proporção de avaliações inconsistentes em cada base, e de forma qualitativa, com exemplos representativos de cada coleção.

Na segunda etapa, investigamos se o comentário contém informações suficientes para representar as preferências dos usuários de forma independente da nota original. Para comparar diretamente o efeito do uso da nota atribuída e da nota inferida como formas de representação do usuário, sem introduzir o viés do comportamento específico dos modelos de recomendação, definimos uma métrica de *cobertura de vizinhança*. Essa métrica quantifica em que proporção os itens vizinhos aos itens mais bem avaliados por um usuário correspondem a itens com os quais ele de fato interagiu na base. Para cada usuário u, identificamos os $k = 5$ itens com as maiores notas em cada cenário avaliado (nota original e nota inferida), formando o conjunto T_u. Para cada item $i \in T_u$, recuperamos seus $n = 5$ vizinhos mais próximos com base na similaridade do cosseno, considerando que cada item é representado pelo vetor de notas recebidas dos usuários, seja no cenário com notas originais, seja no cenário com notas inferidas. O conjunto de vizinhos de todos os itens de T_u constitui o conjunto de candidatos C_u, com $|C_u| = 25$. A cobertura de vizinhança do usuário u é definida como:

$$\text{Cov}(u) = \frac{|\{i \in C_u : i \in H_u\}|}{|C_u|} \tag{1}$$

em que H_u representa o conjunto de itens com os quais o usuário u interagiu na base de dados. A métrica final corresponde à média de $\text{Cov}(u)$ sobre todos os usuários. Valores mais altos indicam que a representação adotada induz uma estrutura de vizinhança mais próxima do comportamento real dos usuários.

Na terceira etapa, investigamos se os modelos RARs são capazes de explorar efetivamente o conteúdo textual dos comentários, especialmente em cenários de possível inconsistência com a nota original. Para isso, cada um dos três modelos selecionados (CARP, CARM e HRDR) foi treinado em dois cenários: (i) com as notas originais e (ii) com todas as notas substituídas pelo valor constante r=3, mantendo os comentários inalterados em ambos os casos. O objetivo do segundo cenário é neutralizar o efeito informacional das notas, permitindo avaliar em que medida os comentários, isoladamente, influenciam as predições dos modelos. A comparação entre os dois cenários é realizada com base na análise das distribuições das notas preditas pelos RARs.

4. Resultados Experimentais

4.1. PP1: Inconsistências entre Nota e Comentário

Os resultados da primeira etapa da avaliação evidenciam inconsistências entre a nota atribuída pelos usuários e a nota inferida pelo BERT a partir dos comentários. Aproximadamente 7,79% das avaliações em *Musical Instruments*, 5,98% em *Digital Music* e 7,84% em *Office Products* apresentaram desalinhamento entre essas duas formas de sinalização de preferência. Embora essas ocorrências não representem a maioria das interações, elas

indicam que uma parcela relevante das avaliações utilizadas no treinamento de modelos de recomendação pode estar em conflito com o conteúdo semântico expresso no texto, o que pode comprometer a modelagem das preferências dos usuários.

A Tabela 2 apresenta seis exemplos representativos de inconsistência, dois por coleção. Os casos ilustram tanto situações em que o usuário atribuiu nota elevada a um comentário claramente negativo quanto o cenário inverso, em que a nota foi baixa apesar de o texto expressar satisfação. Em ambos os casos, a nota inferida pelo BERT mostra-se mais compatível com o teor do comentário do que a nota original, reforçando a hipótese de que nota e comentário nem sempre refletem a mesma percepção sobre o item.

Dataset	Comentário	Nota Original	Nota BERT
Musical Instruments	The second night I had it, "e"string broke. Had a new string on hand; replaced it...	4.0	1.58
Musical Instruments	Great cables - fast service	1.0	4.75
Digital Music	Did not no this was so gross or I would not have bought it....was looking for some jazz.	5.0	1.54
Digital Music	If you ever go to a country music bar, you'll probably hear this song, a real classic! Love this song!	1.0	4.87
Office Products	Had to return printer died	5.0	1.32
Office Products	Great for the price! Would order again	1.0	4.54

Tabela 2. Exemplos representativos de inconsistência entre a nota original e o sentimento expresso no comentário, agrupados por dataset.

