

Entre Nota e Opinião: Análise de Sentimento em Reviews para Recomendação Colaborativa

Naan Vasconcelos¹, Yan Andrade², Elisa Tuler¹, Frederico A. Durão³, Adriano C. M. Pereira²,
Leonardo Rocha¹

¹Universidade Federal de São João del-Rei, ²Universidade Federal de Minas Gerais, ³ Universidade Federal da Bahia,
naan.vasconcelos@aluno.ufsj.edu.br; yan.ribeiro@dcc.ufmg.br; etuler@ufsj.edu.br
fdurao@ufba.br; adrianoc@dcc.ufmg.br; lcrocha@ufsj.edu.br

ABSTRACT

Collaborative recommender systems often treat explicit *ratings* as reliable signals of user preference. However, a numerical score may compress complex experiences and fail to reflect the opinion expressed in a textual *review*. This paper investigates the divergence between *ratings* and *reviews* and analyzes whether sentiment-inferred scores can be used to semantically recalibrate the feedback matrix used by collaborative recommendation models. We apply three sentiment analysis strategies, a lexical approach (VADER) and two approaches based on pre-trained language models (BERT and BART) to infer alternative scores from *reviews* and compare them with the original *ratings*. We then replace the original *ratings* with sentiment-inferred scores and evaluate four collaborative algorithms: UserKNN, ItemKNN, SVD, and EASE. Experiments are conducted on four datasets from two domains, e-commerce and Points of Interest (POI), using two Amazon collections and two Yelp subsets. The results show that *rating-review* divergences occur in all datasets and are more pronounced in collections with longer *reviews*. Replacing original *ratings* with sentiment-inferred scores improves recommendation effectiveness across several scenarios, especially for global collaborative models such as SVD and EASE, whereas neighborhood-based methods exhibit less stable behavior. However, these gains often come at the expense of diversity and novelty, revealing a trade-off between semantic feedback alignment and recommendation variety. Overall, this study shows that textual *reviews* should not be treated only as auxiliary features for predicting *ratings*, but also as independent signals that can question, complement, or recalibrate explicit feedback in collaborative recommendation.

KEYWORDS

recommender systems, sentiment analysis, collaborative filtering, reviews, rating-review divergence, beyond-accuracy evaluation

1 INTRODUÇÃO

Sistemas de recomendação desempenham um papel central no acesso à informação em ambientes digitais. Em cenários como o comércio eletrônico [20], plataformas de mídia [11], redes sociais [12] e serviços de entretenimento [2], esses sistemas ajudam os usuários a lidar com grandes volumes de conteúdo, produzindo listas personalizadas de itens potencialmente relevantes [37]. Entre as estratégias mais utilizadas, destacam-se os modelos colaborativos,

que aprendem padrões de preferência a partir de interações históricas entre usuários e itens [28, 29]. Esses métodos permanecem amplamente adotados por sua simplicidade, eficiência e capacidade de capturar relações estruturais em matrizes usuário-item [26].

Grande parte desses modelos utiliza *ratings* explícitos como principal sinal de preferência. Nesse cenário, a nota numérica atribuída pelo usuário é frequentemente tratada como uma representação direta de sua experiência com o item [5]. No entanto, essa premissa pode ser limitada. Um *rating* concentra em um único valor uma experiência que pode envolver múltiplos aspectos, percepções ambíguas, críticas parciais e graus distintos de satisfação [38]. Assim, embora o *rating* seja um sinal útil e de fácil utilização, ele nem sempre captura a complexidade da opinião do usuário [6].

Reviews textuais oferecem uma fonte complementar para observar essa complexidade, revelando justificativas, críticas e aspectos específicos que o *rating* tende a ocultar [18, 21]. Por isso, diferentes trabalhos incorporam esses textos à recomendação, seja via sistemas sensíveis a *reviews* (*Review-Aware Recommender Systems – RARS*) [3, 7, 31], seja pela extração de características semânticas e sentimentos [34, 38]. Contudo, a maioria dessas abordagens utiliza o texto apenas como informação auxiliar para refinar a predição da própria nota. Assim, o *rating* original permanece como alvo supervisionado, e as representações extraídas são otimizadas estritamente para mimetizá-lo.

Este trabalho parte de uma perspectiva distinta: em vez de usar o *review* para prever ou explicar a nota original, investigamos o texto como um sinal independente de preferência. Essa abordagem assume relevância porque *rating* e *review* frequentemente divergem, como quando notas extremas coexistem com críticas ou elogios textuais opostos [22]. Nesses cenários, basear-se puramente no *rating* mascara conflitos entre a avaliação numérica e a opinião textual [4]. Assim, essa divergência deixa de ser tratada como ruído e passa a ser um objeto de análise central para avaliar a confiabilidade do *feedback* em recomendação.

Com base nessa premissa, propomos uma avaliação controlada na qual modelos de análise de sentimento inferem uma nota semântica alternativa a partir dos *reviews*. Essa nota é escalonada e utilizada para substituir o *rating* original na matriz de interações. Assim, investigamos não apenas a divergência entre o sentimento textual e a nota atribuída, mas também o comportamento de modelos colaborativos ao serem treinados com esse *feedback* recalibrado. Desse modo, avaliamos se a opinião expressa no texto altera a modelagem de preferências e a geração de recomendações.

Para orientar essa investigação, definimos três perguntas de pesquisa respondidas ao longo deste artigo:

- **PP1:** Em que medida há divergências entre os *ratings* atribuídos pelos usuários e os sentimentos expressos em seus *reviews*?
- **PP2:** Qual é o impacto da utilização de notas inferidas a partir dos *reviews* no desempenho de modelos colaborativos em métricas de efetividade?
- **PP3:** Como o uso de notas inferidas a partir dos *reviews* afeta propriedades das recomendações, além da efetividade, como a diversidade e a novidade?

Para responder a essas perguntas, avaliamos três estratégias de análise de sentimento não supervisionadas: uma abordagem lexical (VADER) e duas baseadas em modelos de linguagem pré-treinados (BERT e BART), combinadas com quatro algoritmos colaborativos: UserKNN, ItemKNN, SVD e EASE. Os experimentos são conduzidos em quatro coleções de dados, cobrindo dois cenários de aplicação: comércio eletrônico, com duas bases da Amazon, e Pontos de Interesse (POI), com dois subconjuntos do Yelp. A avaliação considera métricas de efetividade, como Precision, Recall, Hit Rate e NDCG, além de métricas *beyond-accuracy*, como diversidade e novidade.

Os resultados mostram que divergências entre *ratings* e notas inferidas ocorrem em todos os conjuntos de dados, sendo mais expressivas em bases com *reviews* longos. Esse achado sugere que textos extensos comportam nuances e opiniões mistas sub-representadas na nota numérica. Ao substituir os *ratings* originais pelas notas semânticas, a efetividade da recomendação melhorou em diversos cenários, especialmente para modelos globais como SVD e EASE. Em contrapartida, métodos baseados em vizinhança (UserKNN e ItemKNN) apresentaram menor estabilidade, oscilando entre ganhos pontuais, empates estatísticos e perdas.

Os resultados também evidenciam um *trade-off* importante: o uso de notas inferidas a partir dos *reviews* tende a reduzir a diversidade e a novidade. Esse comportamento sugere que alinhar o *feedback* ao sentimento textual pode tornar as recomendações mais precisas, mas também mais concentradas. Assim, a inconsistência entre *rating* e *review*, embora possa prejudicar a fidelidade do *feedback*, também pode introduzir variação útil na matriz de interações. Essa observação reforça que *ratings* e sentimentos textuais não devem ser vistos simplesmente como sinais redundantes, mas como fontes complementares e, em alguns casos, divergentes de preferência.

As principais contribuições deste trabalho são: (i) uma análise empírica da divergência entre *ratings* explícitos e sentimentos extraídos de *reviews* em quatro coleções e dois domínios; (ii) uma avaliação controlada do uso de notas inferidas por análise de sentimento como substitutas dos *ratings* originais em modelos colaborativos; (iii) a demonstração de que o impacto desse sinal depende da família de recomendadores, favorecendo principalmente modelos globais; e (iv) a identificação de um *trade-off* entre efetividade e diversidade, indicando que a recalibração semântica do *feedback* pode melhorar o ranqueamento, mas reduzir a variedade das recomendações.

2 TRABALHOS RELACIONADOS

A proposta deste trabalho se apoia em três eixos da literatura: o uso de *ratings* como sinal explícito de preferência em sistemas de recomendação, a incorporação de *reviews* textuais e de análise de sentimento ao processo de recomendação, e os modelos colaborativos utilizados para aprender padrões de interação em matrizes

usuário-item. Discutimos esses eixos, destacando a lacuna explorada neste estudo: investigar o *review* como um sinal independente, capaz de confirmar e complementar o *rating* atribuído pelo usuário.

