Revista Eletrônica de Iniciação Científica em Computação, 2025, 23:1
ISSN: 3085-8461 • doi: 10.5753/reic.2025.6040

Artigo de Pesquisa/Research Paper

iRev: Um framework de avaliação de sistemas de recomendação baseados comentários textuais

iRev: A framework for evaluating recommender systems based on textual comments

Guilherme Bittencourt [Universidade Federal de São João del Rei | *bittenkourt.gmf@aluno.ufsj.edu.br*]
Naan Vasconcelos [Universidade Federal de São João del Rei |*naan.vasconcelos@aluno.ufsj.edu.br*]
Leonardo Rocha [Universidade Federal de São João del Rei |*lcrocha@ufsj.edu.br*]

Departamento de Computação, Universidade Federal de São João del Rei, Rodovia 494, s/n, Bairro Colônia do Bengo, São João del-Rei, MG, 36301-360, Brasil.

Resumo. Os avanços atuais em Sistemas de Recomendação e Processamento de Linguagem Natural têm motivado estudos recentes a retornarem seu interesse em sistemas de recomendação baseados comentários textuais (RARSs). Nesse sentido, realizamos mapeamento sistemático selecionando 117 artigos publicados nos principais veículos da área, apresentando um resumo dos avanços, destacando os principais algoritmos propostos e detalhando os conjuntos de dados e métricas mais utilizados em configurações experimentais. As implementações e demais artefatos extraídos deste estudo foram consolidados em um framework: iREV. Conduzimos uma comparação experimental entre propostas de última geração, destacando as principais perspectivas para desenvolvimentos futuros.

Abstract. Current advances in Recommendation Systems and Natural Language Processing have motivated recent studies to return their interest in Review-Aware Recommendation Systems (RARSs). In this sense, we employ a systematic mapping approach by selecting 117 papers published on the main vehicles of the area, presenting a summary of the advances, highlighting the main proposal algorithms, and detailing the most used datasets and metrics in experimental setups. All the implementations and other artifacts extracted from this study were consolidated into a framework: iREV. In addition, we conduct a comprehensive experimental comparison among state-of-the-art proposals, highlighting the main directions and new perspectives for future developments.

Palavras-chave: Sistemas de Recomendação Baseados em Comentários, Avaliação Comparativa

Keywords: Review Aware, Recommendation Systems, Comparative Evaluation

Recebido/Received: 16 May 2025 • **Aceito/Accepted:** 24 June 2025 • **Publicado/Published:** 11 July 2025

1 Introdução

Os sistemas de recomendação (SsR) surgiram como uma estratégia eficaz para lidar com a sobrecarga de informações. A sua importância é inegável visto que são amplamente adotados em diversas aplicações web, apresentando potencial para resolver diversos problemas associados à abundância de escolhas. Nos últimos anos, a literatura tem testemunhado a proposta de inúmeras técnicas de recomendação, buscando constantemente melhorar a eficácia destes métodos. Embora muitas abordagens tenham se concentrado em técnicas que empregam avaliações quantitativas (numéricas), é essencial reconhecer o valor significativo do *feedback do usuário*, geralmente expresso como comentários (textuais) (também conhecidos como revisões), a fim de compreender suas preferências. Neste contexto, vários algoritmos foram desenvolvidos para aproveitar efetivamente os comentários como uma valiosa fonte de informação, conhecidos como Sistemas de Recomendação Review-Aware (RARSs), capazes de gerar recomendações alinhadas ao perfil de cada usuário a partir de seus respectivos comentários.

