HPDM: A Hierarchical Popularity-aware Debiased Modeling Approach for Personalized News Recommender

Xiangfu He[1,2] , **Qiyao Peng**[*3] , **Minglai Shao**[3] and **Hongtao Liu**[4]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Ximalaya Inc., ShangHai, China
[3]School of New Media and Communication, Tianjin University, Tianjin, China
[4]Du Xiaoman Financial Technology, Beijing, China
hexiangfu1012, qypeng, shaoml@tju.edu.cn, liuhongtao01@duxiaoman.com

Abstract

News recommender systems face inherent challenges from popularity bias, where user interactions concentrate heavily on a small subset of popular news. While existing debiasing methods have made progress in recommendation, they often overlook two critical aspects: the different granularity of news popularity (across titles, categories, etc.) and how hierarchical popularity levels distinctly influence user interest modeling. Hence, in this paper, we propose a hierarchical causal debiasing framework that effectively captures genuine user interests while mitigating popularity bias at different granularity levels. Our framework incorporates two key components during training: (1) a hierarchical popularity-aware user modeling module to capture user interests by distinguishing popular and unpopular interactions at different granularity news content; and (2) a dual-view structure combining counterfactual reasoning for popular-view news with inverse propensity weighting for unpopular-view news to model user genuine interests. During inference, our framework removes popularity-induced effects to predict relatedness between user and candidate news. Extensive experiments on two widely-used datasets, MIND and Adressa, demonstrate that our framework significantly outperforms existing baseline approaches in addressing both the long-tail distribution challenge. Our code is available at https://github.com/hexiangfu123/HPDM.

1 Introduction

News recommender systems have emerged as a fundamental technological approach for tackling information overload [Wu *et al.*, 2019b; Wu *et al.*, 2019a; An *et al.*, 2020]. While these systems primarily aim to precisely model user interests based on historical click behaviors, they face significant challenges stemming from the inherent nature of news distribution. Specifically, user interaction data typically exhibits a pronounced long-tail distribution [Zheng *et al.*, 2021; Wei *et al.*, 2021], where a majority of user engagements concentrate on a small subset of popular news. This skewed pattern would introduce potential biases in user modeling, as models may overly emphasize popular news when learning user interests, thereby compromising the effectiveness of personalized recommendations.

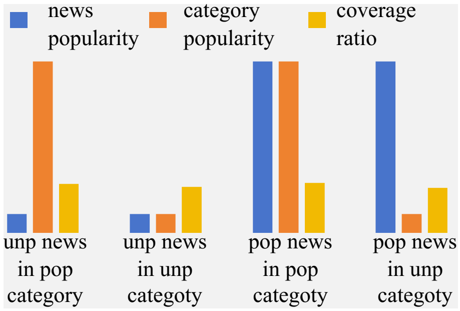


Figure 1: The proportions of four types of news with different popularity on MIND dataset.

While progress has been made in alleviating the long-tail phenomenon across various recommendation domains [Wei *et al.*, 2021], such as PDA [Zhang *et al.*, 2021b] which could effectively de-confound negative bias effects during training while leveraging positive popularity signals at inference time, adapting these methods to news recommendation poses unique challenges:

First, news popularity manifests in a hierarchical manner across different granularity (e.g., titles and categories shown in Figure 1), with heterogeneous effects on user behaviors. Some users may be more susceptible to trending titles, while others are primarily influenced by popular categories. Furthermore, user interactions with news of varying popularity levels often stem from distinct motivations, i.e., clicks on popular news may largely reflect conformity effects, whereas engagement with less popular news typically indicates genuine interests. However, existing approaches typically adopt a simplified view of popularity and treat all user interactions uniformly, failing to capture nuanced relationships between hierarchical popularity patterns and user behaviors.

Second, modeling user interactions with news of different popularity levels needs to be treated differently. As shown in Figure 2, for popular news, user's real interest and conformity could both determine clicks on the news. Hence, the key challenge lies in disentangling genuine user interests from

[*]Corresponding author

the confounding effects of social influence. In contrast, unpopular news interactions, though sparse, often provide more reliable signals of personal preferences and how to leverage these valuable but limited data to model user is important. Current methods typically employ a uniform modeling strategy regardless of news popularity, overlooking these distinct characteristics and modeling requirements.

In this paper, we propose a hierarchical causal debiasing framework to address these challenges. First, we design a hierarchical popularity-aware user interest modeling approach that captures user preferences across different news granularities (from titles to categories). At each granularity level, we introduce a selective masking mechanism that distinctively processes popular and unpopular news interactions, enabling our model to construct more comprehensive user representations that reflect how popularity differently impacts individual users. For instance, when modeling user interests in popular news, we mask unpopular news from the user's click history, and vice versa. Second, we develop a dual-view causal debiasing structure to effectively capture genuine user interests. For popular news interactions, we employ counterfactual reasoning to identify and mitigate conformity effects by comparing observed user behaviors against hypothetical scenarios where clicks are solely influenced by popularity. For unpopular news interactions, we implement an inverse propensity weighting framework to address the exposure bias, which helps appropriately weight these valuable but potentially underrepresented interaction signals. During inference, our framework eliminates both the popularity-induced effects and inverse propensity weights used in training, enabling more accurate capture of genuine user interests while reducing the impact of hierarchical popularity biases.

