Enhancing Diversity in News Recommendations Increases Click-Through Rates: Insights from an Online Experiment and User Study

Robin Verachtert
DPG Media
Antwerp, Belgium
robin.verachtert@dpgmedia.be

Kim Falk
DPG Media
Antwerp, Belgium
kim.falk@dpgmedia.be

Christine Bauer
Department of Artificial Intelligence and Human Interfaces
University of Salzburg
Salzburg, Austria
christine.bauer@plus.ac.at

Abstract

Diversity is a widely studied beyond-accuracy aspect of recommender systems, particularly in the news domain. Extensive research has explored its theoretical foundations and proposed algorithmic strategies to promote it, with most evaluations conducted through offline experiments. This work presents the results of deploying and evaluating diversification methods in a large-scale production news recommender system. Motivated by the goal of upholding editorial values, we compare three diversification methods: *Interleaving* and two implementations of Intra-List Diversification (ILD), relying on Term Frequency-Inverse Document Frequency (TF-IDF) and Bidirectional Encoder Representations from Transformers (BERT) embeddings, respectively. Across a two-week online experiment (A/B test) and a follow-up user study on a large-scale production news platform, ILD with BERT embeddings improved diversity as measured by a reduction in Intra-List Similarity (ILS) and increased Click-Through Rates (CTRs), while also improving users' perceived relevance.

CCS Concepts

• **Information systems** → **Recommender systems**; • **Human-centered computing**;

Keywords

recommender systems, intra-list similarity, diversification, topic diversity, news, user perceptions, online experiment, user study

ACM Reference Format:
Robin Verachtert, Kim Falk, and Christine Bauer. 2026. Enhancing Diversity in News Recommendations Increases Click-Through Rates: Insights from an Online Experiment and User Study. In *34th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '26), June 08–11, 2026, Gothenburg, Sweden.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3774935.3806153

1 Introduction

News plays a vital role in shaping individuals' understanding of public matters and how they connect to society at large. As a result, news is a cornerstone for forming public spheres, which must remain open to a wide range of topics and viewpoints that are essential for societal self-reflection and the development of shared knowledge about current affairs [3]. In this context, the diversity of news is critical to supporting this process [3]. Consequently, diversity has become a widely studied concept in news recommender systems (NRSs) [2, 16]. While diversity in NRSs has been addressed conceptually [e.g., 4, 42] and analyzed through offline evaluation [e.g., 27, 30], comparatively little research has studied diversity in real-world production settings [2, 13].

We extend this line of research by a production-scale comparison of a topic-based interleaving (Round-Robin by topic; *Interleaving*) and item-level diversification (ILD) using Term Frequency-Inverse Document Frequency (TF-IDF) and Bidirectional Encoder Representations from Transformers (BERT) embeddings. We investigate the following resarch questions (RQs):

- RQ1: How do different diversification strategies affect list-level diversity in a production news recommender system, as measured by Intra-List Similarity (ILS) and topic diversity?
- RQ2: What is the impact of these diversification strategies on user engagement, measured by Click-Through Rate (CTR)?
- RQ3: How do users perceive variety and relevance under these strategies?

To address these RQs, our empirical study combines platform-wide A/B testing (online experiment) with a complementary user study. The online experiment provides large-scale behavioral evidence and is used to answer RQ1 (objective diversity) and RQ2 (engagement effects). The user study, motivated by known discrepancies between objective diversity metrics and user perception, addresses RQ3.

Prior studies suggest that perceived diversity and similarity often diverge from their corresponding objective metrics. This misalignment has been observed in the news domain [33, 34] as well as in adjacent areas such as movies or music [7, 14, 25, 35, 43]. To explore this further, we complement our online experiment with a user study that investigates how users perceive the diversification in terms of both diversity and utility in a real-world setting,.

We contribute an integrated, production-scale evaluation of diversification strategies in a live NRS. We examine objective diversity effects, user engagement outcomes, as well as users' perceived variety and relevance, and derive practical and methodological insights into relevance–diversity trade-offs and the operational realities of deploying diversified news recommendations in real-world settings. Our findings show that diversification improved the perceived utility of the recommendations (in terms of *perceived relevance*) as well as user engagement (in terms of *CTRs*).