Como **primeira contribuição**, esse trabalho apresenta um mapeamento sistemático dos estudos sobre sistemas de recomendação *review-aware* com dois objetivos principais: (i) consolidar uma imagem atualizada das principais pesquisas realizadas nessa área recentemente para futuros trabalhos; (ii) destacar as principais limitações, características e orientações que estamos seguindo como comunidade. Identificamos 117 estudos relevantes sobre recomendação *review-aware* publicados nos principais veículos da área (e.g., RecSys, SIGIR, etc.) de 2014 a 2024. Realizamos um estudo detalhado dos principais avanços, dos principais conjuntos de dados e das métricas utilizadas. Observamos que as coleções de dados mais utilizadas dentre os trabalhos relevantes são as bases de dados de produtos da Amazon e de pontos de interesse da Yelp, por disponibilizarem as interações usuário/item e também os comentários provenientes das interações. Em relação às métricas de avaliação, é possível observar que métricas de erro são as formas de metrificação mais utilizadas pelos trabalhos, seguidas pelas métricas de avaliação de ranking. Porém, apesar do consenso na comunidade de recomendação de que é necessário mais do que precisão para avaliar a eficácia dos SsR, a grande maioria dos trabalhos priorizam a precisão sobre outras dimensões de qualidade, tais como serendipidade e diversidade. Outra limitação surge em relação aos algoritmos que são considerados estado-da-arte e suas configurações. Não existe um consenso das linhas de bases a serem consideradas, cada artigo utiliza um conjunto distinto e as configurações dos parâmetros raramente são reportadas. Menos de 50% dos trabalhos analisados disponibiliza

código fonte de suas propostas, e menos de 30% fornece as configurações de parâmetros dos algoritmos propostos.

Como **segunda contribuição** de nosso trabalho, implementamos os 10 principais algoritmos de recomendação *review-aware* (i.e. mais citados e/ou com melhores resultados reportados), consolidando todos os códigos fontes gerados, bem como todos os artefatos levantados durante o mapeamento sistemático (métricas e bases de dados) em um framework aberto e publicamente disponível[1], denominado iRev, com o objetivo de facilitar a pesquisa e comparações entre abordagens na área de *review aware*. Como **terceira contribuição**, realizamos uma análise experimental de diferentes algoritmos, com diversas coleções e métricas, destacando as principais direções para desenvolvimentos futuros.

Esse trabaho teve um artigo publicado no WebMedia 2023 Bittencourt et al. *[2023], foi premiado no CTIC WebMedia 2024 e resultou em um artigo submetido na ACM Computing Survey (A1)*

2 Mapeamento Sistemático

Nesta seção, abordamos o processo de coleta, filtragem e seleção dos artigos para a categorização e avaliação experimental Werneck *et al*. [2020].

2.1 Fase 1: Questões de pesquisa, palavras de busca e fontes digitais

As três questões de pesquisa que deverão ser respondidas com esse mapeamento são:

- **QP1:** Como os algoritmos de recomendação utilizam técnicas de PLN para definir as preferências dos usuários por meio de seus comentários?
- **QP2:** Quais são as bases de dados e métricas mais prevalentes utilizadas na avaliação de algoritmos em estudos relacionados a esse tema?
- **QP3:** Como as avaliações experimentais são conduzidas nos estudos analisados, considerando estados da arte, configurações e parâmetros dos modelos?

Recorremos ao mecanismo de pesquisa do Google Scholar para realizar as três consultas abaixo e formar nossa primeira coleção de artigos:

- **SS-Q1:** ("review based" OR "review aware" OR "review modeling") AND ("recommender systems"OR "recommendation systems"OR "recommender system") AND ("text" OR "textual"OR "review")
- **SS-Q2:** ("review based" OR "review aware" OR "review modeling") AND ("recommender systems" OR "recommendation systems" OR "recommender system") AND ("evaluation" OR "measure"OR "metrics")
- **SS-Q3:** ("review based" OR "review aware" OR "review modeling") AND ("recommender systems"OR "recommendation systems"OR "recommender system") AND ("source code" OR "reproducibility" OR "empirical" OR "experimental")

2.2 Fase 2: Seleção de trabalhos relevantes

A concluir a coleta, aplicamos um filtro de data entre 2014 e 2023, assegurando um escopo de 10 anos de publicações, acumulando um total de 1.190 artigos. Eliminamos as duplicadas e aplicamos um segundo filtro considerando apenas artigos das 100 conferências de maior fator de impacto de acordo com a `research.com`, tais como RecSys, WWW, WSDM, KDD, SIGIR, etc., resultando em 681 artigos. Por fim, empregamos um filtro avançado com critérios de inclusão e exclusão. Realizamos uma análise manual de cada artigo, classificando-os como relevantes ou não, de acordo com os critérios estabelecidos:

