News, With a Twist: Using Contrastive Learning to Improve User Embeddings for Diverse News Recommendations

Zijie Tang[1,†], Manel Slokom[2,‡]

[1]Vrije Universiteit Amsterdam
[2]Centrum Wiskunde en Informatica, Amsterdam

Abstract

News recommender systems (NRS) play a key role in delivering personalised content in fast-paced, high-volume environments. However, models optimised solely for accuracy often overlook important societal objectives such as fairness and diversity, leading to over-personalisation, biased exposure, and narrow content consumption. In this paper, we propose a contrastive learning framework for improving user representations in neural news recommendation.[1] We build upon a bi-encoder architecture and introduce self-supervised objectives that group semantically related news items by theme, encouraging the model to bring similar items closer in the embedding space while pushing dissimilar ones apart. This strategy mitigates embedding collapse and guides the model toward producing recommendations with broader topical coverage.

We evaluate our approach on the MIND dataset, comparing against state-of-the-art neural models, including LSTUR and NAML. Our results show that the proposed method achieves competitive accuracy and yields measurable improvements in beyond-accuracy objectives, particularly in content diversity and exposure fairness. Our results demonstrate the potential of contrastive learning to support more balanced and responsible news recommendations.

Keywords

News recommendation, Contrastive learning, Multi-objective optimisation, Diversity

1. Introduction

Recommender systems (RSs) are designed to provide personalised suggestions for items that users are most likely to find relevant or appealing. By analysing users' preferences and behaviours, these systems extract valuable insights that help deliver suitable products and services [1]. As a result, RSs play an important role in modern business strategies, enabling data-driven decisions by analysing users' historical choices and behavioural patterns [2].

Despite impressive performance in accuracy-based metrics, many NRSs suffer from over-personalisation and embedding bias, leading to narrow content exposure and poor representation of user interests [3]. Prior work has identified that such models often converge to degenerate embedding geometries, where user representations cluster toward dominant topics or highly clicked content [4]. This behaviour not only limits diversity but also hinders fairness by disproportionately amplifying popular content while marginalising niche or minority interests [5].

To address these challenges, researchers have increasingly turned to "beyond-accuracy" objectives, including diversity, fairness, as complementary goals in recommendation [6, 7]. Diversity improves the user experience by reducing content redundancy, while fairness ensures equitable exposure for both users and content providers [8]. Generally, these goals are often intertwined: models that promote diverse content also tend to improve fairness in exposure [9].

In this paper, we focus on news recommender systems (NRS). We focus on improving user representations in NRSs through the lens of contrastive learning (CL). Contrastive learning (CL) is a self-supervised

[1]Our code is publicly available at: https://github.com/tan9zj/xnrs-CL/tree/main

The 13th International Workshop on News Recommendation and Analytics (INRA 2025)

These authors contributed equally.

z.tang3@student.vu.nl (Z. Tang); manel.slokom@cwi.nl (M. Slokom)
0000-0002-9048-1906 (M. Slokom)

learning approach that pulls semantically similar instances closer in the embedding space and pushes dissimilar ones apart [10]. CL has recently gained attention in recommendation tasks, especially for learning robust representations under sparse or implicit feedback [11]. In the news domain, CL-based models such as SFCNR [12] and SimGCL [13] have demonstrated effectiveness in improving generalisation and capturing subtle user interests by introducing content-aware perturbations or graph-based augmentations. We build upon the neural model XNRS [4], which uses a bi-encoder architecture for user-item matching and identifies embedding collapse as a critical limitation. To address this, we introduce a contrastive learning approach that groups news articles based on their semantic *themes* and uses these groupings to construct positive and negative pairs. By encouraging the user encoder to discriminate between thematic clusters, our approach aims to produce user embeddings that are both more expressive and better aligned with diverse topical interests.

This consideration leads us to the research focus of our study. Our main research question is:

- **Main RQ**: How can we optimise user embeddings to improve recommendations accuracy and diversity? We have the following sub-research questions:
- **RQ0** Can we reproduce research [4]?
- **RQ1** How can contrastive learning be applied to improve the quality of user embeddings?
- **RQ2** What is the impact of improved user embeddings on recommendation accuracy?
- **RQ3** How does our approach perform on beyond-accuracy measures such as diversity?

We summarise our contributions as follows:

- We reproduce the neural news recommendation model proposed in [4], verifying its effectiveness and identifying its limitations regarding embedding bias.
- We introduce contrastive learning mechanisms into the user encoder to better learn the user's interest.
- We evaluate our model on both standard accuracy metrics (e.g., NDCG, MRR, AUC) and beyond-accuracy metrics (e.g., KL divergence, JS divergence, fair-nDCG), demonstrating improvements in recommendation quality, diversity, and fairness.

2. Background and Related Work

In this section, we review existing research on news recommendation systems and methods involving contrastive learning.

2.1. News Recommendation Methods

News recommender systems aim to provide users with personalised news content based on their preferences and browsing behaviour. A typical workflow involves three key stages: candidate news retrieval, personalised ranking, and feedback-based profile updating [14]. When a user visits a platform, a small set of candidate articles is recalled and then ranked based on inferred interests from historical interactions. Top-ranked articles are shown to the user, and their click behaviour is used to update profiles [15]. However, user interests are diverse, context-dependent, and dynamic, making accurate modelling challenging [16].

Classical recommender systems are typically divided into content-based, collaborative filtering, and hybrid methods [17]. Content-based systems analyse item features to recommend similar content [17], while collaborative filtering uses the preferences of similar users [18]. Recent work uses deep learning to improve NRS performance. Models like NRMS [19], NAML [20], LSTUR [21], and NPA [22] adopt different neural architectures (e.g., CNNs [23], LSTMs [24], attention mechanisms [25]) to learn high-quality representations of news content and user behaviour. These methods improve performance by encoding contextual semantics and user interest dynamics. More recent approaches further explore graph-based models [26] or user-news co-embedding strategies [4] to better capture structural and semantic relationships in news consumption.