UMAP '26, Gothenburg, Sweden

ACM ISBN 979-8-4007-2311-7/26/06
https://doi.org/10.1145/3774935.3806153

The remainder of this paper is structured as follows: Section 2 presents the conceptual basis and related work on diversity in Recommender Systems (RSs), focusing on the news domain. Section 3 outlines our methodological approach to diversification. Section 4 describes the context of the large-scale production news platform where we deployed and evaluated our approaches. The evaluation setup is detailed in Section 5. We present results in Section 6 and discuss its implications in Section 7. Finally, we conclude our work with a summary of the main contributions (Section 8).

2 Conceptual Background and Related Work

Conceptual Overview of Diversity in Recommender Systems. Diversity in RSs has been extensively studied across various domains, including movies [50], e-commerce [8], and—most notably—news [2, 16, 28]. Promoting diversity in recommendations adds value by ensuring exposure to varied content [1, 6]. Research on diversity in RSs has been approached from multiple perspectives. One line of work focuses on increasing exposure to underrepresented content, such as including articles about small political parties to balance exposure across the political spectrum [11], which may help depolarize democratic societies [12], or balancing editorial content in predefined proportions [1]. Other works emphasize balancing content across topics [52], gender [39], or user intent [44], rather than prioritizing specific content.

Diversity in News Recommender Systems. While diversity has been explored across various domains, the news domain presents unique challenges and opportunities. In NRSs, diversity ensures that users are exposed to a wide range of topics and viewpoints, which is crucial for fostering an informed and reflective public sphere [3, 6]. It has also been explored from humanities and media studies perspectives [10, 12, 22]. For example, Heitz et al. [12] examined how diversifying articles based on political views affects users' political perspectives. A recent review on value-aware NRSs [2] identifies diversity as the most frequently addressed value.

Conceptualizing and Measuring Diversity. Despite its importance, diversity in RSs—both within and beyond the news domain—remains challenging to conceptualize and measure. Indeed, despite numerous calls for more diverse recommendations, these calls often lack clarity on how diversity should manifest [2]. Vrijenhoek et al. [41] highlight that diversity is implemented in various ways—partly due to conceptual ambiguity, as the term often refers to different underlying notions. They argue for a case-by-case conceptualization of diversity rather than pursuing a standardized definition. Others (e.g., [36, 37]) propose that diversity should only be promoted within the boundaries of the accuracy–diversity–fairness (ADF) framework, aiming for a good balance between these three dimensions. A commonly used metric for measuring diversity is ILS, which is calculated as the average pairwise similarity among items in a recommendation list [52]—where diversity is understood as the inverse of similarity. With this, ILS provides an indication of how similar the items in a list are. While other metrics have been proposed [24, 32, 40, 49] (for an overview of diversity metrics, see Kunaver and Požrl [20]), we did not implement them due to constraints related to the production system requirements and implementation efforts (see Section 4).

Accuracy–Diversity Trade-off. The trade-off between accuracy and diversity is widely discussed in the literature [e.g., 15, 23, 26, 42]. Typically, higher accuracy is accompanied by a lower diversity and vice versa—a phenomenon referred to as the "accuracy-diversity dilemma" [51]. This trade-off underscores the challenge of designing RSs that balance relevance with diversity, as it directly affects user satisfaction and system performance.

User Perceptions of Diversity. Although computational metrics like ILS provide a basis for measuring diversity, they often fail to capture how users perceive diversity and the trade-off. Understanding the gap between measured and perceived diversity is critical for designing effective RSs. Diversity metrics often assume a shared understanding of (dis)similarity. However, measuring diversity remains a complex challenge, as research has shown that people's perceptions and understanding of diversity often differ from computational diversity metrics commonly employed in the literature [8, 14, 34, 35]. For instance, Jesse et al. [14] find that while the ILS measure can serve as an effective proxy for user-perceived diversity, its suitability depends heavily on the item space representation used in its computation and the application domain. To address these concerns, we also investigate users' perceptions of the diversity and utility of the recommendations (see Section 5.2).