2.1 *Ratings, reviews* e sinais de preferência

Modelos colaborativos de recomendação tradicionalmente utilizam *ratings* explícitos como principal evidência de preferência dos usuários [5]. Embora essa abordagem simplifique o aprendizado ao mapear a experiência do usuário em valores numéricos padronizados, ela omite nuances importantes. Um *rating* discreto pode ocultar avaliações ambíguas, percepções contraditórias ou aspectos específicos da experiência [14, 38]. Assim, esse sinal numérico isolado frequentemente falha em refletir a real e completa opinião do usuário.

Em contrapartida, *reviews* textuais oferecem uma visão detalhada dessa preferência, permitindo registrar justificativas e aspectos específicos do item [21, 24]. Por isso, são muito explorados para extrair preferências, explicações ou sentimentos [34, 38]. Contudo, a relação entre *rating* e *review* nem sempre é direta: divergências ocorrem quando notas altas acompanham críticas textuais, ou vice-versa [4]. Essa inconsistência sugere que o *review* atua não apenas como explicação da nota, mas como um sinal independente de preferência.

É nesse ponto que este trabalho se diferencia. Em vez de assumir que o *rating* é sempre o alvo mais confiável, investigamos empiricamente a distância entre o *rating* atribuído e a nota inferida a partir do sentimento expresso no *review*. Essa análise permite tratar a divergência entre *rating* e *review* como um fenômeno relevante para a literatura de recomendação, associado à confiabilidade do *feedback* explícito utilizado no treinamento dos modelos.

2.2 Análise de sentimento e RARS

A análise de sentimento tem sido amplamente utilizada para extrair a polaridade e a intensidade subjetiva de textos opinativos [10, 24]. Em sistemas de recomendação, essa técnica permite transformar *reviews* em sinais estruturados, capazes de representar satisfação, insatisfação ou graus intermediários de avaliação. Métodos lexicais, como VADER, utilizam regras e dicionários de polaridade para estimar o sentimento de textos curtos [15]. Embora simples e eficientes, essas abordagens podem apresentar limitações diante de fenômenos linguísticos mais complexos, como a ambiguidade e a ironia [10].

Modelos *Transformers* ampliaram a representação contextual em tarefas de linguagem. O BERT processa o texto de forma bidirecional, capturando o contexto anterior e posterior simultaneamente [13], o que impulsionou seu uso em classificação de sentimento e recomendação, como no BERT4Rec [33]. Por sua vez, arquiteturas como o BART, combinadas a formulações *zero-shot* (e.g., baseadas em MNLI), permitem classificar textos sem a necessidade de ajuste fino por domínio [19, 36]. Essas abordagens viabilizam a estimativa de uma nota semântica a partir do *review*, dispensando classificadores supervisionados previamente treinados na base.

Paralelamente, os sistemas de recomendação sensíveis a *reviews* (*Review-Aware Recommender Systems* – RARS) incorporam textos opinativos ao processo preditivo [3, 7–9, 31], reconhecendo que os textos trazem informações ausentes nos *ratings*. Contudo, a maioria dessas abordagens trata o texto apenas como fonte auxiliar para refinar a predição da nota explícita. Assim, o *rating* original permanece como o alvo supervisionado e as representações textuais terminam otimizadas estritamente para mimetizá-lo.

Este trabalho adota uma perspectiva complementar a essa literatura, tratamos o sentimento expresso no *review* como um sinal independente e avaliamos o que ocorre quando esse sinal substitui o *rating* na matriz de interações. Essa formulação permite investigar uma questão pouco explorada em RARS: o que acontece quando o *review* não confirma a nota atribuída pelo usuário?

2.3 Modelos colaborativos

A filtragem colaborativa baseia-se na premissa de que padrões de interação revelam preferências latentes, independentemente do conteúdo dos itens [1, 17]. Entre os métodos clássicos, os modelos de vizinhança exploram similaridades diretas na matriz usuário-item: o UserKNN foca no comportamento de usuários semelhantes [25], enquanto o ItemKNN avalia a coocorrência de itens [27]. Por operar diretamente sobre as interações, esses métodos de memória tendem a ser altamente sensíveis a variações locais nos *ratings*.

Outra linha importante é a fatoração de matrizes. O SVD projeta usuários e itens em um espaço latente de menor dimensionalidade, permitindo capturar padrões implícitos de preferência a partir da decomposição da matriz de interações [16]. Diferentemente dos métodos de vizinhança, modelos de fatores latentes aprendem representações globais, nas quais alterações pontuais nos valores de entrada podem ser absorvidas por uma estrutura de otimização mais ampla.

Recentemente, modelos lineares têm demonstrado alto desempenho em recomendação, rivalizando com arquiteturas de *Deep Learning* [23]. O EASE (*Embarrassingly Shallow Autoencoders*) é o principal expoente dessa abordagem, formulando o problema como uma reconstrução linear com solução analítica e regularização L_2 [32]. Essa estrutura permite aprender relações item-item de maneira eficiente, servindo como uma linha de base robusta para avaliar o impacto de novos sinais de entrada na matriz de interações.

A comparação entre UserKNN, ItemKNN, SVD e EASE é particularmente útil para os objetivos deste trabalho, pois permite observar como diferentes famílias de recomendadores reagem à substituição do *rating* original por uma nota inferida a partir do *review*. Assim, a análise não se limita a perguntar se o sentimento melhora a recomendação, mas investiga em quais condições e para quais tipos de modelos a recalibração semântica do *feedback* pode ser benéfica.

2.4 Lacuna e posicionamento deste trabalho

A literatura existente mostra que *reviews* podem enriquecer sistemas de recomendação e que técnicas de análise de sentimento permitem extrair sinais estruturados a partir de textos opinativos. Também há uma ampla tradição de modelos colaborativos capazes de aprender preferências a partir de matrizes usuário-item. No entanto, a maior parte dos trabalhos trata o *rating* explícito como o alvo principal a ser predito ou explicado, mesmo quando utiliza o *review* como fonte adicional de informação.

Diante dessa lacuna, buscamos contribuir para a literatura de recomendação em três direções. Primeiro, analisamos empiricamente a divergência entre *ratings* e *reviews* em diferentes domínios. Segundo, avaliamos como essa substituição afeta modelos colaborativos com diferentes mecanismos de aprendizado. Terceiro, examinamos não apenas métricas de efetividade, mas também diversidade e novidade, permitindo compreender os *trade-offs* envolvidos no uso de sentimentos textuais como *feedback* alternativo.

3 METODOLOGIA

Esta seção descreve a metodologia adotada para investigar se *reviews* textuais podem fornecer um sinal de preferência alternativo aos *ratings* explícitos em modelos colaborativos de recomendação. O desenho experimental foi construído para isolar o efeito da substituição do *rating* original por uma nota inferida a partir do sentimento expresso no *review*. Para isso, mantemos constantes as bases de dados, os algoritmos de recomendação, o protocolo de particionamento e as métricas de avaliação, variando apenas o sinal utilizado como valor da interação usuário-item. O fluxo metodológico compreende cinco etapas: (i) seleção e pré-processamento das bases de dados; (ii) inferência de notas semânticas a partir dos *reviews*; (iii) construção de matrizes alternativas de interação; (iv) treinamento e avaliação de modelos colaborativos; e (v) comparação dos resultados em termos de métricas de efetividade, diversidade e novidade.

3.1 Bases de Dados

Os experimentos foram conduzidos em quatro bases de dados, cobrindo dois cenários de aplicação: comércio eletrônico e Pontos de Interesse (POI). No cenário de comércio eletrônico, utilizamos duas coleções do repositório público de avaliações da Amazon¹: *Digital Music* e *Musical Instruments*. Para o cenário de POI, utilizamos dois subconjuntos do Yelp², filtrados pelas cidades de *Philadelphia* e *Tucson*. Essas bases foram selecionadas por conterem, para cada interação, ao menos quatro informações essenciais para este estudo: o usuário, o item, o *rating* explícito e o *review* textual associado. Essa estrutura permite comparar diretamente a nota atribuída pelo usuário com a inferida a partir do conteúdo textual do *review*, além de avaliar como diferentes sinais de preferência afetam o treinamento de modelos colaborativos. Para reduzir cenários extremos de *cold-start* e garantir um volume mínimo de interações por entidade, aplicamos uma filtragem por *k-core*, mantendo apenas usuários e itens com pelo menos cinco interações registradas. Após a filtragem, as estatísticas descritivas das coleções são apresentadas na Tabela 1.

Base de Dados	# Usuários	# Itens	# Revisões	Esparsidade	Mediana (Caract.)
Philadelphia	21.857	14.025	475.224	99,84%	267
Tucson	8.701	8.864	185.774	99,75%	279
Musical Instruments	27.528	10.620	231.344	99,92%	139
Digital Music	16.561	11.797	169.623	99,91%	68

Tabela 1: Estatísticas gerais das coleções utilizadas.