Critérios de Inclusão

- O método principal empregado para realizar as recomendações é o uso dos comentários dos usuários, considerando as avaliações numéricas como um suporte adicional.
- Propõem avanços e inovações no domínio, não se limitando à otimização de algoritmos preexistentes.
- Realizam avaliações experimentais comparativas entre os algoritmos que utilizam comentários do usuário e o método proposto nos artigos correspondentes.

Critérios de Exclusão

- Além dos comentários, empregam outras fontes de informação, como imagens, áudios ou vídeos para prever as preferências do usuário.
- São surveys, casos de estudo, revisões sistemáticas ou experimentais sobre os algoritmos do cenário.
- Utilizam os comentários exclusivamente para justificar as recomendações, focando na explicabilidade..

Após a aplicação dos critérios mencionados anteriormente, restaram 117 artigos que foram identificados como mais relevantes para este mapeamento sistemático. Esses artigos serão o foco discutido no presente trabalho.

2.3 Fase 3: Extração das informações dos artigos

Após finalizarmos a fase de coleta e tratamento dos artigos, realizamos uma leitura detalha de todos os 117 artigos restantes para identificar as principais caracterśiticas das soluções propostas, suas principais inovações e contribuições para a literatura e as metodologias de avaliação utilizadas. Nas próximas seções detalhamos o resultado dessa analise visando responder as três questões de pesquisas levantadas no início dessa seção.

3 Uma Nova Taxnomomia para Sistemas de Recomendação Review-Aware

Esta seção foca em responder a primeira questão de pesquisa. Para tal, construímos uma nova taxonomia para estudar e classificar cada um dos 117 artigos selecionados em nossa SLR. Em nossa proposta de taxonomia dividimos os artigos por meio de quatro principais perspectivas: (1) Modelagem da Informação; (2) Extração da Informação; (3) Objetivo do Modelo e; (4) Arquitetura de Aprendizado dos SsR. A seguir, detalhamos cada uma delas. **A distribuição dos 117 artigos entre essas perspectivas está disponibilizado online**[2].

[1] `https://github.com/guibitten03/iRev`

[2] `https://docs.google.com/spreadsheets/d/1WAp7J67QqoRB2wdJCdTTi3SGcYmafY6krVI2cw9sydw/edit?gid=0#gid=0`

1 - Modelagem da Informação: A primeira perspectiva diz respeito à estratégia de modelagem das informações textuais, na qual propomos três classes distintas: (1) Modelagem por Documento, a qual consitem de trabalhos que concatenam todos os comentários de cada usuário como se fossem um único documento, bem como comentários de itens em um único documento; (2) Modelagem por Sentença, que propõem mecanismos de reconhecimento de palavras chaves nas sentenças e, posteriormente, sentenças chaves, para compor a representação de usuários e itens; e (3) Rating Aggregation que buscam associar as avaliações numéricas dos usuários/itens com seus respectivos comentários, semelhantes aos sistemas de recomendação baseados em filtros colaborativos tradicionais.
2 - Extração da Informação: A segunda perspectiva diz respeito a como as informações, depois de modeladadas, são extraídas no processo de construção dos SsR. Para essa perspectiva, propomos a classificação dos SsR em duas classes distintas: (1) Extração de Sentimento, que correspondem aos trabalhos que utilizam análise de sentimento para a modelagem dos comentários. (2) Extração de Aspectos, em que os trabalhos buscam extrair características latentes que resumam os comentários em baixas dimensões.
3 - Objetivo do Modelo: A terceira perspectiva diz respeito ao objetivo do modelo de recomendação: (1) de predição, que efetivamente realizam recomendações; (2) de explanação, que visam explicar as recomendações realizadas e; (3) híbridos, que fornecem recomendações e suas respectivas explicações. Em nosso trabalho focamos nos modelos preditivos e híbridos.
4 - Arquitetura de Aprendizado: A quarta e perspectiva diz respeito ao processo de aprendizado em si. A maioria dos trabalhos utilizam arquiteturas neurais para o processamento dos textos e/ou para realização das recomendações. Propomos a classificação dos trabalhos em 2 classes distintas. (1) Modelos não-neurais, que utilizam técnicas de fatorização de matrizes ou modelagem de tópicos para realizarem recomendações. E os (2) modelos neurais, que podem utilizar arquiteturas como: (a) Redes convolucionais que utilizam camadas convolucionais como componentes principais no processo de aprendizagem; (b) Modelos de atenção, que utilizam técnicas de atenção neural para diferenciar partes dos textos; (c) Modelos recorrentes, que utilizam redes neurais recorrentes para extrair contexto; e (d) redes neurais baseadas em grafos que utilizam grafo junto com técnicas neurais para modelar as interações usuário-item.