The main contributions of this work are:

(1) We propose a hierarchical causal debiasing framework for news recommendation that effectively addresses long-tail distribution challenges while preserving genuine user interests.

(2) We propose a hierarchical framework that alleviate the popularity bias at different news granularities (from titles to categories) equipped with a selective masking mechanism. Besides, we design a differential causal debiasing structure that employs counterfactual reasoning for popular news to alleviate the conformity effect and inverse propensity weighting for unpopular news to capture genuine user interests.

(3) Extensive experiments on two widely-used benchmarks (MIND and Adressa) demonstrate our framework's superior performance over existing news recommendation and causal recommendation baselines.

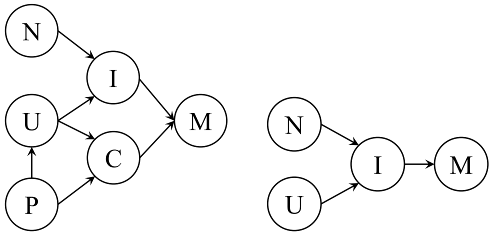
2 Related Work

2.1 News Recommendation

In the early stages of news recommendation research, scholars predominantly relied on hand-crafted feature engineering to construct user profiles [Li *et al.*, 2010; Garcin *et al.*, 2013; Son *et al.*, 2013].

Unlike manual features, Deep Neural Networks (DNNs) [Vaswani *et al.*, 2017] have the ability to learn textual features, which have been widely used in modeling user interest [Wu *et al.*, 2019b; Peng *et al.*, 2020; Peng *et al.*, 2022; Xu *et al.*, 2023]. For example, Wu et al. [Wu *et al.*, 2019b] introduced an innovative neural architecture leveraging multi-head self-attention mechanisms to capture the nuanced interplay between news content and user behavior; Recent studies have investigated the role of popularity in news recommendation systems. For instance, PP-Rec [Qi *et al.*, 2021a] integrated personalized matching with popularity scores through a unified model for popularity prediction, incorporating a personalized aggregator and knowledge-aware news encoder to enhance ranking performance.

However, these approaches treat news popularity uniformly, failing to distinguish between conformity effects in highly popular news interactions and genuine user interests reflected in low-popularity news engagement. This undifferentiated treatment leads to suboptimal user interest modeling.



(a) User's real interest and conformity both determine user click for popular news.

(b) Only user's real interest influence user click for unpopular news.

Figure 2: Causal graph for the popular and unpopular news. Each node represents: (N) News, (U) User, (P) Popularity, (I) Interest Matching, (C) Conformity Matching, (M) User Click.

2.2 Causal Recommendation

Real-world interactions are fundamentally governed by cause-and-effect relationships rather than simple correlations, as correlated events may not necessarily indicate causation [Gao *et al.*, 2022]. Recent research has shifted focus toward understanding various confounding effects and counterfactual analysis [Zhang *et al.*, 2021b]. For example, In addressing popularity bias for general recommendation systems, Zhang et al.[Zhang *et al.*, 2021b] introduced PDA (Popularity-bias Deconfounding and Adjusting), which removes bias during model training and adjusts recommendations through causal intervention at inference time. Additionally, Wang et al. [Wang *et al.*, 2022] proposed UCI (User-Controllable Inference), a framework that integrates causality to enable user alerts, control inputs, and flexible recommendation adjustment.

While these causal approaches have demonstrated promising results in general recommendation scenarios, their direct application to news recommendation systems faces two critical challenges: 1) User-News interactions typically involve multiple engagement points, such as tags and titles. Existing methods lack the capability to differentiate and quantify how popularity at each level influences user interactions; 2) User interactions with news of varying popularity levels carry different implications. Interactions with popular news may primarily stem from conformity effects or social influence, while engagement with less popular news might better reflect

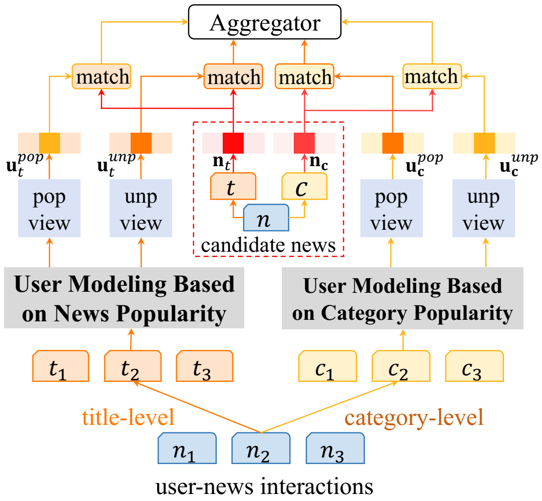


Figure 3: Hierarchical Popularity-aware User Interest Modeling.

genuine user interests. Most existing methods usually treat them equally to model the impact of popularity on user-news interaction.

3 Problem Definition

Given a user u and a news n, the news recommender system aims to calculate the matching score s to evaluate the likelihood of user u clicking on news n. Subsequently, the system ranks all candidate news based on the matching score s and recommends the top-ranked news to the user n. Each news n consists of title $n^t = [w_1, w_1, \cdots, w_L]$ with L words and a specific category, denoted as n^c. Each user u has a history list of behaviors $u_b = [n_1, n_2, \cdots, n_M]$ with the list length of M, representing the news he/she has previously clicked on.