3.2 Inferência de Notas a partir dos Reviews

Para tratar o *review* como um sinal independente de preferência, aplicamos três estratégias de análise de sentimento aos textos das avaliações. O objetivo dessa etapa é transformar cada *review* em uma nota inferida, compatível com a escala dos *ratings* explícitos. Assim, para cada interação usuário-item, mantemos o *rating* original e geramos três valores alternativos, um para cada modelo de análise de sentimento.

Seja r_{ui} o *rating* original atribuído pelo usuário u ao item i , e seja t_{ui} o *review* textual correspondente. Cada modelo de análise de sentimento recebe t_{ui} como entrada e retorna uma nota inferida $\hat{r}_{ui}^m$, em que $m \in \text{VADER, BERT, BART}$. Todas as notas inferidas são normalizadas na escala de 1 a 5, o que permite sua substituição direta na matriz de interações.

¹<https://jmcauley.ucsd.edu/data/amazon/>

²<https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset>

VADER. O VADER é uma abordagem lexical baseada em regras, que retorna um escore composto c no intervalo $[-1, 1]$. Para compatibilizar esse valor com a escala dos *ratings*, aplicamos uma transformação linear para o intervalo $[1, 5]$:

$$\hat{r}_{u,i}^{vader} = \left(\frac{c + 1}{2} \right) \times 4 + 1. \quad (1)$$

BERT. Para o BERT, utilizamos o modelo *bert-base-multilingual-uncased-sentiment*, previamente treinado para classificação de avaliações em cinco níveis. Em vez de utilizar apenas a classe mais provável, calculamos a nota inferida como o valor esperado da distribuição de probabilidades retornada pelo modelo:

$$\hat{r}_{u,i}^{bert} = \sum_{j=1}^5 p_j \cdot j, \quad (2)$$

em que p_j representa a probabilidade atribuída pelo modelo à classe de j estrelas. Essa formulação preserva a incerteza do classificador e permite gerar valores contínuos, como 3,8 ou 4,2.

BART. Para o BART, utilizamos uma configuração *zero-shot* com o modelo *bart-large-mnli*. Definimos cinco hipóteses candidatas associadas à escala de avaliação: *terrible* (1), *poor* (2), *average* (3), *good* (4) e *excellent* (5). O modelo estima a probabilidade de implicação entre o *review* e cada hipótese. A nota inferida é calculada como a média ponderada dessas probabilidades:

$$\hat{r}_{u,i}^{bart} = \sum_k p_k \cdot w_k, \quad (3)$$

em que p_k é a probabilidade atribuída à hipótese k e w_k é o peso numérico associado a essa hipótese.

3.3 Construção das Matrizes de Interação

Com base nas notas inferidas, construímos quatro matrizes de interação para cada base de dados. A primeira corresponde à matriz original, composta pelos *ratings* explícitos atribuídos pelos usuários. As outras três matrizes substituem o *rating* original pela nota inferida a partir do *review*, utilizando, respectivamente, VADER, BERT e BART. Formalmente, para cada interação observada (u, i) , a matriz original contém o valor r_{ui} , enquanto cada matriz semântica contém o valor $\hat{r}_{ui}^m$. A estrutura da matriz permanece inalterada: os mesmos usuários, itens e interações são preservados. Apenas o valor associado a cada interação é substituído. Essa decisão permite comparar os modelos de recomendação sob condições controladas, isolando o impacto do sinal de preferência utilizado no treinamento.

Essa etapa é central para o objetivo do trabalho. Diferentemente de abordagens que usam o *review* como atributo adicional ou como entrada auxiliar para prever o *rating*, nossa metodologia avalia o efeito de substituir diretamente o *rating* por uma nota derivada do conteúdo textual. Assim, investigamos se o sentimento do *review* altera a forma como os modelos colaborativos aprendem preferências.

3.4 Modelos de Recomendação

Para avaliar o impacto das notas inferidas em diferentes recomendadores colaborativos, selecionamos quatro algoritmos: UserKNN, ItemKNN, SVD e EASE, permitindo comparar métodos baseados em vizinhança, modelos de fatores latentes e modelos lineares globais.

UserKNN e ItemKNN. Os modelos baseados em vizinhança utilizam similaridade de cosseno para a escolha dos vizinhos. O UserKNN estima recomendações a partir de usuários com padrões

de interação semelhantes aos do usuário-alvo, enquanto o ItemKNN explora similaridades entre itens avaliados de forma semelhante. Por dependerem diretamente dos valores observados na matriz, esses modelos permitem avaliar como alterações locais nos *ratings* afetam a formação de vizinhanças e o ranqueamento final.

SVD. O SVD decompõe a matriz de interações em fatores latentes de usuários e itens. A pontuação preditiva é obtida a partir do produto escalar entre os vetores latentes do usuário e do item, ajustado pelas médias globais. Esse modelo permite avaliar se notas inferidas podem ser incorporadas a uma representação global de preferências.

EASE. O EASE (*Embarrassingly Shallow Autoencoders*) é um modelo linear raso que aprende relações item-item por meio de uma solução analítica. A partir da matriz de interações X , calcula-se a matriz de Gram $X^T X$, cuja diagonal é regularizada por L_2 antes da inversão. Os pesos item-item são então obtidos normalizando os elementos da matriz inversa pelo respectivo valor na diagonal, com sinal negativo, enquanto a diagonal é fixada em zero. Por aprender relações globais e regularizadas entre itens, o EASE permite avaliar se a substituição dos *ratings* por notas inferidas produz efeitos distintos daqueles observados em métodos baseados em vizinhança.

Neste estudo, fixamos os parâmetros de acordo com a definição original de cada algoritmo. Para os algoritmos ItemKNN e UserKNN, consideramos os $K = 20$ vizinhos mais próximos. Para o SVD, utilizamos $k = 50$ fatores. NO caso do EASE, utilizamos $\lambda = 500$.

3.5 Protocolo de Validação

Adotamos um particionamento temporal por usuário para simular um cenário realista de recomendação. Para cada usuário, as interações foram ordenadas cronologicamente. As 80% interações mais antigas foram utilizadas para treinamento, enquanto as 20% mais recentes foram reservadas para teste. Essa estratégia reduz o risco de vazamento de informações, pois impede que interações futuras sejam usadas para prever interações passadas.

O mesmo particionamento foi utilizado em todas as matrizes de interação: original, VADER, BERT e BART. Assim, as diferenças observadas entre os resultados não decorrem de mudanças no conjunto de treino ou de teste, mas apenas da substituição do *rating* original pela nota inferida. Todos os experimentos foram executados com semente aleatória fixa (*seed* = 42), garantindo reprodutibilidade.

3.6 Métricas de Avaliação

A avaliação foi realizada em listas Top- K , considerando $K \in \{10, 50, 100\}$. Utilizamos duas famílias de métricas: métricas de efetividade, que mensuram a capacidade do modelo de recuperar e ordenar itens relevantes, e métricas *beyond-accuracy*, que avaliam propriedades complementares das listas recomendadas.

Métricas de efetividade. Utilizamos Precision@K, Recall@K, Hit Rate@K e NDCG@K. Precision@K mede a proporção de itens relevantes entre os itens recomendados. Recall@K mede a proporção de itens relevantes do conjunto de teste recuperados na lista recomendada. Hit Rate@K indica a proporção de usuários para os quais ao menos um item relevante foi recomendado. NDCG@K avalia a qualidade do ranqueamento, atribuindo maior peso a itens relevantes posicionados no topo da lista.

Métricas *beyond-accuracy*. Utilizamos as métricas de diversidade e novidade. A diversidade é avaliada por meio da diversidade intra-lista (ILD), calculada a partir da distância de cosseno média entre

os pares de itens recomendados. Enquanto, a novidade é calculada pela auto-informação média dos itens recomendados, atribuindo maior valor a itens menos populares no conjunto de treinamento e penalizando aqueles de maior popularidade. Essas métricas são importantes porque melhorias na efetividade podem vir acompanhadas de recomendações menos exploratórias[30].

Para garantir a confiabilidade das comparações, utilizamos o teste estatístico não paramétrico de Wilcoxon[35], adotando um nível de significância de 5%, com o objetivo de avaliar se as diferenças identificadas entre os modelos são estatisticamente significativas.

3.7 Análise Experimental

A análise experimental foi organizada de acordo com as três perguntas de pesquisa. Para responder à PP1, comparamos os *ratings* originais com as notas inferidas por VADER, BERT e BART, quantificando as divergências entre a avaliação numérica e o sentimento expresso no *review*. Também examinamos exemplos qualitativos de interações divergentes, a fim de ilustrar situações em que o *rating* e o *review* expressam sinais distintos de preferência. Para responder à PP2, comparamos o desempenho dos modelos colaborativos treinados com a matriz original e com as matrizes baseadas em notas inferidas. Essa comparação permite avaliar se a substituição do *rating* pelo sinal textual melhora ou prejudica a efetividade das recomendações. Por fim, para responder à PP3, analisamos a diversidade e a novidade nas listas geradas pelos mesmos modelos e matrizes. Essa etapa permite verificar se eventuais ganhos de efetividade obtidos com notas inferidas são acompanhados por mudanças na variedade e na capacidade de descoberta das recomendações. Todo o código utilizado está disponível em um repositório público³.