4 Avaliação Sistemática

Essa seção visa responder as perguntas de pesquisa QP2 e QP3 que dizem respeito a como esses SsR vêm sendo avaliados por meio de uma inspeção das avaliações experimentais do dos 117 artigos selecionados.

4.1 Bases de dados

Geramos a distribuição das bases de dados na Figura 1. Podemos observar que as bases de dados da *Amazon*, a qual é correspondente a diversos produtos disponibilizados pela Amazon, desde livros, músicas e filmes, e a base da *Yelp*, a qual é referente a estabelecimentos, restaurantes e pontos turísticos nas principais cidades dos Estados Unidos (EUA), são as coleções de dados mais utilizadas no cenário de sistemas de recomendação baseado em comentário por uma notável diferença do terceiro lugar em diante. Ambas tratam-se de cenários clássicos para análises de recomendação *review aware* devido ao grande número de usuários que comentam sobre os itens. A Amazon é composta de subcoleções de acordo com categoria de item e a Yelp composta por diferentes cidades. A grande maioria dos trabalhos não específica qual categoria/cidade utilizada nos experimentos. Outra questão crítica é que essa coleção tem cortes temporais que também não são mencionados. Essas duas questões impactam negativamente na reprodutibilidade desses trabalhos.

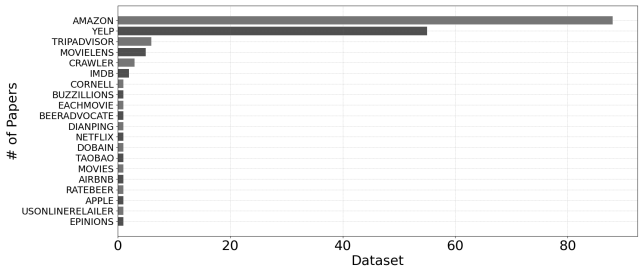

Figura 1. Frequência de bases de dados em experimentações.

4.2 Métricas

Conforme podemos observar na Figura 2, há um constante interesse sob a precisão dos *ratings* preditos. O consenso na comunidade de SR é que a precisão por si só não é suficiente para avaliar a eficácia prática e o valor agregado das recomendações, sendo necessário outras técnicas de avaliação como diversidade e serendipidade. Não identificamos nenhum trabalho dentre os 117 analisado que busca avaliar os modelos nessas dimensões, sendo esse um importante ponto fraco que avaliações futuras precisam considerar.

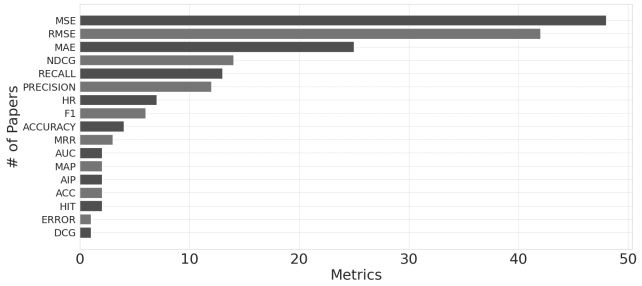

Figura 2. Frequência de métricas em experimentações.