2.2. Contrastive Learning for News Recommendation

Contrastive learning (CL) improves representation by contrasting positive and negative samples from different data views [27]. In the domain of news recommendation, contrastive learning is used to improve the robustness and expressiveness of user and item embeddings, especially under sparse or implicit feedback [28, 29, 30]. For instance, SFCNR [12] applies contrastive learning to cold-start users by aligning news representations under content-based perturbations. [12] shows superior performance and better alignment of virtual and original features. SimGCL [13] uses directed noise to enforce uniformity, outperforming graph-based augmentation models while reducing training time. However, existing methods have limitations [28, 31]. Many rely on simplistic augmentation strategies, such as random masking and item reordering, that may not capture meaningful variations in user behaviour. In addition, different augmentations should be required for different datasets, e.g., properties and task types [32]. They often overlook the temporal nature of user interests in the news domain, where freshness and recency are critical [26]. Some methods also focus heavily on user-side contrast but ignore item-side diversity or semantic drift, leading to suboptimal generalisation.

2.3. Fairness and Diversity in News Recommendations

In the top N recommendations, fairness aims to ensure equitable treatment between users or groupings of elements [33]. Fairness-aware techniques are classified into pre-, in-, or post-processing. Pre-processing addresses bias in data. For instance, data reweighting or augmentation can balance underrepresented groups [34, 35, 36, 37]. Such strategies are beneficial when the model does not directly use sensitive attributes but is correlated with other features in the training data. In-processing methods add fairness constraints into the learning objective or model architecture, often using regularisation [38, 39, 40]. For example, l_1 and l_2 norm regularisation [38] or graph-based relational modelling [39] help control bias. These techniques enable the model to dynamically balance accuracy and fairness but may increase training complexity. Post-processing modifies ranked outputs to satisfy fairness constraints without changing the model, often through re-ranking [41, 42, 43]. FA*IR [41] guarantees minimum group representation, while LLM-based methods such as IFairLRS [43] adjust for semantic bias post-hoc. While post-processing offers practical deployment advantages, it typically requires access to group membership or sensitive attributes to evaluate and enforce fairness constraints.

With respect to *diversity*, the goal is to provide a varied selection of recommendations that reduce redundancy, enrich user experience, and prevent users from being confined to a narrow range of content, often referred to as the *filter bubble* [44]. In recommender systems, diversity generally refers to the recommended items differing from one another, either in terms of content, category, or user intent [45, 46]. Promoting diversity in recommendations is important in news domains, where repeated exposure to similar viewpoints may lead to confirmation bias and reduced information plurality [47, 48]. Diverse recommendations can help users discover novel content, improve user satisfaction, and even enhance long-term engagement [49]. Several strategies have been proposed to improve diversity in recommendation outputs. These include pre-processing methods that modify user profiles by adding or removing items [37]; re-ranking approaches based on pairwise item dissimilarity [45]; topic modelling techniques to promote topical variety [49]; and multi-objective optimisation approaches that balance accuracy and diversity [50]. Some approaches explicitly include diversity-aware regularisation terms in the model's objective function, while others operate in a post-processing stage, adjusting the output list after ranking. Despite its benefits, improving diversity often comes at the cost of recommendation accuracy, leading to a trade-off that must be carefully managed depending on the application context [51].

3. Model Design

This section describes our proposed news recommendation approach, which integrates a content-based bi-encoder architecture with a contrastive learning objective to optimize user representation quality. Then, we describe the training procedure, including the joint loss formulation and optimisation strategy.

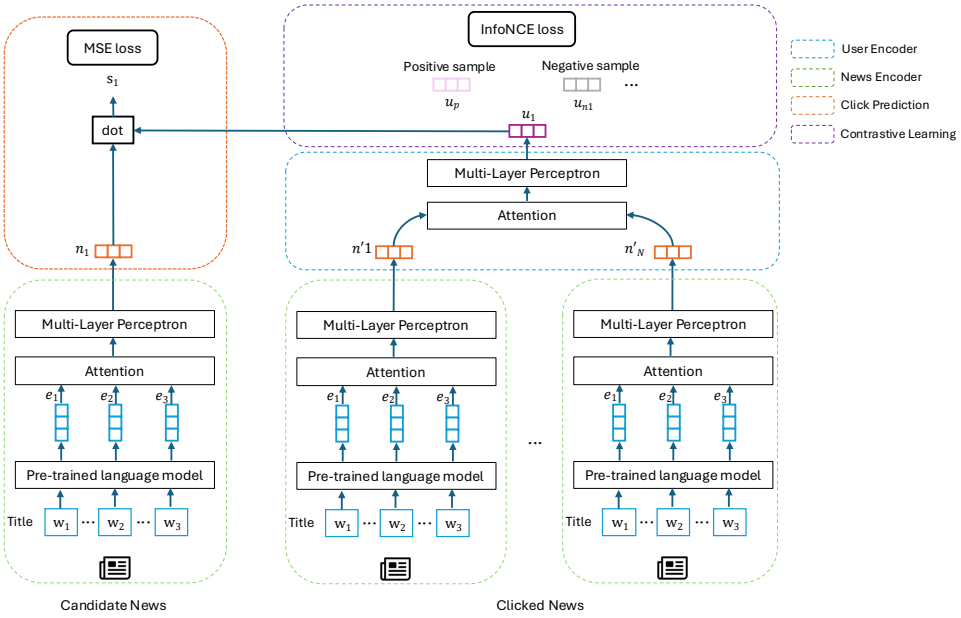

Figure 1: Overview of the Proposed News Recommendation Model.

3.1. Recommendation Model

The overall structure of our model is shown in Figure 1, which consists of a news encoder, a user encoder, and a scoring module. We build upon the work of [4, 52, 20], which proposes a content-based news recommender using a bi-encoder architecture. The model computes the relevance score between a user $\boldsymbol{u}$ and a candidate news article $\boldsymbol{c}$ by taking the inner product of their vector representations:

$$s = \boldsymbol{u}^\top \boldsymbol{c}. \tag{1}$$

News modeling We use the **Siamese sentence-transformer (S-BERT)**, which shows the best performance among RoBERTa, NewsBERT, and FastText [4]. Considering each news title as a sequence of D-dimensional token embeddings $\boldsymbol{e_i}$, with i indexing individual tokens, the sequential representation is represented as:

$$\boldsymbol{n} = MLP\Big(\sum_i \alpha_i \boldsymbol{e_i}\Big), \tag{2}$$

$$\alpha_i = softmax(\boldsymbol{q}^T tanh(W\boldsymbol{e_j} + \boldsymbol{b}))_i, \tag{3}$$

where α_i are the attention weights for the individual token embeddings $\boldsymbol{e_i}$. They are predicted from the respective embeddings as shown in Equation 3, in which W, $\boldsymbol{q}$, and $\boldsymbol{b}$ are learnable parameters. The softmax function ensures that all weights sum to one. After the attention mechanism, the model further transforms the embeddings through a small multi-layer perception (MLP).