Diversification Methods. Many diversification methods have been proposed over the years [31, 37, 45]. For comprehensive overviews, see Kaminskas and Bridge [15] and Kunaver and Požrl [20]. In our work, we focus on comparing *Interleaving* and variations of *Intra-List Diversification (ILD). Interleaving*, equivalent to the 'Round-Robin' diversification method proposed by Silva et al. [31], serves as our baseline. Ziegler et al. [52] proposed a method to minimize ILS for recommendations, referred to as *Topic Diversification*, which uses pairwise (topic) similarity between items. In our work, we deployed ILD with two types of item representations: TF-IDF and BERT embeddings. Prior research suggests that TF-IDF often aligns with user judgments [34], while context-based embeddings like BERT have shown superior performance in other studies [21]. A foundational algorithm in diversity-oriented re-ranking is Maximal Marginal Relevance (MMR) [5], originally introduced for document retrieval but widely adopted in RSs. MMR iteratively selects items that maximize a weighted combination of predicted relevance and novelty relative to the already selected items. Variants of MMR have been used with both categorical topic labels and, more recently, continuous item representations derived from TF-IDF or neural embeddings [21, 29].

Evaluating Diversification Methods. To assess the effectiveness and practical utility of diversification methods, it is essential to rigorously evaluate their impact. For the news domain, the aforementioned review on value-aware NRSs [2] highlights a significant gap in the evaluation of diversification methods, as many studies focus primarily on accuracy-based metrics, while often falling short in evaluating the proposed diversification methods in terms of diversity-related metrics [2]—a trend consistent with earlier reviews on NRSs (e.g., [28]).

While several works evaluate recommendation diversity [8, 9, 14, 46], fewer studies do so in production NRS and link behavioral

indicators to user perceptions under real-world constraints. Our work contributes evidence from such a deployment.

To address this gap, we evaluate the diversity of each method and link these results to online performance metrics and user perceptions. Our evaluation not only addresses critical gaps in the literature but also provides actionable insights into the practical impact of diversification methods in real-world settings.

3 Methods

As we employ re-ranking diversification methods in a production NRS (Section 4), they must be computationally efficient to meet strict time constraints. This affects the choice of methods and implementation details. Specifically, we implemented *Interleaving*, a simple baseline that optimizes topic diversity directly (Section 3.1), and ILD, which optimizes the diversity using item vectors (Section 3.2). We discuss the rationale, practical application and parameter choices in Section 4.

3.1 Interleaving Topics

Interleaving is a crude method that guarantees a 'balanced' spread across a categorical variable, called 'topic' hereafter. It is similar to the Round-Robin technique presented by Silva et al. [31], where the diversification dimension—in our case—is (news) topics. In our work, interleaving (Algorithm 1) is achieved by grouping all items by their topics, sorting the items within the topics by predicted relevance, and then iterating through the topics in a Round-Robin fashion to add the top item from each topic to the list. This continues until the required number of items is added to the re-ranked list. If a topic is exhausted before the required number of items is reached, it is skipped in subsequent iterations. For example, given a sorted list of recommendations (for a user) containing first 10 'Sports' articles, then 3 'Politics' articles, then 2 'Sports' articles again, and finally a single 'Economy' article. If we need to generate a list of six items, the interleaving would generate a list with the following topics: [Sports, Politics, Economy, Sports, Politics, Sports].

Algorithm 1 *Interleaving*

```
sorted_items_per_topic =
    group_items_per_topic(sorted_items)
# Get topics by first appearance in the sorted list
topics = get_topics(sorted_items)
recos: list[Id] = []
while len(recos) < n:
    topic = topics.next()  # round-robin
    if len(sorted_items_per_topic[topic]) > 0:
        top_item = sorted_items_per_topic[topic].pop()
        ranked_items.append(top_item)
```

An advantage of this method is its ability to ensure coverage of the maximum number of topics in the re-ranked list. However, its effectiveness depends on the quality of the topic classification in the metadata, which may be noisy or incomplete. Additionally, this method may include topics in a recommendation list that do not align with a user's preferences, as it prioritizes introducing new topics over providing further recommendations from more relevant ones.