4 Proposed Method

In this section, we introduce our model's overall structure in detail. Specifically, the proposed method consists of three core modules: a *Hierarchical Popularity-aware User Interest Modeling* to learn hierarchical user interests with different news popularity, shown in Figure 3, a *Dual-track Causal Debiasing* to capture genuine user interests, shown in Figure 5, and an *Adaptive Score Aggregator* to predict the final user-news matching score.

4.1 Hierarchical Popularity-aware User Interest Modeling

In this section, we propose to capture personalized popularity-aware user interests. Our method consists of two key components: (1) a hierarchical news encoder that learns comprehensive news representations by incorporating both title and category information, and (2) a popularity-aware user modeling mechanism that leverages a selective masking strategy to distinguish between popular and unpopular news interactions across different news granularities.

Hierarchical News Encoder In our method, news encoder aims to learn news representations. Previous works [Wu *et al.*, 2023] usually use a variety of combinations of news information (e.g., title, category, sub-category, entity, etc.) to learn news representations, which have demonstrated superior performance. In fact, different aspects of news content, specifically titles and categories, represent distinct perspectives of the same news, each with notably different popularity distributions (as illustrated in 1). These varying popularity patterns would significantly influence user engagement in distinct ways. For instance, some users might be drawn to news based on popular categories, while others might be more influenced by trending titles.

Hence, we design a hierarchical news encoder that utilizes the Word2Vec [Mikolov *et al.*, 2013] (denoted as enc_w) to learn the news category embedding and the BERT [Devlin *et al.*, 2018]) (denoted as enc_b) to learn the news title embedding. Thus, the x-th news embedding of title and category with d dimension can be formulated as:

$$\mathbf{n}_x^t = \mathsf{enc}_b(n^t), \; \mathbf{n}_x^c = \mathsf{enc}_w(n^c). \tag{1}$$

In this way, the model could capture news features at different granularities, which helps model users' diverse interests driven by the varying popularity of different news content.

Popular-aware User Representation Based on the learned news embedding $\mathbf{n}^t$ and $\mathbf{n}^c$, a straightforward approach is to encode the category-aware user representation and the title-aware user representation separately and then concatenate them, which is denoted as:

$$\mathbf{u}^t = \mathsf{enc}_u([\mathbf{n}_1^t, \cdots, \mathbf{n}_M^t]), \mathbf{u}^c = \mathsf{enc}_u([\mathbf{n}_1^c, \cdots, \mathbf{n}_M^c]), \tag{2}$$

$$\mathbf{u} = [\mathbf{u}^t; \mathbf{u}^c], \tag{3}$$

where $[\mathbf{n}_1, \cdots, \mathbf{n}_M]$ denotes the embeddings of news in the user's click history and $[\cdot;\cdot]$ is the concatenation, and $\mathsf{enc}_u(\cdot)$ is the user encoder.

However, this approach overlooks a critical aspect: the varying importance of news with different popularity levels in modeling user preferences. User engagement patterns vary fundamentally based on news popularity, i.e., clicks on popular news often result from a combination of conformity effects and genuine interest, while engagement with unpopular news typically indicates authentic personal preferences. Moreover, users exhibit different sensitivity to popularity across content granularities. Some users are more likely to click news due to trending titles, while others are primarily influenced by the popularity of news categories. This heterogeneous popularity influence necessitates separate modeling of user interests at both title and category levels.

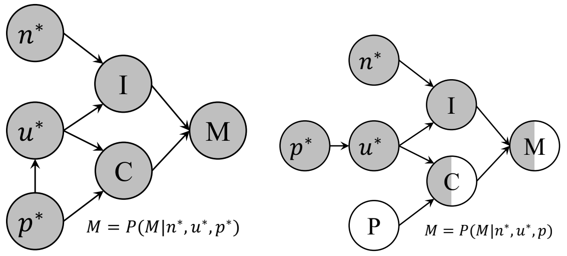
To alleviate this, we propose to learn user representations by considering popularity differences at both title and category levels. Specifically, we employ a selective masking strategy that separates popular from unpopular content within each granularity. For example, for news's title embeddings, we could categorize the news in click history $[n_1, n_2, n_3, n_4]$ into two classes:

$$[n_1^{pop}, n_2^{unp}, n_3^{pop}, n_4^{unp}], \tag{4}$$

in which *pop* and *unp* denote the news classified as popular or unpopular based on news popularity.

Then, based on the popularity of news title, we can obtain the title-aware hierarchical user interest, which could be calculated as follows:

$$\mathbf{u}_t^{pop} = \mathsf{enc}_u([\mathbf{n}_1^t, \texttt{[MASK]}, \mathbf{n}_3^t, \texttt{[MASK]}]), \tag{5}$$



(a) The causal effect, with the variables at the reference status: news n^*, user u^* and popularity p^*.

(b) The counterfactual world, in which user click-through intentions are solely influenced by popularity.

Figure 4: Examples of two causal graphs, where the gray nodes represent the variables at reference.