User modeling For each user, we use the same additive attention mechanism over the user's history news embeddings $\boldsymbol{n_t}$ for the t most recently read news:

$$\boldsymbol{u} = MLP\Big(\sum_t \lambda_t \boldsymbol{n_t}\Big). \tag{4}$$

Here λ_t are the attention weights gained similarly to α_i through Equation 3.

Contrastive learning In our approach, we extend the standard news recommendation pipeline by integrating contrastive learning to improve user representation learning. While the candidate and historical news are encoded as usual via the news encoder and aggregated into a user embedding using the user encoder, we add an auxiliary contrastive loss on the user embeddings. This loss encourages user embeddings with similar interests (e.g., sharing the same category or theme) to be closer in the embedding space, while pushing dissimilar ones apart.

Given a batch of user embeddings $\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_B$ and their corresponding class labels $y_1, y_2, ..., y_B$, the similarity matrix is:

$$S_{ij} = \frac{\boldsymbol{u}_i^T \boldsymbol{u}_j}{\tau} \tag{5}$$

where τ is a temperature scaling hyperparameter, and the embeddings are ℓ_2 normalized before computing dot-product similarity. For each anchor user i, the contrastive loss is defined as:

$$L_{CL}^{(i)} = -log\frac{\sum_{j \in P(i)} exp(S_{ij})}{\sum_{k \neq i} exp(S_{ij})}, \tag{6}$$

where $P(i)$ denotes the set of indices in the batch that share the same label as i (excluding i itself). The final loss is averaged over all users in the batch:

$$L_{CL} = \frac{1}{|L|} \sum_i L_{CL}^{(i)}. \tag{7}$$

3.2. Model training

The proposed model is trained with a joint loss function that combines the standard click prediction objective and the contrastive learning objective described earlier.

Click prediction loss The primary training objective is to predict whether a user will click on a candidate news article. Given a user embedding $\boldsymbol{u}$ and a candidate news embedding $\boldsymbol{c}$, the predicted relevance score is computed as in Equation 1 and the click probability is computed by applying a sigmoid function to their dot product score:

$$\hat{y} = \sigma(\boldsymbol{u}^\top \boldsymbol{c}), \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function. Our approach employs mean squared error (MSE) loss to model user-news interaction scores as continuous relevance values.

Let $y \in [0, 1]$ be the target label representing the interaction (e.g., 1 for clicked, 0 for not clicked), the click prediction loss is defined as:

$$L_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2, \tag{9}$$

where N is the number of training samples in the batch.

Combined training objective To leverage both supervised click signals and self-supervised contrastive signals, we combine the click prediction loss with the contrastive learning loss:

$$L = L_{MSE} + \lambda \cdot L_{CL}, \tag{10}$$

where λ is a hyperparameter that controls the relative weight of the contrastive loss. In practice, λ is chosen based on validation performance to balance accuracy and representation robustness.

4. Experimental Setup

This section presents the experimental design used to evaluate the effectiveness of our proposed approach. We describe the datasets employed, detail the experimental settings, and report key metrics to analyse model performance. For reproducibility, our code is publicly available at: https://github.com/tan9zj/xnrs-CL/tree/main.

4.1. Datasets

We run experiments on a real-world news recommendation dataset **Microsoft News Dataset (MIND)** [15]. MIND is a large-scale English news recommendation dataset constructed from anonymised user behaviour logs on the Microsoft News platform. It contains user click histories, impression logs, and detailed information about the news articles, including titles, abstracts, categories, and entities of the title and abstract. For our experiments, we use the **MIND-small** subset, which is a randomly sampled portion of the full dataset. The detailed statistics are summarised in Table 1.

Table 1
Statistics of MIND-small Dataset

Statistic	Train	Test
# News	51,282	42,416
# Users	49,108	48,593
# Impressions	153,727	70,938
# Categories	17	17
# Subcategories	264	252

To better align with the objectives of our method, we reorganise the original category labels into a set of high-level thematic groups specifically designed for our contrastive learning approach. The mapping between themes and their corresponding categories is presented in Table 2. This reorganisation serves more than one purpose: first, it facilitates the learning of more generalised user representations by grouping semantically similar categories; second, it encourages the model to consider a broader range of content types during training; third, we note that contrastive learning setups on category implicitly force each user embedding to align strongly with a single content category, which may reinforce narrow interest profiles and cause the filter bubble effect. We aim to promote **diversity** by encouraging the inclusion of multiple distinct categories in the final ranked list. By grouping categories into broader themes, we aim to guide the contrastive learning process to be sensitive to semantic-level distinctions. This allows user embeddings to be aligned with higher-level semantic themes rather than overly specific labels, increasing the likelihood that users are exposed to a diverse set of categories. As a result, the model generates recommendations that are both relevant and semantically varied, contributing to improved fairness and diversity.

Table 2
Themes and the corresponding categories in MIND-small dataset

Theme	Categories
News	news
Lifestyle	weather, foodanddrink, health, lifestyle, travel
Entertainment	videos, entertainment, kids, music, tv, movies, autos
World	northamerica, middleeast
Finance	finance
Sports	sports

4.2. Baselines

We follow the work of Möller et al. [4]. In previous work, they compared the proposed model, XNRS, with several established baselines, including LSTUR [21], NPA [22], NAML [20], NRMS [19], CAUM [53], and a late fusion (LF) approach based on NRMS introduced by Iana et al. [52]. XNRS outperformed most baselines overall, although LSTUR and NAML yielded better results on specific metrics. Based on these findings, we select *XNRS*, *LSTUR*, and *NAML* as baseline models for our experiments, as they have previously demonstrated competitive performance.