4 RESULTADOS E DISCUSSÕES

Esta seção apresenta os resultados dos experimentos organizados segundo as três perguntas de pesquisa deste trabalho.

4.1 PP1: Divergência entre *ratings* e *reviews*

A primeira pergunta de pesquisa investiga em que medida os *ratings* explícitos atribuídos pelos usuários estão alinhados ao sentimento expresso em seus *reviews*. Para isso, comparamos os *ratings* originais às notas inferidas pelos modelos de análise de sentimento, considerando divergência uma diferença absoluta mínima de 1 ponto entre as notas (evitando tratar variações como 2,9 e 3,1 como discordância). A Tabela 2 apresenta os percentuais de divergência observados em cada coleção e para cada estratégia de inferência.

Base de dados	Modelo	Divergências
Philadelphia	VADER	12,97%
	BERT	11,29%
	BART	6,54%
Tucson	VADER	14,42%
	BERT	9,76%
	BART	6,39%
Digital Music	VADER	4,46%
	BERT	4,81%
	BART	2,10%
Musical Instruments	VADER	7,67%
	BERT	4,84%
	BART	3,11%

Tabela 2: Percentual de interações reclassificadas com diferença maior ou igual a 1 por modelo e base de dados.

³<https://github.com/NaanVas/webmedia26-codigo.git>

Os resultados mostram que divergências entre *ratings* e notas inferidas ocorrem em todas as bases avaliadas. Esse achado é relevante porque indica que a nota numérica nem sempre reflete diretamente a opinião expressa no *review*. Nas coleções da Yelp, os percentuais de divergência variam entre 6,39% e 14,42%, enquanto, nas coleções da Amazon, os valores variam entre 2,10% e 7,67%. Portanto, embora as interações divergentes não representem a maioria dos casos, ocorrem com recorrência nos dois domínios avaliados, sugerindo que a inconsistência não é um fenômeno isolado.

A diferença entre as bases ajuda a interpretar esse comportamento. As coleções do Yelp apresentam os maiores percentuais de divergência e possuem *reviews* mais longos (medianas de 267 e 279 caracteres). Em contrapartida, a base Digital Music, com mediana de apenas 68 caracteres, concentra as menores divergências. Esse padrão sugere que *reviews* extensos oferecem maior espaço para nuances, críticas mistas e ressalvas, aumentando a probabilidade de o texto revelar aspectos da experiência do usuário que a nota numérica não consegue capturar.

Dataset	Comentário	Nota Original	Nota BERT	Nota BART	Nota VADER
Musical Instruments	Got two, but the first pair broke after a week of playing. :(	4.0	2.10	1.98	1.50
Musical Instruments	Great cables - fast service	1.0	4.75	4.50	4.24
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	1.36	2.04
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	4.66	4.39
Philadelphia	The staff is rude, it's always crowded, the clientele is rude, and it takes years to even get a glass of water	5	1.97	1.35	1.56
Philadelphia	Nice decor, nice ambiance, good neighborhood location, nice staff.	2.0	4.23	4.27	4.76
Tucson	Terrible to be ignored for close to 15-20 minutes, including by 2 cooks who walked by, seeing I had no menu, no Wawa, nuttn.	4.0	1.10	1.38	1.61
Tucson	The server was excellent and so friendly.	2.0	4.71	4.46	4.60

Tabela 3: Exemplos de inconsistência entre a nota original e o sentimento expresso no comentário por coleção.

Também há diferenças entre os modelos de análise de sentimento. O BART apresenta os menores percentuais de divergência em todas as bases, enquanto o VADER tende a produzir os maiores índices — exceto na Digital Music, onde o BERT o supera ligeiramente. Esse comportamento decorre das diferentes formas de captura de sentimento: o VADER baseia-se em regras lexicais, reagindo a termos positivos ou negativos isolados. Em contrapartida, o BERT e o BART exploram representações contextuais profundas, o que resulta em estimativas mais conservadoras ou alinhadas ao *rating* original.

A Tabela 3 apresenta exemplos de divergência entre *rating* e *review*, selecionados de forma aleatória dentre os casos de maior divergência absoluta entre a nota original e a nota inferida. Esses casos ajudam a ilustrar que a inconsistência não decorre apenas de erro nos modelos de análise de sentimento, mas também pode