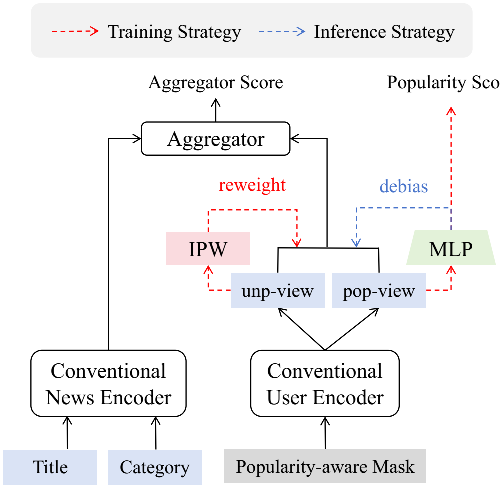


Figure 5: Model training and inference.

$$\mathbf{u}_t^{unp} = \mathsf{enc}_u([\texttt{MASK}], \mathbf{n}_2^t, [\texttt{MASK}], \mathbf{n}_4^t]), \tag{6}$$

where $\mathbf{n}_i^t$ denotes the title embeddings of the i-th news.

Similarly, we can obtain the category-aware hierarchical user interest representations $\mathbf{u}_c^{pop}$ and $\mathbf{u}_c^{unp}$.

Finally, four user's interest representations from four different views would be obtained: (1) popular news title view $\mathbf{u}_t^{pop}$; (2) unpopular news title view $\mathbf{u}_t^{unp}$; (3) popular news category view $\mathbf{u}_c^{pop}$; (4) unpopular news category view $\mathbf{u}_c^{unp}$.

4.2 Dual-view Causal Debiasing

While our previous modeling approach successfully captures user interests across diverse news content and popularity levels, challenges remain in obtaining unbiased interest representations. Specifically, two challenges persist: highly popular news can introduce conformity bias through herd effects, while less popular news often receives insufficient attention during model training due to limited user interactions. To effectively address these contrasting challenges, we propose a novel differential causal debiasing framework including: (1) counterfactual debiasing for popular news to mitigate conformity effects, and (2) inverse propensity weighting for unpopular news to enhance the influence of limited interactions.

Counterfactual Debiasing for Popular News In fact, the news contents and user interest determine the user-news interest matching. Similarly, the interplay between news popularity and user conformity shapes user conformity matching. And the final user-news clicking probability is determined by user conformity and interest. Notably, popularity's influence extends beyond immediate interactions, it plays a crucial role in shaping the user's historical click behavior, so there is an edge from Popularity to User. The causal graph is illustrated in Figure 2 (a).

Within our causal framework, the counterfactual framework aims to mitigate the impact of users' herd mentality on their clicks under the popular view. A feasible solution is to eliminate the influence of popularity within the causal graph. This is because when popularity no longer affects users' clicks, the causal effect generated by path $(U, P) \to C \to M$ is determined solely by the users (which can be regarded as the average click-through rate of users). In the ranking task, for all candidate news items, this influence brought solely by users is the same and will not affect the final ranking result. Therefore, eliminating the effect of Popularity (P) would result in the ranking not being interfered with by users' herd mentality.

There are two distinct conformity-related pathways: indirect path $P \to U \to C$ and direct path $P \to C$. To analyze conformity causal influence, we first investigate the direct causal effect of all factors (i.e., P, N, U). We compare the aggregated score with the reference world, shown in Figure 4 (a), where all factors are set to the average of corresponding vectors. Hence, the Total Effect (TE) is employed to measure the causal effects of all factors, which is defined as:

$$TE = P(M|u, n, p) - P(M|u^*, n^*, p^*)\,. \tag{7}$$

Then, according to the definition, TE can be decomposed into two components: Total Indirect Effect (TIE) and Natural Direct Effect (NDE). NDE represents the influence by the path $P \to C$, which means the difference between the origin variable p and reference variable p^* without the influence of user and news:

$$NDE = P(M|u^*, n^*, p) - P(M|u^*, n^*, p^*)\,. \tag{8}$$

And hence, TIE could be computed by subtracting NDE from TE:

$$TIE = TE - NDE = P(M|u, n, p) - P(M|u^*, n^*, p)\,, \tag{9}$$

where TIE represents the indirect effect ($P \to U \to C$) without the influence of popularity.

We can find that $P(M|u, n, p)$ represents the final prediction score, and $P(M|u^*, n^*, p)$ is a counterfactual result, as demonstrated in Figure 4 (b), which means whether the user will click the news only based on the popularity. As news content and user interest at the reference status, we could assume that the counterfactual result is only determined by news popularity and user conformity:

$$P(M|u^*, n^*, p) \propto s_{Pop} = (pop_u * pop_n) + \bar{\mathbf{u}} \cdot \bar{\mathbf{n}}\,, \tag{10}$$

where pop_n is the real popularity of the news n. $\bar{\mathbf{u}}$ and $\bar{\mathbf{n}}$ denote the mean of all the user and news vectors, which could be ignored. To quantitatively assess the causal impact of conformity on user clicks, we apply the MLP to predict the conformity score pop_u of user u, formulated as:

$$pop_u = MLP(\mathbf{u}_t^{pop}) + MLP(\mathbf{u}_c^{pop}) \ . \tag{11}$$

In this way, we can enable the estimation of model parameters for both $P(M|u,n,p)$ and $P(M|u^*,n^*,p)$ through supervised training, as detailed in Section 4.4. Through the counterfactual learning on the user popular-view histories, the model capture more genuine interests without conformity.