4.3. Hyperparameter tuning of CL

There are two key hyperparameters in our contrastive learning setup: the temperature scaling factor (τ) and the contrastive loss weight (λ). To assess the sensitivity of model performance to these parameters, we perform a grid search over a predefined range.

For the temperature τ, we evaluate values in {0.08, 0.1, 0.9}, and for the contrastive loss weight λ, we explore values in {0.005, 0.01, 0.012, 0.02}. Based on the experimental results, we find that setting $\tau = 0.08$ and $\lambda = 0.01$ yields the best overall performance, and we use these values for all subsequent experiments.

4.4. Metrics

Click-Prediction Evaluation For accuracy metrics, we use the well-established ranking metrics normalised discounted cumulative gain (nDCG), mean reciprocal rank (MRR), click-through rate (CTR), and area under the receiver operating characteristic curve (AUC). These metrics evaluate the quality of ranked recommendations from multiple perspectives: nDCG assesses ranking quality by considering both relevance and position of recommended items [54]. The highly relevant items are more useful when appearing earlier in the ranking than the less relevant items. MRR evaluates the ranking quality based on the position of the first relevant item in the recommendation list. A higher MRR indicates that relevant items appear earlier in the ranking list. CTR measures user engagement by calculating the ratio of clicked recommendations to the total number of displayed recommendations. AUC measures the probability that a randomly chosen relevant item is ranked higher than a randomly chosen irrelevant item. Higher value indicates the effectiveness of a recommender system in distinguishing relevant items from irrelevant ones.

Diversity Evaluation To evaluate the diversity of the recommended results, we adopt distribution-based metrics that compare the statistical difference between the category distributions of the recommended items and a reference distribution, typically the user's historical reading distribution. We use Kullback–Leibler (KL) divergence and Jensen–Shannon (JS) divergence to quantify this dissimilarity [55].

Fairness Evaluation To evaluate the relevance and category-level fairness of recommendations, particularly toward minority categories, we adopt a modified version of the normalised Discounted Cumulative Gain (nDCG) called fair-nDCG [56].

5. Experimental Results

In this section, we analyse our model performance across multiple aspects, including recommendation accuracy, representation geometry, fairness and diversity of recommendations, and robustness.

5.1. Recommendation performance

We start by evaluating the click-prediction performance of our proposed model against several strong neural baselines. Our method, `XNRS+CL`, extends the original XNRS model [4] by incorporating contrastive learning. The goal is to improve user representations by encouraging semantically similar news items to be embedded closer in the latent space. Table 3 shows our experiment results.

Overall, our results in Table 3, show that both contrastive learning variants of XNRS (`XNRS+CL(theme)` and `XNRS+CL(category)`) consistently outperform the original XNRS model across all metrics, nDCG, MRR, CTR, and AUC. This could be explained by the fact that contrastive learning contributes to better user-item matching, likely by improving the quality of user embeddings.

Notably, `XNRS+CL(category)` shows slightly stronger performance than the theme-based variant on several metrics, including nDCG@5, MRR, and CTR@1. Although the differences are modest, this suggests that the type of semantic grouping used to define positive contrastive pairs may affect specific

Table 3
Click Prediction Performance of Our Model and Baselines (Bold = Best, Underlined = Second Best).

	XNRS+CL(theme)	XNRS+CL(category)	XNRS	LSTUR	NAML
nDCG@5	0.3624 ± 0.0014	0.3625 ± 0.0013	0.3591 ± 0.0012	0.3502 ± 0.0049	**0.3650 ± 0.0052**
nDCG@10	0.4244 ± 0.0009	0.4246 ± 0.0008	0.4215 ± 0.0011	0.4139 ± 0.0041	**0.4272 ± 0.0042**
MRR	0.3767 ± 0.0010	0.3787 ± 0.0006	0.3750 ± 0.0003	0.3684 ± 0.0031	**0.3797 ± 0.0029**
CTR@1	0.2032 ± 0.0020	**0.2074 ± 0.0023**	0.2034 ± 0.0014	0.1994 ± 0.0029	0.2038 ± 0.0003
CTR@10	0.1267 ± 0.0002	0.1265 ± 0.0003	0.1261 ± 0.0004	0.1247 ± 0.0009	**0.1270 ± 0.0007**
AUC	0.6805 ± 0.0028	0.6800 ± 0.0039	0.6783 ± 0.0040	0.6696 ± 0.0051	**0.6822 ± 0.0057**

aspects of model behaviour. Further exploration of grouping strategies could provide additional insight into the relationship between content semantics and recommendation effectiveness.

5.2. Geometry of the embeddings

To gain insight into how contrastive learning shapes the user representation space, we visualise the user embeddings before and after applying our theme-based contrastive learning objective. Specifically, we project the high-dimensional user embeddings onto a polar coordinate plot to enable qualitative comparison, following the approach of Möller et al. [4].

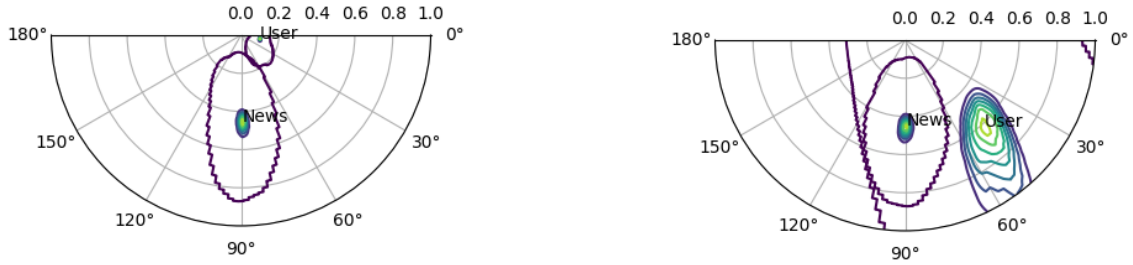

Figure 2: Kernel density plots of the distributions of user embeddings and news embeddings (`Left`) before and (`Right`) after applying theme-based contrastive learning. Angles are relative to the mean user.