3.2 Intra-List Diversification

As a second method, we use ILD to minimize the ILS of the recommended list, adapting the approach from the original work by Ziegler et al. [52]. While the original work used intra-list topic-similarity as the target metric, we extend this to intra-list vector-similarity. The implemented method assumes that each item has a vector representation, which can be derived from textual features (e.g., text embeddings) or by collaborative filtering models that generate item embeddings. An important advantage of using embeddings, compared to *Interleaving* or other topic-diversification methods, is that it does not depend on the quality of topic classification. Instead, it leverages vector representations, which are less susceptible to incomplete information or mislabeling than topics are. In order to achieve diversification, we iteratively remove items from the list of top-m candidates until a desired number of n items is reached ($m > n$). At each iteration, the item with the highest total similarity with all other items in the list is removed (Algorithm 2).

Algorithm 2 *ILD*

```
top_items = sorted_items[:m]
while len(top_items) > n:
    sim_mat = embedding_cosine_similarity(top_items)
    item_to_delete = sim_mat.sum(axis=column).argmax()
    top_items.remove(item_to_delete)
```

Unlike *Interleaving*, this method only re-ranks the top m most relevant items due to performance constraints and to account for user preferences. This method differs from other commonly used approaches [5, 45, 52] in that it removes items, instead of adding them one by one. A comparative disadvantage of this method is that it does not consider relevance for the user when removing articles after the top-m cutoff. In our previous example, any one of the 12 'Sports' articles could be removed (assuming they form a tightly clustered group) regardless of similarity. Although it constitutes a theoretical disadvantage when diversified articles are more likely to appear at the bottom of the list, we note that it also fosters a natural discovery experience, desirable in production systems. Users are presented with their most similar articles first, while more diverse items are displayed further down the list. If users are not interested in their most similar articles, the diverse items create opportunities for discovery. In this way, the diversified items are not imposed on users ahead of their preferred topics.

4 Production System Context

We deployed the models described in Section 3 and conducted our experiment on the Dutch large-scale production news platform NU.nl. The method choices were guided by the platform's specific characteristics, practical implementation considerations, and the requests from news editors. This section outlines the production system's context and constraints, details about the available data, and the design choices for implementing the diversification methods and evaluation setup. The news platform NU.nl is freely accessible, with all articles available without subscription plans. The platform features international and domestic news, with a substantial section dedicated to sports.

4.1 Candidate Pool and Recency Constraint

As news items get out-of-date quickly [2, 19, 28] and CTR falling sharply after less than two days [47], the platform enforces a recency policy: only articles published within the previous 24 hours are eligible for recommendation. All algorithms evaluated in this paper operate exclusively on this recency-filtered candidate set. In practice, this policy yields a pool of roughly 150–300 eligible articles at any given time during the study period.

4.2 Personalization

In this work, we considered personalization at three different positions on the news platform. The first position, located on the homepage, consists of 5 items displayed in positions 9–13 of the top articles shortlist (3M daily impressions, 180K daily clicks). The items curated by editors (positions 1–8) are removed from the personalized recommendations. The second position, also on the homepage but below the fold, contains 5 items; here without de-duplication (1M daily impressions, 12.5K daily clicks). The third position is a fully personalized page—called the 'For you page'—containing 20 items (5K daily impressions, 300 daily clicks). This 'For you page', only available on iOS and Android apps, had relatively low daily traffic, as it was not strongly advertised yet. For all three platform positions, the primary goal is to deliver content tailored to a user's specific interests, complementing the editorially curated content on the homepage, which highlighted the fundamentally important news. Additionally, a key requirement for personalization is to ensure high coverage of the daily published articles.