Coleção:		Yelp - Philadelphia											
Algoritmo	Variação	Precision			Recall			Ndeg			Hit Rate		
		10	50	100	10	50	100	10	50	100	10	50	100
ItemKNN	Original	0,0031	0,0030	0,0029	0,0089	0,0424	0,0822	0,0060	0,0157	0,0247	0,0297	0,1318	0,2330
	BERT	0,0028▼	0,0029▼	0,0028-	0,0076▼	0,0419-	0,0835-	0,0051▼	0,0148-	0,0241-	0,0266▼	0,1264▼	0,2326-
	BART	0,0031-	0,0032▲	0,0031▲	0,0089-	0,0446-	0,0866▲	0,0059-	0,0162▲	0,0257▲	0,0296-	0,1372▲	0,2450▲
	VADER	0,0029-	0,0032▲	0,0031▲	0,0081-	0,0443-	0,0876▲	0,0055-	0,0159-	0,0257▲	0,0285-	0,1367▲	0,2487▲
UserKNN	Original	0,0059	0,0056	0,0046	0,0140	0,0696	0,1055	0,0103	0,0268	0,0359	0,0559	0,2271	0,3177
	BERT	0,0056-	0,0056-	0,0046-	0,0131-	0,0687-	0,1060-	0,0096-	0,0260-	0,0355-	0,0530-	0,2233-	0,3179-
	BART	0,0059-	0,0056-	0,0048▲	0,0141-	0,0695-	0,1091▲	0,0105-	0,0269-	0,0369-	0,0560-	0,2267-	0,3249▲
	VADER	0,0057-	0,0057-	0,0047▲	0,0134-	0,0702-	0,1084-	0,0098-	0,0265-	0,0361-	0,0538-	0,2265-	0,3240▲
SVD	Original	0,0206	0,0124	0,0094	0,0451	0,1278	0,1908	0,0379	0,0625	0,0782	0,1651	0,3674	0,4778
	BERT	0,0213▲	0,0126▲	0,0096▲	0,0467▲	0,1304▲	0,1945▲	0,0401▲	0,0647▲	0,0807▲	0,1705▲	0,3732▲	0,4842▲
	BART	0,0212▲	0,0127▲	0,0096▲	0,0463▲	0,1306▲	0,1941▲	0,0402▲	0,0652▲	0,0810▲	0,1700▲	0,3736▲	0,4844▲
	VADER	0,0213▲	0,0127▲	0,0096▲	0,0471▲	0,1318▲	0,1960▲	0,0405▲	0,0655▲	0,0815▲	0,1712▲	0,3753▲	0,4864▲
EASE	Original	0,0259	0,0134	0,0096	0,0573	0,1357	0,1888	0,0503	0,0731	0,0866	0,2009	0,3788	0,4702
	BERT	0,0264▲	0,0137▲	0,0098▲	0,0580-	0,1389-	0,1927▲	0,0514▲	0,0750▲	0,0886▲	0,2035-	0,3872▲	0,4782▲
	BART	0,0263-	0,0137▲	0,0098▲	0,0581-	0,1403-	0,1950▲	0,0515-	0,0755▲	0,0894▲	0,2031-	0,3897▲	0,4817▲
	VADER	0,0259-	0,0136▲	0,0097-	0,0576-	0,1390-	0,1911-	0,0512-	0,0749-	0,0882▲	0,2014-	0,3874▲	0,4768▲
Coleção:		Yelp - Tucson											
Algoritmo	Variação	Precision			Recall			Ndeg			Hit Rate		
		10	50	100	10	50	100	10	50	100	10	50	100
ItemKNN	Original	0,0019	0,0019	0,0019	0,0048	0,0230	0,0466	0,0033	0,0088	0,0143	0,0190	0,0861	0,1584
	BERT	0,0020-	0,0021▲	0,0021▲	0,0053-	0,0259▲	0,0519▲	0,0036-	0,0097▲	0,0158▲	0,0195-	0,0927▲	0,1714▲
	BART	0,0021-	0,0022▲	0,0021▲	0,0053-	0,0270▲	0,0524▲	0,0037-	0,0102▲	0,0162▲	0,0209-	0,0980▲	0,1741▲
	VADER	0,0023▲	0,0023▲	0,0022▲	0,0061▲	0,0280▲	0,0542▲	0,0041▲	0,0107▲	0,0169▲	0,0226▲	0,1031▲	0,1819▲
UserKNN	Original	0,0049	0,0048	0,0045	0,0107	0,0548	0,0993	0,0083	0,0214	0,0324	0,0469	0,1925	0,3080
	BERT	0,0053-	0,0049-	0,0045-	0,0115-	0,0561-	0,0998-	0,0087-	0,0220-	0,0329-	0,0501-	0,1968-	0,3069-
	BART	0,0052-	0,0050-	0,0047▲	0,0117-	0,0565-	0,1019▲	0,0091-	0,0226-	0,0339-	0,0493-	0,1955-	0,3146-
	VADER	0,0057▲	0,0054▲	0,0047▲	0,0127▲	0,0604▲	0,1025-	0,0094-	0,0238▲	0,0343▲	0,0537▲	0,2111▲	0,3185▲
SVD	Original	0,0158	0,0107	0,0083	0,0325	0,1082	0,1676	0,0274	0,0503	0,0652	0,1294	0,3321	0,4442
	BERT	0,0164▲	0,0108-	0,0084▲	0,0335-	0,1088-	0,1694-	0,0279-	0,0507-	0,0659▲	0,1353▲	0,3323-	0,4498-
	BART	0,0167▲	0,0109▲	0,0084-	0,0337-	0,1103-	0,1690-	0,0282▲	0,0512▲	0,0660▲	0,1364▲	0,3394▲	0,4487-
	VADER	0,0165-	0,0108-	0,0085▲	0,0333-	0,1100-	0,1700-	0,0280-	0,0510-	0,0661▲	0,1364▲	0,3346-	0,4470-
EASE	Original	0,0192	0,0111	0,0082	0,0381	0,1078	0,1577	0,0337	0,0543	0,0671	0,1533	0,3331	0,4260
	BERT	0,0189-	0,0113▲	0,0085▲	0,0376-	0,1098-	0,1627▲	0,0340-	0,0557▲	0,0691▲	0,1495-	0,3351-	0,4327▲
	BART	0,0190-	0,0113▲	0,0085▲	0,0380-	0,1112-	0,1623▲	0,0342-	0,0561▲	0,0693▲	0,1512-	0,3412▲	0,4325▲
	VADER	0,0188-	0,0112-	0,0085▲	0,0378-	0,1088-	0,1632▲	0,0338-	0,0553-	0,0691▲	0,1519-	0,3344-	0,4357▲
Coleção:		Amazon - Musical Instruments											
Algoritmo	Variação	Precision			Recall			Ndeg			Hit Rate		
		10	50	100	10	50	100	10	50	100	10	50	100
ItemKNN	Original	0,0031	0,0026	0,0023	0,0180	0,0762	0,1346	0,0102	0,0240	0,0350	0,0297	0,1171	0,2025
	BERT	0,0020▼	0,0028▲	0,0024▲	0,0119▼	0,0845▲	0,1384▲	0,0065▼	0,0238▼	0,0341▼	0,0197▼	0,1261▲	0,2079▲
	BART	0,0020▼	0,0028▲	0,0024▲	0,0113▼	0,0825▲	0,1384▲	0,0060▼	0,0230-	0,0337▼	0,0191▼	0,1240▲	0,2082▲
	VADER	0,0019▼	0,0026-	0,0024-	0,0119▼	0,0783-	0,1353-	0,0063▼	0,0220▼	0,0329-	0,0189▼	0,1168-	0,2028-
UserKNN	Original	0,0088	0,0041	0,0024	0,0511	0,1135	0,1329	0,0406	0,0564	0,0602	0,0785	0,1764	0,2072
	BERT	0,0092▲	0,0041-	0,0023▼	0,0533▲	0,1143-	0,1298▼	0,0368▼	0,0526▼	0,0556▼	0,0810▲	0,1780-	0,2033▼
	BART	0,0094▲	0,0041-	0,0023▼	0,0548▲	0,1147-	0,1305▼	0,0411-	0,0564▲	0,0595-	0,0836▲	0,1776-	0,2036▼
	VADER	0,0086-	0,0042▲	0,0024-	0,0501-	0,1173-	0,1334-	0,0321▼	0,0492▼	0,0524▼	0,0766-	0,1810▲	0,2072-
SVD	Original	0,0126	0,0047	0,0032	0,0716	0,1264	0,1680	0,0617	0,0758	0,0839	0,1111	0,2036	0,2704
	BERT	0,0129▲	0,0047▲	0,0032▲	0,0729▲	0,1281▲	0,1698▲	0,0630▲	0,0772▲	0,0853▲	0,1136▲	0,2065▲	0,2731▲
	BART	0,0131▲	0,0047▲	0,0032▲	0,0737▲	0,1278▲	0,1702▲	0,0652▲	0,0791▲	0,0874▲	0,1135▲	0,2051-	0,2723-
	VADER	0,0130▲	0,0047-	0,0032-	0,0732▲	0,1275-	0,1686-	0,0653▲	0,0791▲	0,0871▲	0,1119-	0,2027-	0,2684-
EASE	Original	0,0194	0,0062	0,0038	0,1090	0,1653	0,2016	0,0958	0,1106	0,1177	0,1685	0,2610	0,3169
	BERT	0,0196-	0,0062▲	0,0038-	0,1100-	0,1672▲	0,2028-	0,0958-	0,1108▲	0,1178▲	0,1697-	0,2630-	0,3188-
	BART	0,0197▲	0,0063▲	0,0039▲	0,1108▲	0,1690▲	0,2038▲	0,0964▲	0,1117▲	0,1185▲	0,1715▲	0,2665▲	0,3201▲
	VADER	0,0193-	0,0061-	0,0038▼	0,1088-	0,1648-	0,1999-	0,0950▼	0,1097▼	0,1165▼	0,1678-	0,2590-	0,3141▼
Coleção:		Amazon - Digital Music											
Algoritmo	Variação	Precision			Recall			Ndeg			Hit Rate		
		10	50	100	10	50	100	10	50	100	10	50	100
ItemKNN	Original	0,0029	0,0029	0,0026	0,0155	0,0784	0,1342	0,0091	0,0245	0,0355	0,0276	0,1297	0,2181
	BERT	0,0026▼	0,0029-	0,0027-	0,0139▼	0,0804-	0,1347-	0,0082-	0,0244-	0,0353-	0,0240▼	0,1297-	0,2181-
	BART	0,0027-	0,0029-	0,0027-	0,0142-	0,0795-	0,1353-	0,0081-	0,0241-	0,0352-	0,0247▼	0,1278-	0,2202-
	VADER	0,0027-	0,0029-	0,0026-	0,0147-	0,0785-	0,1326-	0,0082-	0,0237-	0,0345-	0,0258-	0,1271-	0,2154-
UserKNN	Original	0,0039	0,0037	0,0027	0,0198	0,0905	0,1344	0,0109	0,0286	0,0374	0,0365	0,1569	0,2233
	BERT	0,0036-	0,0039▲	0,0027▼	0,0184-	0,0949▲	0,1309▼	0,0095▼	0,0289▲	0,0360-	0,0339-	0,1628▲	0,2165▼
	BART	0,0033▼	0,0038▲	0,0027▼	0,0167▼	0,0933▲	0,1309▼	0,0087▼	0,0282-	0,0357-	0,0314▼	0,1615▲	0,2180▼
	VADER	0,0033▼	0,0038▲	0,0027▼	0,0165▼	0,0937▲	0,1303▼	0,0086▼	0,0282-	0,0355-	0,0311▼	0,1604-	0,2167▼
SVD	Original	0,0051	0,0028	0,0022	0,0238	0,0655	0,1024	0,0163	0,0273	0,0346	0,0482	0,1226	0,1846
	BERT	0,0053▲	0,0029-	0,0023▲	0,0247-	0,0663-	0,1060▲	0,0166▲	0,0276▲	0,0354▲	0,0499-	0,1250▲	0,1894▲
	BART	0,0052-	0,0029▲	0,0023▲	0,0242-	0,0674▲	0,1048▲	0,0165-	0,0279▲	0,0353▲	0,0490-	0,1260▲	0,1864-
	VADER	0,0050-	0,0029▲	0,0023-	0,0236-	0,0668-	0,1029-	0,0161-	0,0275-	0,0347-	0,0472-	0,1261▲	0,1861-
EASE	Original	0,0157	0,0059	0,0039	0,0773	0,1377	0,1751	0,0597	0,0761	0,0837	0,1380	0,2395	0,2944
	BERT	0,0159-	0,0060-	0,0039-	0,0781-	0,1382-	0,1775▲	0,0604-	0,0767▲	0,0846▲	0,1394-	0,2392-	0,2978▲
	BART	0,0158-	0,0060▲	0,0039▲	0,0772-	0,1388-	0,1769▲	0,0604-	0,0771▲	0,0848▲	0,1379-	0,2403-	0,2968-
	VADER	0,0158-	0,0060-	0,0039▲	0,0772-	0,1389-	0,1781▲	0,0592-	0,0758-	0,0837-	0,1376-	0,2406-	0,2985-

Tabela 4: Métricas de Precisão e Ranqueamento consolidadas para todas as coleções. Resultados são avaliados com teste de Wilcoxon ($p = 0,05$). Símbolos: ▲ (ganho significativo), ● (empate estatístico), ▼ (perda significativa).

refletir situações reais em que o usuário expressa uma avaliação textual distinta da nota atribuída. Há casos em que *ratings* altos são acompanhados de críticas relevantes, assim como casos em que *ratings* baixos aparecem associados a *reviews* com sinais positivos. Nesses exemplos, a nota inferida a partir do texto aproxima-se mais do conteúdo semântico do *review* do que do valor numérico original.