As illustrated in Figure 2, the contrastive learning objective encourages the user embeddings to expand and become more geometrically structured. In the original embedding space (Figure 5.2), user embeddings are relatively concentrated and show limited separation. After contrastive learning (Figure 5.2), the embeddings are more widely dispersed. This geometric expansion suggests that contrastive learning promotes better separation among users. Figure 3 provides an additional t-SNE visualisation of user embeddings. After applying contrastive learning, the embeddings show clearer clustering structures and are more directionally aligned.

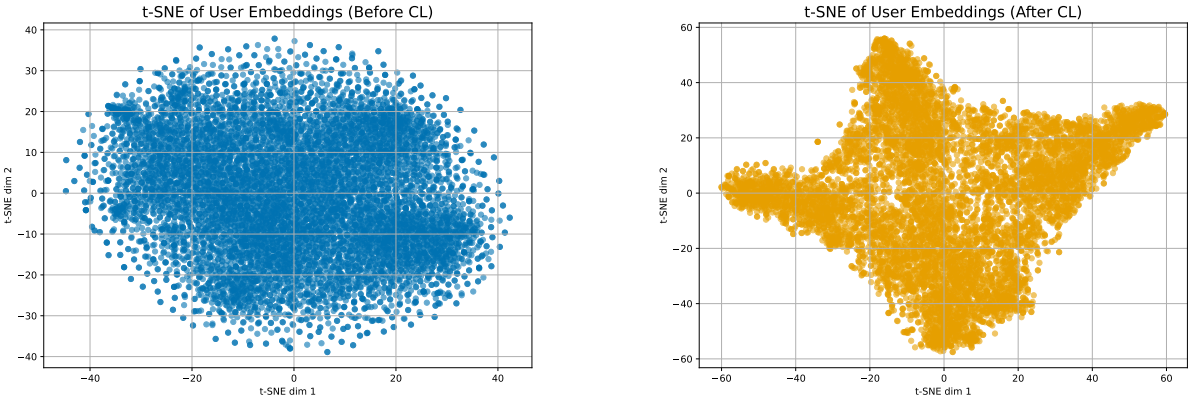

Figure 3: t-SNE plots of user embeddings (`Left`) before and (`Right`) after applying them-based contrastive learning.

5.3. Fairness and Diversity of the Recommendations

In addition to accuracy, we evaluate our model on fairness and diversity criteria to assess its ability to provide balanced exposure across categories, especially minority ones.

Fairness To evaluate the fairness of the recommendation lists, we adopt the fair-nDCG metric, a modification of the traditional nDCG that assigns relevance only to items belonging to predefined underrepresented or minority categories. In our experiments, we define the minority categories as: `tv`, `entertainment`, `music`, `kids`, `movies`, `middleeast`, `games`, `weather`, and `autos`.

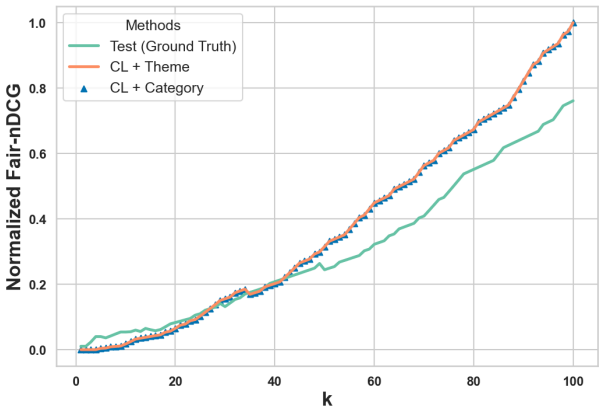

Figure 4: Fair-nDCG@k scores for different models. Higher scores indicate more equitable representation of minority categories. Recommendations range from $K = \{1, ..100\}$

As shown in Figure 4, we compare the fairness performance across top-k recommendations between different model variants. The green line represents the ground-truth distribution from the test set, while the orange and blue lines correspond to models trained with contrastive learning (CL) based on theme-level and category-level grouping, respectively.

We observe that both contrastive variants outperform the baseline in terms of minority category coverage, especially as the value of k increases. The curves for `CL+Theme` and `CL+Category` almost entirely overlap, indicating that both variants contribute similarly to improved fairness, despite using different semantic groupings during training. This trend is also reflected in Table 3, where the NDCG scores of these two variants are nearly identical. This consistency may suggest that the introduction of contrastive signals, rather than the specific grouping schema, is the primary factor driving improved fairness in recommendation ranking.

Diversity We quantify the diversity of the recommendations using statistical divergence measures. To evaluate the topical diversity of the recommendation lists, we compute the average Kullback–Leibler (KL) divergence and Jensen–Shannon (JS) divergence between the category distributions of the recommended items and a reference distribution, either derived from the user's historical reading behaviour or the global news category distribution. Here, we consider the user's reading history distribution as the reference one. We observe that both contrastive learning variants, whether based on category or theme, produce remarkably similar diversity outcomes. Specifically, the average KL divergence across all users is 2.2192, and the mean JS divergence is 0.4941 for both variants. Figure 5 illustrates the distribution of divergence scores alongside the user count. These findings suggest that, despite differences in how user representations are constructed, the overall category diversity of the generated recommendation lists remains comparable.

However, when we restrict the evaluation to the top-5 or top-10 recommended items, the theme-based variant shows better diversity. Table 4 presents a comparison between `XNRS+CL(theme)` and `XNRS+CL(category)` under Top-5 and Top-10 cutoffs.