At the core of the RS, we use a content-based approach that matches items to users based on the similarity between the text embedding of an article and the topic embedding of the articles that the user has read. Candidate retrieval selects items from within the last 12h for the first position on top of the homepage and from the last 24h for the other two positions.

4.3 Topics

The news platform organizes its content into six high-level "navigational" topics, also referred to as "main sections". These topics are assigned to articles and relate to how the articles are shown on the website[1]. In addition to these six main sections, articles are assigned sub-section labels, which provide a more fine-granular categorization. The number of sub-sections varies by main section. For example, 'General' contains seven sub-sections, while 'Economy' has only two. The sub-sections are organized in a tree taxonomy; as such, each sub-section belongs to exactly one main section. During the experiment period, the platform featured a total of 25 sub-sections. In the remainder of this paper, we will refer to the main sections as "topics" and the sub-sections as "subtopics".

4.4 Diversification

The editorial team raised a concern that personalization could narrow the set of topics shown to users, potentially fostering filter bubbles, where users would miss out on the broader range of news topics available on the platform [38]. To address their concerns, we implemented a topic-based *Interleaving* approach (Algorithm 1) that maximizes topic coverage within a recommendation list while prioritizing the most relevant items within each topic. Interleaving by topic mirrors how newsrooms structure navigational sections across the site (and legacy media in general). This approach is simple, fast, and interpretable for editors; however, it ignores intra-topic content similarity, motivating our ILD variants.

However, with only six topics available, the limitations of this approach became evident despite improvements in the diversity of the recommended list. Frequently, topics ranked in positions 4 to 6 were irrelevant to the user but were still included to guarantee maximal topic coverage. The small number of topics also constrained the overall level of diversification. For example, the second article selected from a given topic was often very similar to the first. Using subtopics as an alternative was not viable because users who are most interested in a topic with many subtopics would still receive recommendations dominated by that single overarching topic, which the editorial team deemed unacceptable.

To overcome these limitations, we implemented ILD (Algorithm 2), which leverages embeddings generated from the textual content of the items, instead of relying on the quality of topic labeling. For our experiment, we used *two* types of content-based embeddings: (i) BERT, a transformer-based, pre-trained language model, that generates dense 512-dimensional vectors for each article based on its title and text [29]. (ii) TF-IDF, which uses a corpus of 50 days of articles to compute the IDF values, generating sparse 35K dimensional vectors. We chose these representations because prior research in the news domain shows that TF-IDF often aligns with user judgments [34], while context-based embeddings like BERT have demonstrated superior performance in other studies [21]. For the hyperparameter m, we set it to $2n$ (i.e., 10 for the two homepage positions and 40 for the personalized page). Through experimentation, we found that further increasing m beyond $2n$ rarely improved diversity in the final list but did increase inference time.

In terms of computational overhead, *Interleaving* has a marginal impact on latency, whereas ILD takes around 50–75ms, which remains within our production constraints.

5 Evaluation

We evaluated the methods on a large-scale production news platform using two complementary methods: an online experiment (A/B test) (Section 5.1) and a user study (Section 5.2).

5.1 Online Experiment

We conducted an online experiment (A/B test) in which the diversification approaches were integrated into the RS deployed in a real-world, live setting. Online experiments provide a realistic evaluation scenario because users are self-motivated and interact with the platform naturally [17, 18]. The online experiment employed a *between-subjects* design, dividing platform users equally into four groups: a control group *NoDiversification*, an *Interleaving* treatment group, and two ILD-based treatment groups (*ILD-BERT* and *ILD-TFIDF*). Users were randomly assigned to one of the groups, and assignments remained fixed for the duration of the experiment. The experiment was carried out for two full weeks in January 2025 and

[1]The six topics are 'General', 'Sports', 'Media & Culture', 'Economy', 'Interact with us', and 'Other'. The 'Interact with us' topic contains games and interactive content for the readers (e.g., questions).

was deployed across all three positions on the platform. According to the platform's editorial records, traffic patterns remained normal during the study period. To verify the effectiveness of the diversification methods, we computed the ILS scores of the recommended lists using the BERT item embeddings and analyzed the distribution of recommendations across the available topics. User engagement was measured using the widely adopted CTR metric. For brevity, this evaluation concentrates on the aggregate results across all three lists. Each list was also analyzed independently, and the findings were consistent across all of them.