Inverse Propensity Weighting for Unpopular News In the context of unpopular news, user engagement patterns differ significantly from those of popular content. As depicted in Figure 2 (b), users' interactions with unpopular news tend to be driven more by genuine interest rather than conformity bias. However, this presents a challenge in our training: unpopular news receives limited exposure in our training data, potentially leading to under representation in the model learning process. Besides, the popularity influences exposure rates, creating an imbalance between exposed (treatment) and unexposed (control) news.

To address this bias, we adopt an Inverse Probability Weighting (IPW) approach. We propose a weighting mechanism that adjusts for this imbalance by calculating an exposure-based propensity score for each user:

$$Ratio_{imp}(u) = \left(\frac{\sum_{i \in \mathcal{D}_{unp}^{u}} impr_{n_i}}{\sum_{j \in \mathcal{D}} impr_{n_j}}\right)^{\gamma}, \tag{12}$$

where $impr_{n_i}$ denotes the impression number of news n_i, γ is the hyper-parameter, $\mathcal{D}_{unp}^{u}$ and $\mathcal{D}$ represent the unpopular news set for the user u's click and total news dataset.

Finally, each user-news match score calculated from the unpopular news embeddings would be re-weighted by the following approach:

$$W_u = \frac{1}{Ratio_{imp}(u)}. \tag{13}$$

In this way, the model could effectively compensate for the exposure bias by giving more weight to interactions with unpopular news items, thereby improving the model's ability to capture genuine user interests.

4.3 Adaptive Score Aggregator

Given the four different view user embeddings, $\mathbf{u}_t^{pop}$, $\mathbf{u}_t^{unp}$, $\mathbf{u}_c^{pop}$, $\mathbf{u}_c^{unp}$, we compute two title-aware matching scores based on the candidate news title representation $\mathbf{n}_t^{candidate}$ and two category-aware matching scores based on the candidate news category representation $\mathbf{n}_c^{candidate}$. Then, we design an adaptive score aggregator to personalized fuse these four scores to obtain the final user-news matching score. For simplicity, we denote the representations of different view user representations as $\mathbf{u}$, which could be written as:

$$\mathbf{u} \in \{\mathbf{u}_t^{pop}, \mathbf{u}_c^{pop}, \mathbf{u}_t^{unp}, \mathbf{u}_c^{unp}\}, \tag{14}$$

The final fusion matching score $P(M|u,n,p)$ could be denoted by:

$$s_{Agg} = \sum W_s(\mathbf{u}) * W_u * (\mathbf{u} \cdot \mathbf{n}), \tag{15}$$

where $W_s(\mathbf{u})$ is a MLP and softmax layer to compute the fusion weights. W_u is obtained via Eq.13 when $\mathbf{u} \in \{\mathbf{u}_t^{unp}, \mathbf{u}_c^{unp}\}$ for unpopular-view news and equals 1 when $\mathbf{u} \in \{\mathbf{u}_t^{pop}, \mathbf{u}_c^{pop}\}$ for popular-view news.

4.4 Training & Inference

Training As denoted above, the model parameters for both $P(M|u,n,p)$ and $P(M|u^*,n^*,p)$ need to be trained.

For $P(M|u,n,p)$, we utilize the cross-entropy objective following the previous works,, which is denoted as:

$$\mathcal{L}_{Agg} = -\sum_{s_{Agg}}^{|\mathcal{D}|} log \frac{exp(s_{Agg}^{+})}{exp(s_{Agg}^{+}) + \sum_{\mathcal{D}_u} exp(s_{Agg}^{-})}, \tag{16}$$

where s_{Agg}^{+} and s_{Agg}^{-} final matching scores of positive and negative samples. $\mathcal{D}$ and $\mathcal{D}_u$ represent the training dataset and candidate negative news set for the user u.

Similar to Eq. 16, we still apply the cross-entropy objective for training $P(M|u^*,n^*,p)$, which is denoted as:

$$\mathcal{L}_{Pop} = -\sum_{s_{Pop}}^{|\mathcal{D}|} log \frac{exp(s_{Pop}^{+})}{exp(s_{Pop}^{+}) + \sum_{\mathcal{D}_u} exp(s_{Pop}^{-})}. \tag{17}$$

Then, we utilize a multi-task learning paradigm with a hyperparameter η to train these two loss functions, which is defined as:

$$\mathcal{L} = \eta * \mathcal{L}_{Pop} + \mathcal{L}_{Agg}. \tag{18}$$

Inference As previously analyzed, the popularity effect (path $P \rightarrow C \rightarrow M$) for the popular news should be removed for inference. The actual score without the popularity effect could be accessed via Eq. 9. $P(M|u,n,p)$ is the click probability determined by user u, news n, and popularity p, which is the model output S_{Agg}. Besides, $P(M|u^*,n^*,p)$ is affected only by popularity with the user and news at reference. It can be simplified like Eq. 10, denoted as:

$$P(M|u^*,n^*,p) \propto s_{Pop} = (pop_u * pop_n) + C \ , \tag{19}$$

where C is an invariant constant derived from the reference status of user and news, which could be ignored in inference.

Moreover, since the IPW aims to re-weight the loss function to mitigate the sample bias, we remove the IPW weight in Eq. 15 for inference. The aggregating score for inference could be written as:

$$s_{Agg} = \sum W_s(\mathbf{u}) * (\mathbf{u} \cdot \mathbf{n}) \ . \tag{20}$$

Hence, the final matching score without popularity interference can be calculated by the following:

$$TIE = P(M|u,n,p) - P(M|u^*,n^*,p) = s_{Agg} - \tau * s_{Pop} \ , \tag{21}$$

where τ is the hyper-parameter.