These results suggest that our theme-based methods lead to a more diverse set of recommended news articles, particularly in the top-ranked positions where user attention is highest. Although absolute

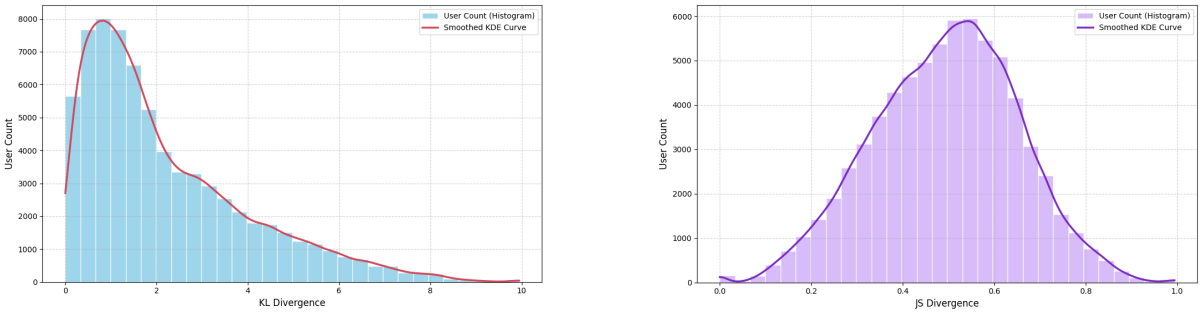

Figure 5: Diversity Evaluation Based on CL with Theme Modelling. `(Left)` KL Divergence; `(Right)` JS Divergence

Table 4
Diversity Evaluation Based on KL and JS Divergence Across Model Variants.

Model	Top-5		Top-10	
	Mean KL	Mean JS	Mean KL	Mean JS
XNRS+CL(theme)	**5.7758**	**0.6901**	**4.3402**	**0.6152**
XNRS+CL(category)	5.7620	0.6883	4.3217	0.6130

differences are relatively small, the consistent advantage of the theme-based variant at the top position indicates that higher semantic granularity in contrastive learning, such as grouping by theme rather than a single category, can encourage more varied and semantically rich recommendation lists.

5.4. Ablation study

To further understand the impact of different model components, we conduct an ablation study comparing multiple architectural and training variations, including different user representation strategies (e.g., scoring function) and contrastive grouping granularity (e.g., themes vs. categories). We name these variants following the setup from previous work [4].

First, we remove the MLP inside the user encoder (Equation 4), obtaining the *base* model. Then, we replace the attention mechanisms in both the news and user encoders (Equations 2 and 4) with a simple average pooling over the input embeddings, naming this variant *Mean*. For the final matching layer between the user representation $\boldsymbol{u}$ and candidate news items $\boldsymbol{c}_i$, we experiment with two types of scoring functions:

- Dot Product (*dot*): This baseline computes the interaction score as the dot product between the user and item embeddings:

$$s_i = \boldsymbol{u}^\top \boldsymbol{c}_i \tag{11}$$

  Optionally, the embeddings can be ℓ_2-normalised before computing the similarity. This is efficient and widely used in many recommendation models.
- Bilinear (*bilin*): Inspired by prior work [57, 4] on learning user-item interactions, we introduce a trainable bilinear transformation via a learnable matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$:

$$s_i = \boldsymbol{u}^\top \mathbf{W} \boldsymbol{c}_i \tag{12}$$

Our results are summarised in Table 5. As shown, we observe that `base+dot+theme` consistently outperforms `base+bilin+theme`. This indicates that the dot-product scoring function is more effective than the bilinear alternative in our contrastive learning setup. Among the model variants, the *standard* configuration outperforms the *mean* variant across most metrics, consistent with prior findings [4] that identify the standard model as the strongest among the three.

We note that when contrastive learning is applied to the lightweight *base* model, its performance improves substantially, outperforming even NAML on several metrics. In the case of the *base* model, contrastive learning is applied directly to the user embeddings produced by the encoder, without

Table 5
Performance comparison across different model variants.

	standard+dot(XNRS)	standard+dot+theme(Ours)	base+bilin+theme	mean+dot+theme	base+dot+theme
NDCG@5	0.3591 ± 0.0012	0.3624 ± 0.0014	0.3572 ± 0.0071	0.3608 ± 0.0040	**0.3667 ± 0.0043**
NDCG@10	0.4215 ± 0.0011	0.4244 ± 0.0009	0.4201 ± 0.0065	0.4224 ± 0.0038	**0.4275 ± 0.0038**
MRR	0.3750 ± 0.0003	0.3767 ± 0.0010	0.3723 ± 0.0078	0.3750 ± 0.0041	**0.3810 ± 0.0032**
CTR@1	0.2034 ± 0.0014	0.2032 ± 0.0020	0.1990 ± 0.0098	0.2024 ± 0.0035	**0.2095 ± 0.0041**
CTR@10	0.1261 ± 0.0004	**0.1267 ± 0.0002**	0.1261 ± 0.0007	0.1257 ± 0.0006	0.1262 ± 0.0008
AUC	0.6783 ± 0.0040	0.6805 ± 0.0028	0.6786 ± 0.0053	0.6805 ± 0.0024	**0.6830 ± 0.0051**

additional architectural complexity. Surprisingly, this straightforward setup yields strong performance. However, the reasons behind its effectiveness remain unclear and warrant further investigation.

6. Discussion

Our results demonstrate that incorporating contrastive learning into a bi-encoder news recommendation architecture improves user representation quality and overall recommendation accuracy. Both theme-based and category-based grouping strategies outperform the base model, with particularly strong performance from the theme-based variant.

Beyond accuracy, we observe measurable gains in diversity and fairness metrics, especially at deeper ranks (e.g., top-30). However, these improvements are less pronounced at top-5 and top-10, where real-world user attention is typically concentrated. This discrepancy highlights a key challenge in designing recommendation models that balance fairness and diversity with real-world usability.

Another important observation is the strong performance of the lightweight *base* model when combined with contrastive learning. Despite its architectural simplicity, it competes closely with, and sometimes outperforms, more complex models like NAML. This suggests that contrastive learning can be a powerful technique even in low-complexity setups and motivates further exploration into contrastive signals and pairing strategies.

7. Conclusion and Future Work

In this paper, we propose a contrastive learning framework for content-based news recommendation using a bi-encoder architecture.

Outlook Our method builds upon the XNRS model [4] and introduces self-supervised learning objectives that group semantically similar news items by themes or categories. The resulting user embeddings lead to improved performance across standard ranking metrics and beyond-accuracy objectives. Extensive experiments on the MIND dataset show that our contrastive learning variants achieve competitive or superior performance compared to strong neural baselines such as LSTUR and NAML. We found that even lightweight model variants benefit from contrastive learning supervision.