To assess the significance of the CTR and ILS metrics, we use both the parametric 95% confidence intervals (CIs) of the difference between the two metrics at each time point (significant if the interval does not include 0) and the non-parametric Wilcoxon signed-rank test for paired samples (given that our samples are paired by time), with $p = 0.05$ and Bonferroni correction (with $m = 6$ to account for the multiple comparisons). We pair samples by hour because the natural temporal variance of the metrics has a stronger effect than any differences attributable to the chosen method. Therefore, to isolate the impact of the variants, we control for temporal variance by pairing samples within the same hour.

5.2 User Study

Following the online experiment (A/B test), we conducted a small-scale user study on the production news platform to evaluate users' perceptions of the diversified recommendations. Guided by the A/B test results, we focused on two variants: *NoDiversification* and the best-performing diversification variant, *ILD-BERT*. Participants rated two agreement items on a 5-point Likert scale ($[1, 5]$), where higher values indicate stronger agreement: (i) *The items in this list provide a rich variety*; (ii) *The items in this list are relevant to me.* To maximize participation in the production setting, we kept the questionnaire minimal. A small pilot with platform users indicated that the term 'variety' was better aligned with the study's intent than 'diversity'. In public discourse, 'diversity' is often associated with sensitive attributes (e.g., gender, ethnicity, political viewpoints), whereas our focus was topic diversity. This wording also aligns with prior questionnaire phrasing (e.g., [14]). The questionnaire consisted of two mentioned Likert-type items and does not constitute a validated psychometric scale; items were designed for face validity and pretested in the pilot for clarity.

The questions were presented as a pop-up on the fully personalized page (i.e., at the third platform position, as described in Section 4.2) on the production news platform's iOS and Android apps. This page displayed a total of 20 personalized recommendations; with no further items shown. To ensure that users had sufficient time to consider the items on the recommendation list, the pop-up appeared after 20 seconds of page activity. This threshold was suggested by the platform's user-study team to ensure surveying engaged users. The pop-up questionnaire targeted all users visiting the page, making the entire user base the sample population for this evaluation.

The user study was conducted over two consecutive full weeks at the beginning of March 2025. During the study period, on average, approximately 3500 unique users visited the page daily. Conducting the user study in a live, large-scale production environment increases external (ecological) validity due to the realistic usage context, but it also imposed typical production constraints (e.g., limited instrumentation and tooling; cf. Zangerle and Bauer [48]). Notably, the survey tool on the production site does *not* allow user identification, preventing linkage of responses across conditions.

Given these constraints, we employed a *quasi-experiment* using a *posttest-only, non-equivalent groups design*. All platform users were exposed to the *ILD-BERT* variant in week 1 and to the *NoDiversification* setup in week 2. Because users could not be identified, we treat responses as independent repeated cross-sections across weeks (i.e., between-subjects across weeks), while acknowledging potential period effects (history/seasonality) and selection differences. No major newsworthy events (e.g., sports or politics) occurred during the study period, reducing the risk of short-term seasonality/history bias; according to the platform's editorial records, traffic patterns remained normal during the study period. To avoid frustration and duplicate responses, users were excluded from receiving repeated pop-ups within the same week. Although some users could, in principle, have responded in both weeks, the overlap is unlikely given the low response rate. Approximately 11K surveys were sent out each week; of these, 182 users responded in week 1 and 184 in week 2, resulting in a response rate of ≈1.5%. This response rate is in line with previous user studies on the platform and makes cross-week overlap improbable, though not impossible. We therefore treat the two weekly samples as independent for analysis.

6 Results

Our main results, summarized in Table 1, show that *ILD-BERT* outperforms the other three variants in terms of CTR and *perceived relevance*. However, it performs worse in *perceived variety*. We discuss these results in Sections 6.1–6.3.