4.5 Time Complexity

The HPDM user modeling framework introduces an MLP layer and an aggregator module (with time complexity $\mathcal{O}(5d^2)$, where d denotes the embedding dimension) to the base user encoder. Consequently, the overall time complexity of our

Datasets	MIND				Adressa			
Methods	AUC	MRR	NDCG@5	NDCG@10	AUC	MRR	NDCG@5	NDCG@10
NRMS	66.10	31.48	34.35	40.56	75.31	42.24	44.66	48.46
NAML	65.72	30.63	34.01	40.08	73.09	44.27	43.51	50.02
LSTUR	66.42	31.67	34.49	41.23	68.37	38.76	38.11	44.33
FIM	65.98	30.78	34.13	40.49	76.58	44.18	46.88	50.67
Hie-Rec	67.18	31.69	35.32	41.61	78.67	47.22	48.72	56.67
PP-Rec	67.36	33.25	36.04	42.77	79.32	48.54	48.93	56.71
UNBERT	67.38	31.93	34.72	41.34	79.18	46.44	48.45	54.44
NRMS-BERT	68.84	33.07	36.81	42.98	78.24	47.17	48.46	54.84
IPS	68.25	33.15	36.69	42.96	78.16	46.72	48.19	56.90
MACR	68.99	33.65	36.29	42.53	79.05	47.14	48.18	56.99
PPAC	68.20	34.04	37.16	42.98	79.42	47.6	48.65	57.48
PA	69.14	33.1	36.43	42.97	79.11	47.18	49.01	56.72
PAD	69.31	33.59	37.20	43.62	79.42	47.12	48.98	56.71
HPDM	**70.79**	**34.58**	**38.24**	**44.69**	**80.53**	**47.63**	**49.4**	**57.45**

Table 1: The main experiment results compared with the baselines. Best results are in **bold**.

Datasets	MIND	Adressa
#News	65,238	20,428
#Users	94,057	640,503
#Categories	20	24
#Impressions	230,117	-
#Clicks	347,727	3,101,991

Table 2: Statistics of the datasets.

framework depends on the choice of news and user encoders. For instance, the NAML architecture exhibits $\mathcal{O}(Md + Nd^2)$ complexity; NRMS demonstrates $\mathcal{O}(Md + 3Nd^2 + Nd^2)$ complexity; Our HPDM framework achieves $\mathcal{O}(Md+3Nd^2+Nd^2 + 5d^2)$ complexity, where M represents the candidate news number, and N denotes the length of the user's historical news sequence. Therefore, we argue that HPDM maintains acceptable inference efficiency.

5 Experiments

5.1 Experiment Setting

We evaluate our model on two real-world benchmark datasets: MIND [Wu *et al.*, 2020] and Adressa [Gulla *et al.*, 2017]. Detailed statistics of both datasets are presented in Table 2.

We compare our method with several benchmark baselines. We select two representative groups of baselines for evaluation. The first group is the traditional news recommendation models, including: (1) NRMS [Wu *et al.*, 2019b]; (2) NAML [Wu *et al.*, 2019a]; (3) LSTUR [An *et al.*, 2020]; (4) FIM [Wang *et al.*, 2020]; (5) Hie-Rec [Qi *et al.*, 2021b]; (6) PP-Rec [Qi *et al.*, 2021a]; (6) UNBERT [Zhang *et al.*, 2021a]; (7) NRMS-BERT [Wu *et al.*, 2021].

The second group consists of several model-agnostic methods of popularity debias in the recommender system: (1) IPS [Liang *et al.*, 2016]; (2) MACR [Wei *et al.*, 2021]; (3) PPAC [Ning *et al.*, 2024]; (4) PA & PAD [Zhang *et al.*, 2021c].

Following the previous works, we choose four widely used metrics to evaluate the performance of our method, including AUC, MRR, NDCG@5, and NDCG@10.

5.2 Main Results

The main results of our model are illustrated in Table 1. Based on these results, we have the following observations:

Firstly, the bert-based model (e.g., UNBERT, NRMS-BERT) usually demonstrates better performance than other models without bert (e.g., NRMS, NAML). This is because the extensive pre-training on large-scale corpora enables bert to capture rich linguistic knowledge and semantic representations, which serve as a strong foundation for understanding news and user preferences.

Secondly, our experimental results demonstrate that incorporating popularity debiasing mechanisms into news recommendation frameworks yields substantial performance improvements. Specifically, by augmenting the baseline models (e.g., NRMS-BERT) with various popularity debiasing frameworks (IPS, MACR, PA, and PAD), we observe consistent and significant enhancements across all evaluation metrics on both the MIND and Adressa datasets. This is because these models could help calibrate the training signals by reweighting samples based on their popularity, preventing the model from overfitting to popularity patterns and focusing more on news intrinsic content and user preferences.