In addition to improvements in accuracy, our method contributes positively to diversity and fairness. While gains are more prominent at deeper ranks, the theme-based contrastive setup shows promise for encouraging a more balanced exposure of content.

Future Work One key limitation of the current model is its reliance solely on news titles for content encoding. Other textual fields, such as subtitles, abstracts, and metadata (e.g., keywords, subcategories), offer rich semantic information that could improve representation learning. Incorporating these modalities may help capture finer-grained topical signals and disambiguate semantically similar content.

Future work should also address fairness in the top-ranked positions, where user engagement is highest. Although our method improves fairness at top-30 positions, the impact is limited at top-5 or top-3 ranks. Fairness-aware re-ranking techniques or adaptive contrastive loss functions that prioritise exposure fairness at higher ranks may help bridge this gap. Another promising direction is the

exploration of alternative contrastive learning strategies. These may include hierarchical or dynamic positive pair generation, temporal-aware contrastive signals, or dual user-item contrastive frameworks. Such extensions could further enhance both the performance and generalisability of contrastive learning in news recommendation systems.

Acknowledgments

This publication is part of the AI, Media & Democracy Lab. For more information about the lab and its further activities, visit https://www.aim4dem.nl/. The authors thank Bo-Chan Jack, Sven Lankester, and Natalie Halaskova for their valuable advice.

Declaration on Generative AI

The authors declare that GenAI is only used to identify and correct grammatical errors, typos, and other writing mistakes. This helps improve the clarity and professionalism of the writing.

References

[1] F. E. Zaizi, S. Qassimi, S. Rakrak, Multi-objective optimization with recommender systems: A systematic review, Information Systems 117 (2023) 102233.

[2] R. Kuo, C.-K. Chen, S.-H. Keng, Application of hybrid metaheuristic with perturbation-based k-nearest neighbors algorithm and densest imputation to collaborative filtering in recommender systems, Information Sciences 575 (2021) 90–115.

[3] E. Pariser, The filter bubble: What the Internet is hiding from you, Penguin UK, 2011.

[4] L. Möller, S. Padó, Explaining neural news recommendation with attributions onto reading histories, ACM Transactions on Intelligent Systems and Technology 16 (2025) 1–25.

[5] Y. Zhao, Y. Wang, Y. Liu, X. Cheng, C. C. Aggarwal, T. Derr, Fairness and diversity in recommender systems: a survey, ACM Transactions on Intelligent Systems and Technology 16 (2025) 1–28.

[6] S. M. McNee, J. Riedl, J. A. Konstan, Being accurate is not enough: how accuracy metrics have hurt recommender systems, in: CHI '06 Extended Abstracts on Human Factors in Computing Systems, CHI EA '06, Association for Computing Machinery, 2006, p. 1097–1101. doi:10.1145/1125451.1125659.

[7] D. Jannach, G. Adomavicius, Recommendations with a purpose, in: Proceedings of the 10th ACM conference on recommender systems, 2016, pp. 7–10.

[8] T. Qi, F. Wu, C. Wu, P. Sun, L. Wu, X. Wang, Y. Huang, X. Xie, Profairrec: Provider fairness-aware news recommendation, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1164–1173.

[9] M. D. Ekstrand, A. Das, R. Burke, F. Diaz, Fairness in Recommender Systems, Springer US, 2022, pp. 679–707.

[10] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International conference on machine learning, PmLR, 2020, pp. 1597–1607.

[11] Y. Wei, X. Wang, Q. Li, L. Nie, Y. Li, X. Li, T.-S. Chua, Contrastive learning for cold-start recommendation, in: Proceedings of the 29th ACM international conference on multimedia, 2021, pp. 5382–5390.

[12] H. Jiang, C. Li, J. Cai, R. Tian, J. Wang, Self-supervised contrastive enhancement with symmetric few-shot learning towers for cold-start news recommendation, in: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, 2023, pp. 945–954.

[13] J. Yu, H. Yin, X. Xia, T. Chen, L. Cui, Q. V. H. Nguyen, Are graph augmentations necessary? simple graph contrastive learning for recommendation, in: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, 2022, pp. 1294–1303.

[14] C. Wu, F. Wu, Y. Huang, X. Xie, Personalized news recommendation: Methods and challenges, ACM Transactions on Information Systems 41 (2023) 1–50.
[15] F. Wu, Y. Qiao, J.-H. Chen, C. Wu, T. Qi, J. Lian, D. Liu, X. Xie, J. Gao, W. Wu, et al., Mind: A large-scale dataset for news recommendation, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 3597–3606.
[16] C. Feng, M. Khan, A. U. Rahman, A. Ahmad, News recommendation systems-accomplishments, challenges & future directions, IEEE Access 8 (2020) 16702–16725.
[17] G. Adomavicius, A. Tuzhilin, 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 (2005) 734–749.
[18] J. B. Schafer, D. Frankowski, J. Herlocker, S. Sen, Collaborative filtering recommender systems, in: The adaptive web: methods and strategies of web personalization, Springer, 2007, pp. 291–324.
[19] C. Wu, F. Wu, S. Ge, T. Qi, Y. Huang, X. Xie, Neural news recommendation with multi-head self-attention, 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), 2019, pp. 6389–6394.
[20] C. Wu, F. Wu, M. An, J. Huang, Y. Huang, X. Xie, Neural news recommendation with attentive multi-view learning, arXiv preprint arXiv:1907.05576 (2019).
[21] M. An, F. Wu, C. Wu, K. Zhang, Z. Liu, X. Xie, Neural news recommendation with long-and short-term user representations, in: Proceedings of the 57th annual meeting of the association for computational linguistics, 2019, pp. 336–345.
[22] C. Wu, F. Wu, M. An, J. Huang, Y. Huang, X. Xie, Npa: neural news recommendation with personalized attention, in: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019, pp. 2576–2584.
[23] R. Yamashita, M. Nishio, R. K. G. Do, K. Togashi, Convolutional neural networks: an overview and application in radiology, Insights into imaging 9 (2018) 611–629.
[24] A. Graves, A. Graves, Long short-term memory, Supervised sequence labelling with recurrent neural networks (2012) 37–45.
[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).
[26] S. Ge, C. Wu, F. Wu, T. Qi, Y. Huang, Graph enhanced representation learning for news recommendation, in: Proceedings of the web conference 2020, 2020, pp. 2863–2869.
[27] J. Giorgi, O. Nitski, B. Wang, G. Bader, DeCLUTR: Deep contrastive learning for unsupervised textual representations, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, 2021, pp. 879–895. doi:10.18653/v1/2021.acl-long.72.
[28] L. Wu, H. Lin, C. Tan, Z. Gao, S. Z. Li, Self-supervised learning on graphs: Contrastive, generative, or predictive, IEEE Transactions on Knowledge and Data Engineering (2021).
[29] A. Iana, G. Glavaš, H. Paulheim, Train once, use flexibly: A modular framework for multi-aspect neural news recommendation, arXiv preprint arXiv:2307.16089 (2023).
[30] X. Ren, W. Wei, L. Xia, C. Huang, A comprehensive survey on self-supervised learning for recommendation, ACM Computing Surveys (2024).
[31] W. He, G. Sun, J. Lu, X. S. Fang, Candidate-aware graph contrastive learning for recommendation, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR, 2023, p. 1670–1679. doi:10.1145/3539618.3591647.
[32] W. Jin, T. Derr, H. Liu, Y. Wang, S. Wang, Z. Liu, J. Tang, Self-supervised learning on graphs: Deep insights and new direction, arXiv preprint arXiv:2006.10141 (2020).
[33] E. Pitoura, K. Stefanidis, G. Koutrika, Fairness in rankings and recommendations: an overview, The VLDB Journal (2022) 1–28.
[34] F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, K. R. Varshney, Optimized pre-processing for discrimination prevention, Advances in neural information processing systems 30