4.2. PP2: O Comentário como Fonte Primária de Feedback

Para avaliar se os comentários carregam informação suficiente para representar as preferências dos usuários de forma independente da nota original, comparamos as notas atribuídas pelos usuários com as notas inferidas pelo BERT por meio da métrica de cobertura de vizinhança, definida na Equação 1. A comparação foi realizada em dois cenários: (i) utilizando as notas originais e (ii) utilizando as notas inferidas a partir dos comentários. Os resultados são apresentados na Tabela 3. As notas inferidas pelo BERT apresentaram cobertura superior em todas as coleções avaliadas, com ganhos de 7,63% em *Musical Instruments*, 4,00% em *Digital Music* e 2,31% em *Office Products*. Em todos os casos, as diferenças foram estatisticamente significativas pelo teste de Wilcoxon [Wilcoxon 1992]. Esses resultados indicam que, ao utilizar as notas inferidas diretamente dos comentários, obtém-se uma estrutura de vizinhança de itens mais aderente ao comportamento real dos usuários do que a induzida pelas notas originais. Esse achado sugere que o comentário textual contém informação relevante sobre as preferências dos usuários que a avaliação numérica nem sempre consegue captar.

Coleção	Musical Instruments	Digital Music	Office Products
Notas Originais	0,1377	0.1254	0.0979
Notas BERT	0,1482 ▲	0.1304 ▲	0.1002 ▲

Tabela 3. Cobertura de vizinhança média nos dois cenários avaliados. Resultados são avaliados com teste de Wilcoxon ($p = 0{,}05$). Símbolos: ▲ (ganho significativo), • (empate estatístico), ▼ (perda significativa).

4.3. PP3: Sensibilidade dos RARs à Nota Original

Com relação à terceira etapa da avaliação, a Figura 1 apresenta a distribuição das predições dos modelos para a coleção *Musical Instruments* em dois cenários: com as notas originais e com todas as notas fixadas em $r{=}3$. No primeiro caso, as predições apresentam uma distribuição mais ampla, refletindo a variabilidade natural dos dados. No segundo, observa-se que as predições de todos os modelos convergem fortemente para o valor fixado, com variância próxima de zero, mesmo com os comentários mantidos

inalterados. Resultados semelhantes foram observados nas demais coleções, embora não sejam apresentados por limitação de espaço. Em conjunto, esses achados indicam que os modelos avaliados utilizam o texto predominantemente como sinal auxiliar para aproximar a nota original, e não como fonte principal de informação sobre as preferências dos usuários. Essa dependência constitui uma limitação importante, sobretudo em cenários em que há inconsistência entre a nota atribuída e o conteúdo semântico do comentário.

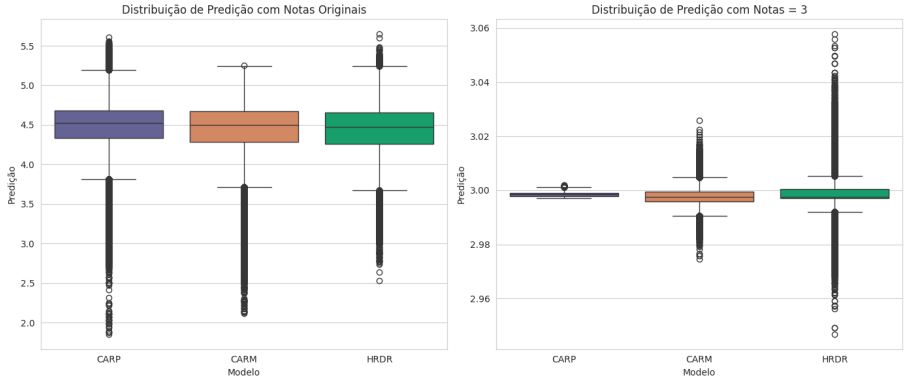

Figura 1. Distribuição das predições dos modelos RARs com notas originais (esquerda) e notas fixadas em $r = 3$ (direita) para a coleção *Musical Instruments*.

4.4. Síntese dos Resultados

De forma conjunta, os resultados revelam três achados principais. Primeiro, uma parcela não desprezível das avaliações apresenta desalinhamento entre nota e comentário, o que indica que essas duas formas de interação nem sempre expressam a mesma percepção do usuário sobre o item. Segundo, quando utilizadas como base para representar preferências, as notas inferidas a partir dos comentários apresentam cobertura de vizinhança superior à das notas originais em todas as coleções avaliadas, sugerindo que o texto captura sinais relevantes que a nota atribuída nem sempre preserva. Por fim, os modelos RARs analisados permanecem fortemente dependentes da nota original, empregando o comentário principalmente como informação auxiliar. Em conjunto, esses resultados apontam para a necessidade de investigar arquiteturas que tratem os comentários textuais de forma mais central no processo de recomendação.