4.2 PP2: Impacto na efetividade

A segunda pergunta avalia se a substituição dos *ratings* originais por notas inferidas a partir dos *reviews* melhora a efetividade dos modelos. A Tabela 4 apresenta os resultados de Precision, Recall, Hit Rate e NDCG, considerando diferentes configurações (algoritmo, coleção, métrica etc.). Os resultados mostram que o impacto das notas inferidas não é uniforme entre os algoritmos. De modo geral, os ganhos de efetividade são mais consistentes em modelos colaborativos globais, especialmente SVD e EASE. Esses modelos aprendem representações latentes ou pesos globais a partir da matriz de interações, o que parece favorecer a incorporação de um sinal semântico contínuo derivado dos *reviews*. Ao substituir o *rating* original pela nota inferida, esses algoritmos conseguem incorporar o novo sinal em uma estrutura de aprendizado mais ampla, suavizando alterações pontuais e capturando padrões gerais de preferência. O comportamento difere entre os modelos baseados em vizinhança. UserKNN e ItemKNN apresentam resultados mais instáveis, com ganhos pontuais, empates estatísticos e perdas em alguns cenários. Isso pode estar relacionado à forma como esses métodos utilizam diretamente os valores

observados na matriz para calcular similaridades entre usuários ou itens. Alterações nas notas, ainda que semanticamente justificadas pelo conteúdo dos *reviews*, podem modificar as relações locais de vizinhança e afetar o ranqueamento de forma menos previsível. Assim, os resultados indicam que o benefício da nota inferida não depende apenas da qualidade da análise de sentimento, mas também da forma como o algoritmo de recomendação utiliza o sinal de entrada.

Nas bases da Yelp, observa-se um cenário mais favorável ao uso de notas inferidas. Em Philadelphia, o SVD apresenta ganhos estatisticamente significativos em praticamente todas as métricas e tamanhos de lista, independentemente do modelo de análise de sentimento utilizado. O EASE também apresenta ganhos recorrentes, principalmente em listas maiores, com melhorias em $k = 50$ e $k = 100$. Em Tucson, os ganhos do ItemKNN tornam-se mais evidentes, especialmente com o VADER, que melhora todas as métricas avaliadas. Nas bases da Amazon, os efeitos são mais variados. Em Musical Instruments, SVD e EASE continuam apresentando os comportamentos mais positivos, com destaque para o BART no EASE, que obtém ganhos significativos em todas as métricas e tamanhos de lista. No entanto, os modelos baseados em vizinhança apresentam perdas em alguns cenários, especialmente em listas curtas. Em Digital Music, os ganhos são mais discretos e concentrados novamente em SVD e EASE, enquanto UserKNN e ItemKNN apresentam muitos empates estatísticos e perdas. Esse contraste acompanha as características textuais das coleções apresentadas na Tabela 1: as bases da Yelp, com maiores ganhos, possuem *reviews*

Coleção:		Yelp - Philadelphia						Coleção:		Amazon - Musical Instruments					
Algoritmo	Variação	Diversity			Novelty			Algoritmo	Variação	Diversity			Novelty		
		10	50	100	10	50	100			10	50	100	10	50	100
ItemKNN	Original	0,9310	0,9554	0,9678	14,6805	14,7595	14,8004	ItemKNN	Original	0,9805	0,9886	0,9925	14,4115	14,2896	14,3025
	BERT	0,8964▼	0,9480▼	0,9676●	14,6969▲	14,7491●	14,7626▼		BERT	0,9700▼	0,9864▼	0,9928●	14,3306▼	14,1203▼	14,2610▼
	BART	0,8917▼	0,9445▼	0,9653▼	14,5342▼	14,6255▼	14,6768▼		BART	0,9698▼	0,9860▼	0,9926▲	14,3408▼	14,1404▼	14,2764▼
	VADER	0,8987▼	0,9486▼	0,9677▼	14,6066●	14,6828▼	14,7030▼		VADER	0,9644▼	0,9836▼	0,9914▼	14,3261▼	14,1607▼	14,2770▼
UserKNN	Original	0,9718	0,9741	0,9805	12,0668	11,9652	12,5370	UserKNN	Original	0,9895	0,9931	0,9965	12,5498	12,6563	13,4548
	BERT	0,9730▲	0,9745▲	0,9806▲	12,0798▲	11,9601●	12,5468▲		BERT	0,9860▼	0,9930●	0,9967▲	12,2382▼	12,6857▲	13,5280▲
	BART	0,9698▼	0,9735▼	0,9796▼	11,8669▼	11,9480▼	12,5333●		BART	0,9872▼	0,9929▼	0,9966▲	12,2961▼	12,6966▲	13,5273▲
	VADER	0,9700▼	0,9738▼	0,9800▼	11,9246▼	11,9456▼	12,5315●		VADER	0,9861▼	0,9926▼	0,9965▼	12,5484●	12,6543▼	13,4886▲
SVD	Original	0,9150	0,9347	0,9447	9,8489	10,2980	10,6129	SVD	Original	0,9754	0,9830	0,9867	9,6273	10,2035	10,6348
	BERT	0,9160▲	0,9358▲	0,9457▲	9,8479●	10,2937▼	10,6078▼		BERT	0,9761▲	0,9843▲	0,9877▲	9,6323▲	10,1958▼	10,6327▼
	BART	0,9132▼	0,9332▼	0,9434▼	9,8381▼	10,2884▼	10,6019▼		BART	0,9760▲	0,9836▲	0,9871▲	9,6528▲	10,2081▲	10,6340▼
	VADER	0,9145▼	0,9345▼	0,9445▼	9,8506●	10,2935▼	10,6080▼		VADER	0,9766▲	0,9834▲	0,9863▼	9,7172▲	10,2709▲	10,7007▲
EASE	Original	0,9345	0,9554	0,9639	10,3469	10,9460	11,2874	EASE	Original	0,9815	0,9910	0,9934	10,8459	11,7905	12,1768
	BERT	0,9344●	0,9552▼	0,9637▼	10,3181▼	10,9047▼	11,2454▼		BERT	0,9810▼	0,9909▼	0,9934▼	10,7734▼	11,7250▼	12,1298▼
	BART	0,9315▼	0,9530▼	0,9619▼	10,2882▼	10,8734▼	11,2169▼		BART	0,9809▼	0,9909▼	0,9934▼	10,7539▼	11,7103▼	12,1306▼
	VADER	0,9337▼	0,9545▼	0,9632▼	10,3198▼	10,8957▼	11,2374▼		VADER	0,9815●	0,9908▼	0,9933▼	10,9114▲	11,8257▲	12,2012▲
Coleção:		Yelp - Tucson						Coleção:		Amazon - Digital Music					
Algoritmo	Variação	Diversity			Novelty			Algoritmo	Variação	Diversity			Novelty		
		10	50	100	10	50	100			10	50	100	10	50	100
ItemKNN	Original	0,9111	0,9484	0,9648	14,1061	14,2023	14,2310	ItemKNN	Original	0,9549	0,9741	0,9842	14,2196	14,2222	14,2006
	BERT	0,8760▼	0,9390▼	0,9637▼	14,1117●	14,1463▼	14,1808▼		BERT	0,9346▼	0,9684▼	0,9835▼	14,1984▼	14,2033▼	14,1765▼
	BART	0,8702▼	0,9355▼	0,9617▼	14,0376▼	14,0777▼	14,1259▼		BART	0,9331▼	0,9674▼	0,9830▼	14,2177▼	14,2219●	14,1960▼
	VADER	0,8774▼	0,9382▼	0,9630▼	14,0001▼	14,0732▼	14,1166▼		VADER	0,9170▼	0,9578▼	0,9777▼	14,1647▼	14,1770▼	14,1595▼
UserKNN	Original	0,9580	0,9689	0,9751	11,6150	11,5932	11,8942	UserKNN	Original	0,9806	0,9862	0,9908	12,5243	12,9015	13,2068
	BERT	0,9606▲	0,9693●	0,9753▲	11,5597▼	11,5633▼	11,9190▲		BERT	0,9784▼	0,9849▼	0,9910▲	12,5592▲	12,9023●	13,2577▲
	BART	0,9581▲	0,9684▼	0,9746▼	11,3960▼	11,5557▼	11,9075▲		BART	0,9794▼	0,9855▼	0,9910▲	12,5389▲	12,9027●	13,2617▲
	VADER	0,9557▼	0,9671▼	0,9740▼	11,3681▼	11,4892▼	11,8913●		VADER	0,9732▼	0,9834▼	0,9900●	12,5742▲	12,8950▼	13,2543▲
SVD	Original	0,9072	0,9311	0,9424	9,5461	9,9971	10,3404	SVD	Original	0,9594	0,9695	0,9742	10,7011	11,3031	11,6500
	BERT	0,9102▲	0,9323▲	0,9433▲	9,5051▼	9,9596▼	10,3045▼		BERT	0,9578▼	0,9689▼	0,9738▼	10,6985▼	11,3204▲	11,6644▲
	BART	0,9063▼	0,9299▼	0,9412▼	9,5107▼	9,9711▼	10,3146▼		BART	0,9580▼	0,9686▼	0,9736▼	10,6814▼	11,2986▼	11,6436●
	VADER	0,9075▲	0,9307▼	0,9420▼	9,5065▼	9,9585▼	10,3059▼		VADER	0,9510▼	0,9593▼	0,9656▼	10,7676▲	11,4409▲	11,7769▼
EASE	Original	0,9350	0,9552	0,9638	10,0381	10,5897	10,9444	EASE	Original	0,9711	0,9854	0,9890	12,1906	12,5614	12,7495
	BERT	0,9338▼	0,9541▼	0,9629▼	9,9700▼	10,5209▼	10,8769▼		BERT	0,9702▼	0,9849▼	0,9887▼	12,1608▼	12,5438▼	12,7367▼
	BART	0,9317▼	0,9524▼	0,9615▼	9,9423▼	10,4928▼	10,8535▼		BART	0,9703▼	0,9850▼	0,9887▼	12,1519▼	12,5320▼	12,7284▼
	VADER	0,9328▼	0,9532▼	0,9621▼	9,9547▼	10,5012▼	10,8577▼		VADER	0,9681▼	0,9838▼	0,9878▼	12,2002▲	12,5596●	12,7454▼