4.3 Algoritmos e Configuração de Parâmetros

Nossa terceira análise sobre a configuração dos ambientes experimentais utilizados pelos trabalhos recentes da literatura está relacionada aos algoritmos utilizados. Com respeito a disponibilidade do código fonte dos algoritmos propostas, temos que pelo menos 50% dos trabalhos não disponibilizam código fonte de suas propostas, o que complica severamente a utilização dos mesmos como linhas de base em novas propostas. Um novo algoritmo é comparado, em média, com apenas três outros algoritmos e, no máximo, nove outros algoritmos.

Essas duas questões impactam diretamente na avaliação do real impacto de uma nova abordagem para o avanço da área. Outra observação importante diz respeito à documentação do processo de calibração dos algoritmos utilizados na experimentação (linhas de base e nova proposta). Menos de 30% dos trabalhos selecionados apresentam em detalhes os processos de calibração de parâmetros dos algoritmos considerados. Podemos observar que a grande maioria dos trabalhos que fornecem as especificações dos parâmetros dos algoritmos e detalhes do treinamento são algoritmos não-neurais, o que impacta diretamente na replicabilidade dos trabalhos que propõem inovações utilizando arquiteturas neurais, pois a calibragem de parâmetros trata-se de um processo essencial para se definir o sucesso ou insucesso de uma tarefa.

5 iRev

Nessa seção detalhamos nossa proposta de um framework de avaliação de SsR baseados comentários textuais: iRev (`https://github.com/guibitten03/iRev`).

5.1 Algoritmos Implementados

Dos 117 algoritmos examinados, selecionamos e implementamos todas as abordagens utilizadas em pelo menos dois artigos diferentes. Os 10 algoritmos selecionados são apresentados na Tabela 1, onde a coluna *'Linhas de Base'* representa quantas vezes cada algoritmo foi utilizado como linha de base em outros trabalhos.

Algoritmo	# Linhas de Base	# Citações
DeepCoNN	28	908
Narre	14	455
D-ATTN	9	421
Daml	7	123
MPCN	6	268
CARL	4	163
ANR	3	126
CARP	2	97
HRDR	2	75
RGNN	2	20

Tabela 1. Algoritmos mais utilizados como linhas de base.

O DeepCoNN utiliza redes neurais convolucionais para capturar as informações relevantes nos comentários Zheng *et al*. [2017]. O MPCN, por sua vez, emprega uma arquitetura de co-atenção multi-pontual para capturar o contexto em diferentes níveis de granularidade Tay *et al*. [2018]. O D-ATTN adota uma rede neural de atenção dupla considerando os comentários e as características do usuário/item Seo *et al*. [2017]. O NARRE utiliza uma abordagem baseada em redes neurais para modelar a atenção e as interações entre aspectos nos comentários Chen *et al*. [2018]. O DAML propõe uma abordagem de aprendizado mútuo de atenção entre avaliações e comentários Liu *et al*. [2019]. O CARL utiliza redes neurais convolucionais em cápsulas para gerar recomendações e fornecer explicações sobre as preferências do usuário Wu *et al*. [2019]. O CARP introduz uma estrutura de rede neural para incorporar a atenção contextual na modelagem de avaliações e comentários Li *et al*. [2019]. O ANR adota uma abordagem baseada em aspectos para recomendação, capturando a relação entre aspectos e usuários/itens Chin *et al*. [2018]. O HRDR realiza uma abordagem conjunta de representações de aprendizado profundo de avaliações e comentários Liu *et al*. [2020]. Por fim, o RGNN propõe uma representação hierárquica de comentários de avaliações em forma de grafo para aprimorar a precisão das recomendações Liu *et al*. [2021].

5.2 Configuração dos Algoritmos

Utilizamos os códigos dos algoritmos provenientes no GitHub dos respectivos autores e realizamos uma tunagem de parâmetros, de acordo com o apresentado na Tabela 2.