5. Conclusão e Direções Futuras

Este trabalho evidencia uma limitação importante dos *Review-Aware Recommender Systems*: embora incorporem comentários textuais, esses modelos ainda permanecem estruturados em torno da nota explícita como principal sinal de preferência. Ao problematizar a relação entre nota e comentário, mostramos que essa dependência merece ser revista e que o texto pode desempenhar um papel mais informativo do que o habitualmente assumido. Esse resultado abre espaço para uma agenda de pesquisa voltada a modelos que tratem o comentário textual de forma mais central na representação de preferências, especialmente em cenários de inconsistência com a nota original.

Agradecimentos

Este trabalho foi apoiado por CNPq, Capes, Fapemig, Fapesp, AWS e INCT-TILD-IAR.

Declaração de Uso de IA Generativa.

Ferramentas de IA foram utilizadas apenas como apoio à revisão gramatical; os autores revisaram todo o texto e assumem responsabilidade pela integridade das informações.

Referências

Al-Ghuribi, S., Mohd Noah, S. A., and Mohammed, M. (2023). An experimental study on the performance of collaborative filtering based on user reviews for large-scale datasets. *PeerJ Comput Sci*, 9:e1525.

Bittencourt, G., Vasconcelos, N., Andrade, Y., Silva, N., Cunha, W., Colombo Dias, D. R., Gonçalves, M. A., and Rocha, L. (2026). Review-aware recommender systems (rarss): Recent advances, experimental comparative analysis, discussions, and new directions. *ACM Comput. Surv.*, 58(1).

Dang, C., Moreno García, M., and De La Prieta, F. (2021). An approach to integrating sentiment analysis into recommender systems. *Sensors*, 21:5666.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Ganu, G., Elhadad, N., and Marian, A. (2009). Beyond the stars: improving rating predictions using review text content. In *12th International Workshop on the Web and Databases (WebDB)*, pages 1–6.

Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G. (2010). *Recommender systems: an introduction*. Cambridge University Press.

Li, C., Quan, C., Peng, L., Qi, Y., Deng, Y., and Wu, L. (2019). A capsule network for recommendation and explaining what you like and dislike. In *Proceedings of the 42nd ACM SIGIR*, SIGIR'19.

Li, D., Liu, H., Zhang, Z., Lin, K., Fang, S., Li, Z., and Xiong, N. N. (2021). Carm: Confidence-aware recommender model via review representation learning and historical rating behavior in the online platforms. *Neurocomputing*, 455:283–296.

Liu, H., Wang, Y., Peng, Q., Wu, F., Gan, L., Pan, L., and Jiao, P. (2020). Hybrid neural recommendation with joint deep representation learning of ratings and reviews. *Neurocomputing*, 374:77–85.

Liu, J., Li, T., Yu, M., Yang, S., Tang, Z., and Yang, Z. (2025). A multi-factor collaborative prediction for review-based recommendation. In *Proceedings of the Nineteenth ACM RecSys*.

Ni, J., Li, J., and McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings o EMNLP-IJCNLP*.

Ricci, F., Rokach, L., and Shapira, B. (2015). Recommender systems: introduction and challenges. *Recommender systems handbook*, pages 1–34.

Sayeed, M. S., Roji, V., and Anbananthen, K. (2023). Bert: A review of applications in sentiment analysis. *HighTech and Innovation Journal*, 4:453–462.

Srifi, M., Oussous, A., Ait Lahcen, A., and Mouline, S. (2020). Recommender systems based on collaborative filtering using review texts—a survey. *Information*, 11(6):317.

Town, N. (2019). bert-base-multilingual-uncased-sentiment.

Werneck, H., Silva, N., Viana, M. C., Mourão, F., Pereira, A. C., and Rocha, L. (2020). A survey on point-of-interest recommendation in location-based social networks. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, pages 185–192.

Wilcoxon, F. (1992). *Individual Comparisons by Ranking Methods*, pages 196–202. Springer New York, New York, NY.

Wu, H., Guo, G., Yang, E., Luo, Y., Chu, Y., Jiang, L., and Wang, X. (2024). Pesi: Personalized explanation recommendation with sentiment inconsistency between ratings and reviews. *Knowledge-Based Systems*, 283:111133.