Tabela 5: Métricas Beyond-Accuracy consolidadas. Resultados são avaliados com teste de Wilcoxon ($p = 0,05$). Símbolos: ▲ (ganho significativo), ● (empate estatístico), ▼ (perda significativa).

mais longos, enquanto as da Amazon apresentam textos mais curtos e ganhos mais discretos. Isso sugere que *reviews* mais longas fornecem informações mais detalhadas sobre as preferências dos usuários, contribuindo para inferências de nota mais precisas.

4.3 PP3: Impacto em diversidade e novidade

A terceira pergunta avalia se o uso de notas inferidas dos *reviews* afeta propriedades das listas recomendadas além da efetividade. Essa análise é importante, pois sistemas de recomendação não devem ser avaliados apenas pela relevância dos itens recuperados, mas também pela variedade e pela capacidade de promover descoberta. A Tabela 5 apresenta os resultados de diversidade e novidade.

Os resultados mostram que os ganhos de efetividade observados em PP2 frequentemente vêm acompanhados de reduções na diversidade e na novidade. Essa tendência aparece nas quatro bases, embora com intensidade variável entre algoritmos e modelos de análise de sentimento. O padrão geral sugere um trade-off: ao substituir o *rating* original por uma nota mais alinhada ao sentimento textual, os modelos podem tornar as recomendações mais efetivas, mas também mais concentradas em itens semelhantes ou mais populares.

Nas bases da Yelp, esse comportamento é evidente. Em Philadelphia, o ItemKNN apresenta perdas recorrentes de diversidade em quase todas as variações, enquanto a novidade também é reduzida na maioria dos cenários, especialmente com BART e VADER. O EASE, também sofre perdas frequentes em diversidade e novidade. Em Tucson, observa-se um padrão semelhante: ItemKNN e EASE apresentam perdas consistentes nessas métricas, enquanto UserKNN e SVD exibem um comportamento mais misto, com alguns ganhos pontuais em diversidade, mas perdas frequentes em novidade. Nas bases da Amazon, os resultados também indicam a presença desse trade-off. Em Musical Instruments, o ItemKNN apresenta perdas claras em diversidade e novidade, enquanto o EASE reduz a diversidade na maioria dos cenários e apresenta ganhos de novidade apenas com VADER. O SVD mostra um comportamento mais equilibrado, com ganhos frequentes em diversidade e ganhos de novidade principalmente com VADER. Em Digital Music, as perdas de diversidade são ainda mais frequentes, afetando praticamente todos os algoritmos e variações. A novidade, por outro lado, apresenta comportamento mais dependente do modelo de análise de sentimento e do algoritmo de recomendação.

4.4 Discussão Geral

Os resultados respondem às três perguntas de pesquisa ao mostrar que *ratings* e *reviews* não devem ser tratados como sinais equivalentes. Em relação à PP1, observamos divergências entre a nota atribuída pelo usuário e o sentimento inferido a partir do texto, especialmente em bases com *reviews* mais longos. Esse achado indica que o *review* pode revelar nuances da experiência que não aparecem no *rating*, funcionando como um sinal independente para avaliar a confiabilidade do *feedback* explícito. Quanto à PP2, as notas inferidas podem melhorar a efetividade, mas esse efeito depende do algoritmo. Modelos globais, como SVD e EASE, tendem a se beneficiar mais desse sinal, enquanto métodos baseados em vizinhança apresentam comportamento mais instável. Já a PP3 mostra que esses ganhos podem vir acompanhados de perdas em diversidade e novidade, sugerindo um trade-off entre o alinhamento semântico do *feedback* e a variedade das recomendações. Em conjunto, os

resultados indicam que a substituição direta do *rating* pela nota inferida não deve ser vista como solução geral. A direção mais promissora está em combinar os dois sinais de forma adaptativa, considerando a consistência entre *rating* e *review*, a confiança da análise de sentimento e os objetivos do sistema de recomendação.

5 CONCLUSÃO E TRABALHOS FUTUROS

Este trabalho investigou a relação entre a nota atribuída por usuários e a opinião expressa em seus *reviews*, analisando em que medida sinais de sentimento podem ser utilizados como alternativa ou complemento ao *rating* explícito em sistemas de recomendação colaborativa. A motivação do estudo parte da constatação de que a avaliação numérica, embora adotada como principal sinal de preferência, pode não representar integralmente a experiência relatada pelo usuário.

A partir de experimentos conduzidos em diferentes domínios, com múltiplos modelos de análise de sentimento e algoritmos de recomendação, os resultados evidenciam discrepâncias relevantes entre nota e opinião, especialmente em bases nas quais as *reviews* tendem a ser mais longas e descritivas. Além disso, observou-se que notas inferidas a partir do texto podem, em determinados cenários, substituir o *rating* original sem comprometer a efetividade da recomendação e, em alguns casos, até melhorar o desempenho dos modelos avaliados. Esses achados indicam que a opinião textual não apenas reproduz a avaliação numérica, mas pode revelar sinais de preferência distintos daqueles capturados diretamente pelo *rating*.

A principal contribuição deste artigo consiste em deslocar a análise do texto de *reviews* de uma posição auxiliar para uma perspectiva em que ele é tratado como fonte independente de evidência sobre a preferência do usuário. Diferentemente de abordagens que utilizam o texto como complemento supervisionado pela nota, este estudo contrasta o *rating* informado pelo usuário com notas inferidas por análise de sentimento, permitindo observar quando esses sinais convergem, quando divergem e como essa divergência impacta a recomendação. Com isso, o trabalho contribui para uma compreensão mais crítica do papel dos *ratings* explícitos e reforça que o desalinhamento entre nota e opinião deve ser considerado no projeto, na avaliação e na interpretação de sistemas de recomendação.

Como desdobramentos futuros, pretende-se investigar estratégias híbridas que combinem *ratings* explícitos e sinais extraídos das *reviews*, explorando diferentes formas de integração entre esses dois tipos de evidência. Também se considera relevante aprofundar métodos capazes de capturar aspectos mais específicos da opinião do usuário, como a polaridade por atributo, a intensidade do sentimento e as variações de comportamento entre domínios. Esses caminhos podem contribuir para sistemas de recomendação mais sensíveis às nuances da experiência relatada pelos usuários e menos dependentes de uma única forma de representar a preferência.