(2017).
[35] L. Chen, L. Wu, K. Zhang, R. Hong, D. Lian, Z. Zhang, J. Zhou, M. Wang, Improving recommendation fairness via data augmentation, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 1012–1020.
[36] K. Balasubramanian, A. Alshabanah, E. Markowitz, G. Ver Steeg, M. Annavaram, Biased user history synthesis for personalized long-tail item recommendation, in: Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 189–199.
[37] M. Slokom, S. Daniil, L. Hollink, How to diversify any personalized recommender?, in: Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part IV, 2025, p. 307–323. URL: https://doi.org/10.1007/978-3-031-88717-8_23. doi:10.1007/978-3-031-88717-8_23.
[38] R. Burke, N. Sonboli, A. Ordonez-Gauger, Balanced neighborhoods for multi-sided fairness in recommendation, in: Conference on fairness, accountability and transparency, PMLR, 2018, pp. 202–214.
[39] B. Yang, D. Liu, T. Suzumura, R. Dong, I. Li, Going beyond local: Global graph-enhanced personalized news recommendations, in: Proceedings of the 17th ACM Conference on Recommender Systems, 2023, pp. 24–34.
[40] L. Boratto, F. Fabbri, G. Fenu, M. Marras, G. Medda, Fair augmentation for graph collaborative filtering, in: Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 158–168.
[41] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, R. Baeza-Yates, Fa* ir: A fair top-k ranking algorithm, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 1569–1578.
[42] S. C. Geyik, S. Ambler, K. Kenthapadi, Fairness-aware ranking in search & recommendation systems with application to linkedin talent search, in: Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining, 2019, pp. 2221–2231.
[43] M. Jiang, K. Bao, J. Zhang, W. Wang, Z. Yang, F. Feng, X. He, Item-side fairness of large language model-based recommendation system, in: Proceedings of the ACM on Web Conference 2024, 2024, pp. 4717–4726.
[44] T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, J. A. Konstan, Exploring the filter bubble: the effect of using recommender systems on content diversity, in: Proceedings of the 23rd ACM international conference on World wide web, 2014, pp. 677–686.
[45] S. Vargas, P. Castells, Rank and relevance in novelty and diversity metrics for recommender systems, in: Proceedings of the 5th ACM conference on Recommender systems, 2011, pp. 109–116.
[46] M. Kunaver, T. Požrl, Diversity in recommender systems–a survey, Knowledge-Based Systems 123 (2017) 154–162.
[47] E. Bozdag, Bias in algorithmic filtering and personalization, Ethics and Information Technology 15 (2013) 209–227.
[48] N. Helberger, K. Karppinen, L. D'Acunto, Exposure diversity as a design principle for recommender systems, Information, Communication & Society 21 (2018) 191–207.
[49] C.-N. Ziegler, S. M. McNee, J. A. Konstan, G. Lausen, Improving recommendation lists through topic diversification, in: Proceedings of the 14th international conference on World Wide Web, 2005, pp. 22–32.
[50] M. Zhang, N. Hurley, J. Peng, Avoiding monotony: improving the diversity of recommendation lists, Proceedings of the 2008 ACM conference on Recommender systems (2008) 123–130.
[51] P. Castells, S. Vargas, J. Wang, Novelty and diversity in recommender systems, Recommender Systems Handbook (2021) 845–884.
[52] A. Iana, G. Glavas, H. Paulheim, Simplifying content-based neural news recommendation: On user modeling and training objectives, in: Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval, 2023, pp. 2384–2388.
[53] T. Qi, F. Wu, C. Wu, Y. Huang, News recommendation with candidate-aware user modeling, in: Proceedings of the 45th international ACM SIGIR conference on research and development in

information retrieval, 2022, pp. 1917–1921.
[54] K. Järvelin, J. Kekäläinen, Ir evaluation methods for retrieving highly relevant documents, in: ACM SIGIR Forum, volume 51, 2017, pp. 243–250.
[55] H. Steck, Calibrated recommendations, in: Proceedings of the 12th ACM Conference on Recommender Systems, RecSys '18, 2018, p. 154–162.
[56] E. Pitoura, K. Stefanidis, G. Koutrika, Fairness in rankings and recommendations: an overview, The VLDB Journal 31 (2021) 431–458.
[57] L. Möller, S. Padó, Understanding the relation of user and news representations in content-based neural news recommendation, The 10th International Workshop on News Recommendation and Analytics (INRA) (2022).