Thirdly, our proposed HPDM framework demonstrates superior performance compared to existing popularity debiasing approaches across both MIND and Adressa datasets. This superior performance can be attributed to two key modules: This is because: 1) Unlike existing approaches that rely on single-dimensional popularity metrics, our method introduces hierarchical popularity patterns (title-level, category-level), enabling the model to capture more nuanced user interests. For instance, when a user clicks on a news with an unpopular title but popular category, HPDM can better discern whether this interaction stems from genuine topic interest or category-level conformity effects. 2) Rather than applying uniform popularity adjustments, HPDM utilizes the dual-view causal debiasing structure that considers the nuanced popularity patterns of both popular and unpopular news news. This granular approach

Methods	AUC	MRR	NDCG@5	NDCG@10
HPDM	70.79	34.58	38.24	44.69
w/o PV	68.13	32.57	36.81	43.28
w/o UV	67.92	32.52	36.7	42.74
w/o CF	69.51	33.69	37.29	43.82
w/o IPW	69.28	33.91	37.44	43.21
w/o NE	67.23	32.02	35.72	42.15
w/o CE	68.37	32.77	36.47	42.82

Table 3: The results of the ablation experiments.

enables more precise calibration of recommendation signals, leading to better alignment with users' genuine interests.

5.3 Ablation Study

In this section, we conduct ablation studies to evaluate the contribution of each component within our framework. For experiments, we use the MIND dataset as the example dataset, and design six variants: (1) w/o PV removes the popular-view user representations ($\mathbf{u}_t^{pop}$ and $\mathbf{u}_c^{pop}$) and the counterfactual module; (2) w/o UV removes the unpopular-view user representations ($\mathbf{u}_t^{unp}$ and $\mathbf{u}_c^{unp}$) and the inverse propensity weighting module; (3) w/o CF maintains user representations with different views but removes the counterfactual module; (4) w/o IPW maintains user representations with different views but removes the inverse propensity weighting module; (5) w/o NE represents removing the title-aware news encoder and only employing the category-aware news encoder; (6) w/o CE represents removing the category-aware news encoder and only employing the title-aware news encoder.

The experimental results are shown in Table 3. We can find that: 1) The removal of hierarchical popularity level architecture components leads to more substantial performance degradation, with the unpopular view (UV) and popular view (PV) modules showing significant AUC decreases to 67.92 and 68.13, respectively, while the removal of counterfactual reasoning (CF) and inverse probability weighting (IPW) modules resulted in relatively moderate decreases to 69.51 and 69.28. This hierarchical pattern suggests the fundamental importance of our dual-view architecture in capturing user preferences, where UV effectively extracts genuine interest signals from less popular news items, and PV models the conformity influence in clicking popular news; 2) The notable performance drops associated with the removal of debiasing strategies underscore their crucial role in mitigating popularity bias, with CF simulating counterfactual user behaviors to reduce conformity effects and IPW ensuring balanced attention distribution across the popularity spectrum; 3) The performance degradation of w/o NE and w/o CE indicate the different importance of title perspective and category perspective in modeling news representation and user interest respectively. In all, these findings validate our framework's architectural, which could effectively combine different view user representation for capturing hierarchical popularity patterns with debiasing mechanisms.

5.4 Hyper-parameter Analysis

We investigate two critical parameters in the framework: γ for scaling exposure rate in inverse probability weighting, and τ for adjusting the debiasing weight in the counterfactual framework. The optimal value of γ was observed around 0.1, striking a balance between mainstream and niche user interests, while τ achieved peak performance at approximately 0.12. Notably, in our experiments, we find that performance degrades at $\tau = 0$, which could demonstrate the necessity of counterfactual debiasing in news recommendation. Hence, we carefully set the γ as 0.1 and the τ as 0.12.

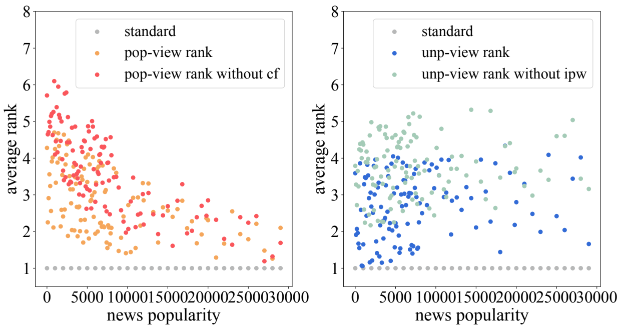


Figure 6: Visualization of popularity debiasing.

5.5 Popularity Debias Visualization

We use visualization to more intuitively analyze the debiasing effect of popularity from different views. In Figure 6, the standard (grey) represents the real click behaviors. Since these news have an interactive relationship with users, they need to be ranked first in terms of standard results. In the popularity-view shown in Figure 6 (a), we observe that "pop-view rank without cf" (red) exhibits excessive sensitivity to popularity signals, potentially due to the confounding effects. As the popularity decreases, the recommended ranking is lower, and the recommendation effect is worse compared with the standard results. Our approach (orange) achieves more balanced ranking scores across the popularity spectrum. Similarly, in the unpopularity view shown in Figure 6(b), "unp-view rank without ipw" approaches struggle with sparse interaction signals from less popular news. The integration of inverse propensity weighting in our framework (blue) could effectively address this challenge. These observations validate our motivation to design a dual-track framework.