6 CUIDADOS ÉTICOS

Utilizamos neste trabalho apenas *datasets* públicos e abertos para fins de pesquisa, já devidamente pseudoanonimizados por seus respectivos criadores originais, não permitindo a reidentificação direta dos usuários. Portanto, esta pesquisa está em conformidade com as diretrizes da Lei Geral de Proteção de Dados Pessoais (LGPD) (Lei n.º 13.709/2018) no que tange ao tratamento de dados para fins acadêmicos. Este estudo não envolveu nova coleta de dados com seres humanos, atendendo, portanto à Resolução CNS n.º 510/2016).

ACKNOWLEDGMENTS

Esta pesquisa foi financiada por: CNPq, Capes, Fapemig, Fapesp, AWS, NVIDIA, CIIA-Saúde e INCT-TILD-IAR (grant 408490/2024-1). Ferramentas de Inteligência Artificial Generativa foram utilizadas apenas como suporte à revisão gramatical de todo o conteúdo. Os autores realizaram uma revisão completa do texto e assumem total responsabilidade pela integridade das informações apresentadas.

REFERÊNCIAS

- [1] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering* 17, 6 (2005), 734–749.
- [2] Charu C. Aggarwal. 2016. *Recommender Systems: The Textbook* (1 ed.). Springer, Cham. doi:10.1007/978-3-319-29659-3
- [3] Sumaia Mohammed Al-Ghuribi and Shahruil Azman Mohd Noah. 2019. Multi-Criteria Review-Based Recommender System—The State of the Art. *IEEE Access* 7 (2019), 169446–169468. doi:10.1109/ACCESS.2019.2954861
- [4] Amal Almansour, Reem Aloiaibi, and Hajar Alharbi. 2022. Text-Rating Review Discrepancy (TRRD): An Integrative Review and Implications for Research. *Future Business Journal* 8, 1 (2022), 3. doi:10.1186/s43093-022-00114-y
- [5] Xavier Amatriain and Justin Basilico. 2015. *Recommender Systems in Industry: A Netflix Case Study*. 385–419. doi:10.1007/978-1-4899-7637-6_11
- [6] Konstantin Bauman, Bing Liu, and Alexander Tuzhilin. 2017. Aspect Based Recommendations: Recommending Items with the Most Valuable Aspects Based on User Reviews. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) (KDD '17). Association for Computing Machinery, New York, NY, USA, 717–725. doi:10.1145/3097983.3098170
- [7] Guilherme Bittencourt, Guilherme Fonseca, Yan Andrade, Nicollas Silva, and Leonardo Rocha. 2023. A Survey on Review - Aware Recommendation Systems. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web* (Ribeirão Preto, Brazil) (WebMedia '23). 198–207. doi:10.1145/3617023.3617050
- [8] Guilherme Bittencourt, Naan Vasconcelos, Yan Andrade, Nicollas Silva, Washington Cunha, Diego Roberto Colombo Dias, Marcos André Gonçalves, and Leonardo Rocha. 2025. Aware Recommender Systems (RARs): Recent Advances, Experimental Comparative Analysis, Discussions, and New Directions. *Comput. Surveys* 58, 1 (2025), 1–49.
- [9] Guilherme Bittencourt, Naan Vasconcelos, and Leonardo Rocha. 2025. iRev: Um framework de avaliação de sistemas de recomendação baseados comentários textuais. *Revista Eletrônica de Iniciação Científica em Computação* 23 (2025), 130–135.
- [10] Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE intelligent systems* 32, 6 (2017), 74–80.
- [11] Oscar Celma and Paul Lamere. 2011. Music recommendation and discovery revisited. In *Proceedings of the Fifth ACM Conference on Recommender Systems* (Chicago, Illinois, USA) (RecSys '11). Association for Computing Machinery, New York, NY, USA, 7–8. doi:10.1145/2043932.2043936
- [12] Chien Chin Chen, Shun-Yuan Shih, and Meng Lee. 2016. Who should you follow? Combining learning to rank with social influence for informative friend recommendation. *Decision Support Systems* 90 (2016), 33–45. doi:10.1016/j.dss.2016.06.017
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *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)*. 4171–4186.
- [14] Nan Hu, Paul A Pavlou, and Jennifer Zhang. 2006. Can online reviews reveal a product's true quality? Empirical findings and analytical modeling of online word-of-mouth communication. In *Proceedings of the 7th ACM conference on Electronic commerce*. 324–330.
- [15] Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, Vol. 8. 216–225.
- [16] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009), 30–37.
- [17] Yehuda Koren, Steffen Rendle, and Robert Bell. 2021. Advances in collaborative filtering. *Recommender systems handbook* (2021), 91–142.
- [18] Balraj Kumar and Neeraj Sharma. 2016. Approaches, Issues and Challenges in Recommender Systems: A Systematic Review. *Indian Journal of Science and Technology* 9 (12 2016). doi:10.17485/ijst/2015/v8i1/94892
- [19] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 7871–7880.
- [20] Greg Linden, Brent Smith, and Jeremy York. 2003. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing* 7, 1 (2003), 76–80.
- [21] Washington Luiz, Felipe Viegas, Rafael Alencar, Fernando Mourão, Thiago Salles, Dárlinton Carvalho, Marcos Andre Gonçalves, and Leonardo Rocha. 2018. A Feature-Oriented Sentiment Rating for Mobile App Reviews. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1909–1918. doi:10.1145/3178876.3186168
- [22] Emanuelle Marreira, Tiago Melo, Miguel Oliveira, and Carlos Mauricio. 2025. Detectando Incoerências Avaliativas em E-commerce com LLMs - Um Estudo de Caso na Amazon Brasil. In *Proceedings of the 31st Brazilian Symposium on Multimedia and the Web* (Rio de Janeiro/RJ). SBC, Porto Alegre, RS, Brasil, 535–539. doi:10.5753/webmedia.2025.15952
- [23] Jaewon Moon, Hye-young Kim, and Jongwuk Lee. 2023. It's Enough: Relaxing Diagonal Constraints in Linear Autoencoders for Recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) (SIGIR '23). Association for Computing Machinery, New York, NY, USA, 1639–1648. doi:10.1145/3539618.3591704
- [24] Bo Pang and Lillian Lee. 2008. *Opinion mining and sentiment analysis*. Now Publishers Inc.
- [25] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM conference on Computer supported cooperative work*. 175–186.
- [26] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web* (Hong Kong, Hong Kong) (WWW '01). Association for Computing Machinery, New York, NY, USA, 285–295. doi:10.1145/371920.372071
- [27] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. 285–295.
- [28] J. Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. 2007. *Collaborative Filtering Recommender Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, 291–324. doi:10.1007/978-3-540-72079-9_9
- [29] Alex Shtof, Michael Videman, Naama Haramaty-Krasne, Oren Somekh, Ariel Raviv, and Tularam Ban. 2024. Low Rank Field-Weighted Factorization Machines for Low Latency Item Recommendation. In *Proceedings of the 18th ACM Conference on Recommender Systems* (Bari, Italy) (RecSys '24). Association for Computing Machinery, New York, NY, USA, 238–246. doi:10.1145/3640457.3688097
- [30] Rodrigo Souza and Marcelo Manzato. 2024. Uma Abordagem em Etapa de Processamento para Redução do Viés de Popularidade. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web* (Juiz de Fora/MG). SBC, Porto Alegre, RS, Brasil, 310–317. doi:10.5753/webmedia.2024.241542
- [31] Mehdi Srifi, Ahmed Oussous, Ayoub Ait Lahcen, and Salma Mouline. 2020. Recommender Systems Based on Collaborative Filtering Using Review Texts—A Survey. *Information* 11, 6 (2020). doi:10.3390/info11060317
- [32] Harald Steck. 2019. Embarrassingly shallow autoencoders for sparse data. In *The World Wide Web Conference*. 3251–3257.
- [33] Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Ji-ang. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM international conference on information and knowledge management*. 1441–1450.
- [34] Naan Vasconcelos, Davi Reis, Thiago Silva, Nicollas Silva, Washington Cunha, Elisa Tuler, and Leonardo Rocha. 2024. Integrando Avaliações Textuais de Usuários em Recomendação baseada em Aprendizado por Reforço. In *Proceedings of the 30th Brazilian Symposium on Multimedia and the Web* (Juiz de Fora/MG). SBC, Porto Alegre, RS, Brasil, 390–394. doi:10.5753/webmedia.2024.241405
- [35] Joao Vinagre, Alípio Mário Jorge, Conceição Rocha, and Joao Gama. 2019. Statistically robust evaluation of stream-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering* 33, 7 (2019), 2971–2982.
- [36] Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 3914–3923.
- [37] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Comput. Surv.* 52, 1, Article 5 (Feb. 2019), 38 pages. doi:10.1145/3285029
- [38] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval* (Gold Coast, Queensland, Australia) (SIGIR '14). Association for Computing Machinery, New York, NY, USA, 83–92. doi:10.1145/2600428.2609579