Parâmetros	Valores
Épocas de treinamento	10, 20, 50
Função de perda	MSE
Otimizador	ADAM
Dimensões dos vetores de usuário e item	32
Dimensões dos vetores de palavras	300
Codificadores utilizados	TF-IDF, Word2Vec e FastText.
Taxa de dropout	0.5
Weight decay	$1e^{-3}$
Tamanho do lote	128 a 32
Tamanho máximo dos documentos	500 palavras
Taxa de aprendizado	$2e^{-3}$
# filtros nas camadas convolucionais	100

Tabela 2. Configurações dos parâmetros dos algoritmos

5.3 Coleções de Dados

Conforme observado anteriormente, as bases da Amazon e da Yelp são as mais utilizadas nas avaliações experimentais. A Tabela 4 apresenta alguns detalhes sobre cada coleção. Para os experimentos, dividimos os dados em conjuntos de treino, teste e validação. O conjunto de treino foi composto por 80% dos dados, enquanto os conjuntos de validação e teste possuem 10% cada. Os *reviews* presentes nos dados foram pré-processados utilizando a biblioteca NLTK, que permitiu realizar tratamentos nos textos, tais como remoção de *stopwords* e lematização. Buscamos manter o pré-processamento fiel ao proposto nos respectivos artigos afim de obtermos resultados fidedignos de cada recomendador.

5.4 Métricas

Para avaliar as recomendações dos algoritmos consideramos quatro métricas de precisão: duas métricas de erro (i.e., MSE e MAE) que avaliam a diferença entre o *rating* real e o previsto pelos algoritmos; e duas de efetividade (i.e. Accuracy e F1 Score) que avaliam o quão bem o algoritmo aprendeu o comportamento do usuário. O consenso na comunidade de SsR é que a precisão por si só não é suficiente para avaliar a eficácia prática e o valor agregado das recomendações. Assim, além da precisão, exclusivamente considerada em praticamente todos revisados neste artigo, consideramos outras duas métricas: serendipidade e diversidade. A serendipidade se refere a descoberta de itens úteis e inesperados e

Coleção	Amazon - Video Games						Yelp - Tampa						Yelp - Philadelphia					
Measure	MSE	MAE	Acc	F1@10	Ser	Div	MSE	MAE	Acc	F1@10	Ser	Div	MSE	MAE	Acc	F1@10	Ser	Div
DeepCoNN	1.541	0.928	0.206	0.323	0.141	**0.197**▲	1.337	0.892	**0.361**▲	0.216	0.147	0.097	1.561	0.994	0.300	0.125	**0.159**●	**0.224**●
D ATTN	1.127	**0.727**▲	0.228	0.428	0.048	0.060	1.346	0.912	0.329	0.197	0.159	0.088	1.207	0.871	0.343	0.198	0.141	0.099
MPCN	1.636	0.993	0.121	0.262	0.069	0.1598	1.447	0.965	0.288	0.124	**0.164**●	0.163	1.322	0.913	0.323	0.121	0.145	0.125
NARRE	1.075	0.691	0.255	**0.459**▲	0.061	0.141	1.302	0.892	0.348	0.218	0.147	0.049	1.172	0.844	0.364	0.218	0.146	0.063
DAML	1.149	0.744	0.234	0.411	0.035	0.094	1.364	0.935	0.308	0.177	0.146	0.037	1.270	0.899	0.329	0.173	0.152	0.037
CARL	1.286	0.839	0.326	0.152	0.021	0.081	1.525	0.995	0.293	0.134	**0.166**●	0.125	1.306	0.921	0.332	0.171	**0.153**●	**0.224**●
CARP	1.262	0.824	0.340	0.170	0.213	0.124	1.583	0.987	0.285	0.107	0.147	**0.225**▲	1.336	0.902	0.324	0.136	0.152	0,172
ANR	1.171	0.780	0.381	0.226	**0.242**▲	0.071	1.288	0.899	0.338	0.215	0.148	0.061	**1.115**▲	**0.813**▲	**0.397**▲	**0.267**▲	0.124	0.050
HRDR	**1.039**▲	0.751	**0.456**▲	0.248	0.082	0.081	**1.257**▲	**0.879**▲	0.355	**0.249**▲	0.146	0.050	1.482	0.964	0.324	0.209	0.146	0.057
RGNN	1.179	0.792	0.297	0.150	0.101	0.148	1.364	0.916	0.314	0.186	0.142	0.124	1.296	0.896	0.287	0.128	0.150	0.075

Tabela 3. Resultados validados com o teste de Wilcoxon com um valor p = 0,05. ▲ representa ganhos significativos e ● empates estatísticos.