6 Conclusion

In this paper, we investigated the critical issue of popularity bias in news recommendation systems, with emphasis on its hierarchical manifestation across different news granularities, from fine-grained titles to coarse-grained categories. Our framework introduces two key modules: a multi-granular user interest modeling approach with selective popularity-aware masking mechanism that distinctively processes interactions at different granularity levels to capture more comprehensive user preferences, and a dual-track causal debiasing structure that employs counterfactual reasoning for popular news while leveraging inverse propensity weighting for unpopular news to effectively disentangle genuine user interests from popularity-induced behaviors. Extensive experiments on the MIND and Adressa datasets demonstrated the superior performance, validating its effectiveness in capturing genuine user interests while mitigating popularity bias at different view levels.

References

[An *et al.*, 2020] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. Neural news recommendation with long- And short-term user representations. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 336–345, 2020.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2018.

[Gao *et al.*, 2022] Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. Causal inference in recommender systems: A survey and future directions. *ACM Transactions on Information Systems*, 2022.

[Garcin *et al.*, 2013] Florent Garcin, Christos Dimitrakakis, and Boi Faltings. Personalized news recommendation with context trees. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 105–112, 2013.

[Gulla *et al.*, 2017] Jon Atle Gulla, Lemei Zhang, Peng Liu, Özlem Özgöbek, and Xiaomeng Su. The adressa dataset for news recommendation. In *Proceedings of the International Conference on Web Intelligence*, page 1042–1048, New York, NY, USA, 2017. Association for Computing Machinery.

[Li *et al.*, 2010] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[Liang *et al.*, 2016] Dawen Liang, Laurent Charlin, and David M Blei. Causal inference for recommendation. In *Causation: Foundation to Application, Workshop at UAI. AUAI*, volume 6, page 108, 2016.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Internation Conference on Neural Information Processing Systems*, pages 3111–3119, 2013.

[Ning *et al.*, 2024] Wentao Ning, Reynold Cheng, Xiao Yan, Ben Kao, Nan Huo, Nur Al Hasan Haldar, and Bo Tang. Debiasing recommendation with personal popularity. In *Proceedings of the ACM on Web Conference 2024*, pages 3400–3409, 2024.

[Peng *et al.*, 2020] Qiyao Peng, Hongtao Liu, Yang Yu, Hongyan Xu, Weidi Dai, and Pengfei Jiao. Mutual self attention recommendation with gated fusion between ratings and reviews. In *Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part III 25*, pages 540–556. Springer, 2020.

[Peng *et al.*, 2022] Qiyao Peng, Hongtao Liu, Yinghui Wang, Hongyan Xu, Pengfei Jiao, Minglai Shao, and Wenjun Wang. Towards a multi-view attentive matching for personalized expert finding. In *Proceedings of the ACM Web Conference 2022*, pages 2131–2140, 2022.

[Qi *et al.*, 2021a] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. Pp-rec: News recommendation with personalized user interest and time-aware news popularity. *arXiv preprint arXiv:2106.01300*, 2021.

[Qi *et al.*, 2021b] Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. HieRec: Hierarchical user interest modeling for personalized news recommendation. In *ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 5446–5456, 2021.

[Son *et al.*, 2013] Jeong-Woo Son, A-Yeong Kim, and Seong-Bae Park. A location-based news article recommendation with explicit localized semantic analysis. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 293–302, 2013.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the International Conference of Neural Information Processing Systems*, pages 5998–6008, Long Beach, CA, USA, 2017. OpenReview.net.

[Wang *et al.*, 2020] Heyuan Wang, Fangzhao Wu, Zheng Liu, and Xing Xie. Fine-grained interest matching for neural news recommendation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 836–845, Online, July 2020. Association for Computational Linguistics.

[Wang *et al.*, 2022] Wenjie Wang, Fuli Feng, Liqiang Nie, and Tat-Seng Chua. User-controllable recommendation against filter bubbles. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1251–1261, 2022.

[Wei *et al.*, 2021] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, page 1791–1800. ACM, August 2021.

[Wu *et al.*, 2019a] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. Neural news recommendation with attentive multi-view learning. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2019-Augus, pages 3863–3869, 2019.

[Wu *et al.*, 2019b] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. Neural news recommendation with multi-head self-attention. In *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 6389–6394, 2019.

[Wu *et al.*, 2020] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. MIND: A large-scale dataset for news recommendation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606, Online, July 2020. Association for Computational Linguistics.

[Wu *et al.*, 2021] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. Empowering news recommendation with pre-trained language models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1652–1656, 2021.

[Wu *et al.*, 2023] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. Personalized news recommendation: Methods and challenges. *ACM Transactions on Information Systems*, 41(1):1–50, 2023.

[Xu *et al.*, 2023] Hongyan Xu, Qiyao Peng, Hongtao Liu, Yueheng Sun, and Wenjun Wang. Group-based personalized news recommendation with long-and short-term fine-grained matching. *ACM Transactions on Information Systems*, 2023.

[Zhang *et al.*, 2021a] Qi Zhang, Jingjie Li, Qinglin Jia, Chuyuan Wang, Jieming Zhu, Zhaowei Wang, and Xiuqiang He. UNBERT: User-News Matching BERT for News Recommendation. In *IJCAI International Joint Conference on Artificial Intelligence*, pages 3356–3362, 2021.

[Zhang *et al.*, 2021b] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 11–20, 2021.

[Zhang *et al.*, 2021c] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal Intervention for Leveraging Popularity Bias in Recommendation. In *SIGIR 2021 - Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, volume 1, pages 11–20. Association for Computing Machinery, 2021.

[Zheng *et al.*, 2021] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021*, pages 2980–2991, 2021.