Coleção	# Usuários	# Itens	Esparsidade
Amazon - Video Games	10.000	17.005	99.99%
Yelp - Tampa	18.437	8.664	99,99%
Yelp - Philadelphia	32.376	14.226	99.99%

Tabela 4. Visão geral das coleções utilizadas na avaliação.

a diversidade aos itens recomendados diferentes do histórico de consumo. Todas elas estão descritas na Tabela 5.

Métrica	Fórmula
MSE	$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
MAE	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Accuracy	$\frac{TP+TN}{TP+TN+FP+FN}$
F1 Score	$\frac{2 \cdot precision \cdot recall}{precision+recall}$
Serendipidade	$\frac{\text{Itens recomendados inéditos}}{\text{Total de itens recomendados}}$
Diversidade	$1 - \frac{sim(\text{Total de pares de itens diferentes})}{\text{Total de pares de itens comparados}}$

Tabela 5. Definição das métricas utilizadas na avaliação.

5.5 Avaliação Experimental

Com o objetivo de validar o framework proposto, realizamos uma avaliação experimental de todos os algoritmos implementados, considerando as cinco métricas nas três coleções disponibilizadas e os resultados são apresentados na Tabela 3 apresentamos os resultados obtidos. Observamos que não há um destaque único. Na coleção Amazon, por exemplo, dos 10 algoritmos analisados, cinco deles se destacaram em distintas métricas. Enquanto os algoritmos HRDR, D-ATTN e o NARRE se destacaram em métricas de precisão, o DeepCONN e o ANR se destacaram em diversidade e serendipidade.

Os resultados também variam de acordo com as coleções. Na Yelp - Tampa, o segundo algoritmo mais recente proposto, HRDR, obteve melhores resultados nas métricas de precisão, corroborando com os experimentos mencionados no artigo original. Além disso, observou-se que os algoritmos CARP, CARL e MPCN não obtiveram resultados significativos nessas mesmas métricas, o que contradiz as afirmações de seus respectivos artigos. Por outro lado, esses algoritmos foram destaque em termos de serendipidade e diversidade. O ANR não obteve o melhor resultado em nenhuma métrica, mas apresentou resultados consistentes e estáveis em termos de desempenho. O algoritmo RGNN, embora seja o mais recente em termos de proposta, apresentou resultados inferiores a muitas outras estratégias avaliadas. Na coleção Yelp - Philadelphia, o algoritmo ANR mostrou os melhores resultados nas quatro métricas relacionadas à efetividade, mais uma vez alinhados com o que foi apresentado no artigo original. O algoritmo CARL foi melhor nessa coleção em comparação com a anterior, obtendo os melhores resultados de serendipidade e diversidade, juntamente com o DeepCoNN.

Grande parte dos algoritmos apresentaram resultados consistentes com seus estudos originais. No entanto, alguns algoritmos não obtiveram bons resultados em comparação com o que foi descrito pelos autores. Algoritmos como CARL, CARP, MPCN e RGNN, que teoricamente deveriam superar metodologias como DATTN, ANR e NARRE, não tiveram sucesso em nossa avaliação empírica. Esses resultados reforçam a importância do iRev por avançar na questão da reprodutibilidade, por ser um repositório público não apenas de coleções de dados, como também dos próprios algoritmos e seus processos de tunagem de parâmetros.

6 Conclusões e Trabalhos Futuros

Como primeira contribuição, esse trabalho apresenta um mapeamento sistemático dos estudos sobre sistemas de recomendação *review-aware* (RARSs) selecionando e investigando os 117 artigos relevantes publicados nos principais veículos da área (e.g., RecSys, SIGIR, WWW, etc.), identificando esforços, resultados, contribuições e limitações relevantes. A partir desse levantamento, propomos e disponibilizamos um framework, denominado iRev, contendo a implementação dos 10 principais RARSs, bem como todos os artefatos levantados durante o mapeamento sistemático (métricas e bases de dados) com o intuito de mitigar a limitação atual de falta de reprodutibilidade devido à ausência de códigos fontes e de confiabilidade devido à ausência de distintas métricas de avaliação. Para validar o iRev, realizamos uma avaliação completa das principais abordagens, considerando diferentes coleções de dados e métricas. Nossos resultados mostram que as SsR baseadas em redes neurais, especialmente as que utilizam mecanismos de extração de atenção e aspecto, obtiveram os resultados mais competitivos. Por outro lado, tais resultados também reforçam que não há um único algoritmo que se destaque de forma absoluta, deixando claro que ainda há espaço de melhora considerável a ser explorado por novas estratégias. Como trabalhos futuros, visamos complementar o iRev com a implementação de outras estratégias, tornando o repositório uma referência para que pesquisadores da área.

Declarações complementares

Financiamento

Este trabalho foi financiado por CNPq, CAPES, Fapemig, FAPESP e AWS.

Contribuições dos autores

Todas as implementações, execuções dos resultados foram realizadas pelo aluno Guilherme Bittencourt, com auxílio do aluno Naan Vasconcelos, sob a orientação do professor Leonardo Rocha. A concepcção do projeto e as análises de resultados foram feitas em conjunto, alunos e professor.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesse.

Disponibilidade de dados e materiais

Os conjuntos de dados (e/ou softwares) gerados e/ou analisados durante estudo atual serão feitos mediante solicitação.

Referências

Bittencourt, G., Fonseca, G., Andrade, Y., Silva, N., and Rocha, L. (2023). A survey on review - aware recommendation systems. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web*, WebMedia '23. DOI: 10.1145/3617023.3617050.

Chen, C., Zhang, M., Liu, Y., and Ma, S. (2018). Neural attentional rating regression with review-level explanations. In *Proceedings of the 2018 world wide web conference*, pages 1583–1592. DOI: 10.1145/3178876.3186070.

Chin, J. Y., Zhao, K., Joty, S., and Cong, G. (2018). Anr: Aspect-based neural recommender. In *27th ACM CIKM*, pages 147–156. DOI: 10.1145/3269206.3271810.

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 *42nd ACM SIGIR*, pages 275–284. DOI: 10.1145/3331184.3331216.

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 *25th ACM SIGKDD*, pages 344–352. DOI: 10.1145/3292500.3330906.

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. DOI: 10.1016/j.neucom.2019.09.052.

Liu, Y., Yang, S., Zhang, Y., Miao, C., Nie, Z., and Zhang, J. (2021). Learning hierarchical review graph representations for recommendation. *IEEE TKDE*, 35(1):658–671. DOI: 10.1109/TKDE.2021.3075052.

Seo, S., Huang, J., Yang, H., and Liu, Y. (2017). Interpretable convolutional neural networks with dual local and global attention for review rating prediction. In *11st ACM RecSys*, pages 297–305. DOI: 10.1145/3109859.3109890.

Tay, Y., Luu, A. T., and Hui, S. (2018). Multi-pointer co-attention networks for recommendation. In *24th ACM SIGKDD*, pages 2309–2318. DOI: 10.1145/3219819.3220086.

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. DOI: 10.1145/3428658.3430970.

Wu, L., Quan, C., Li, C., Wang, Q., Zheng, B., and Luo, X. (2019). A context-aware user-item representation learning for item recommendation. *ACM TOIS*, 37(2):1–29. DOI: 10.1145/3298988.

Zheng, L., Noroozi, V., and Yu, P. (2017). Joint deep modeling of users and items using reviews for recommendation. In *ACM WSDM*, pages 425–434. DOI: 10.1145/3018661.3018665.