Table 1: Summary of main performance metric and user study results. Metric values are averaged over the entire experiment duration and all three platform positions. User study responses are reported as means. Mann–Whitney U test for user study; non-significant results ($p = 0.95$) are marked "n.s." and best-performing values are in bold.

Method	CTR	Perceived Relevance	Perceived Variety
NoDiversification	11.61%	3.38	**3.67**$^{n.s.}$
Interleaving	10.74%	–	–
ILD-BERT	**11.91%**	**3.69**	3.48
ILD-TFIDF	11.62%	–	–

6.1 Diversity Results

To verify that the investigated diversification methods operate as intended, we report how the methods affect the diversity of the recommended items. By explicitly measuring diversity outcomes, we address a limitation noted in prior work [e.g., 2, 28], namely that many studies proposing diversification methods in NRS do not assess or report whether their methods actually increase diversity.

Table 2 presents diversity outcomes in terms of ILS and topic diversity, where the latter is measured as the number of unique topics

per list (mode and percentage). These metrics serve a descriptive role: They document how each method changes the composition of recommendation lists and confirm that the intended diversification behavior is achieved. Because some methods explicitly optimize one of these metrics, these are not interpreted as comparative performance measures. Instead, effectiveness is assessed based on user behavior (CTR).

Table 2: Summary statistics of diversity analysis. ILS computation is based on the BERT embedding of the items recommended; mode gives the number of topics recommended.

Method	ILS (average)	Number of topics recommended@5 (mode)	(percentage)
NoDiversification	0.17	2	(40%)
Interleaving	0.16	**5**	**(70%)**
ILD-BERT	**0.11**	3	(50%)
ILD-TFIDF	0.13	2	(40%)

As illustrated in Fig. 1, pairwise relative comparisons of ILS (with 95% CIs[2]) indicate that *ILD-BERT* achieves the highest diversity, reflecting a diversity improvement of 37% compared to the control variant *NoDiversification*. As *ILD-BERT* directly optimizes this metric, it is no surprise that it performs best here. We do see that all diversification methods improve over the *NoDiversification* variant.

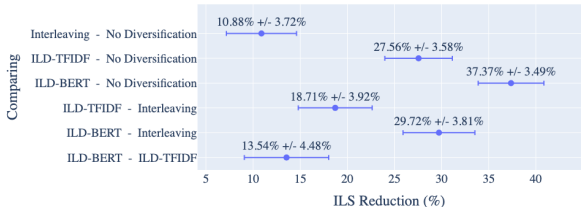

Figure 1: Pairwise relative comparison of ILS with 95% CIs. Values indicate the relative change when comparing the left variant to the right. E.g., *Interleaving* shows a 10.88% relative reduction in ILS compared to *NoDiversification*. A Wilcoxon signed-rank test ($p < 0.05$) confirmed the significance of the results, consistent with the CIs.

The results in Fig. 2 show the impact the various diversification methods have on the number of unique topics recommended in the top-5 recommendations. We chose top 5, as it matches our most prominent and most requested list on the homepage. *Interleaving* tends to maximize the number of topics (as it is designed to do), as fewer than 5 unique topics are recommended only in rare cases. These rare cases occur when only articles from 4 topics were available ('Interact with us' and 'Other' were not available on multiple occasions). Both *ILD-BERT* and *ILD-TFIDF* shift the distribution towards more unique topics, yet they almost never reach the extreme case of recommending 5 unique topics. Herein lies a big strength

[2]CIs are computed using time-aligned, paired samples, with each sample representing an hourly average value of the metric

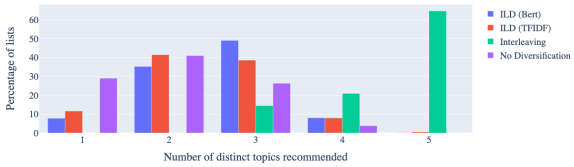

Figure 2: Distribution of recommendation lists by number of unique topics in the top-5 items (1–5). Bars indicate the frequency of lists in each category (e.g., 50% of lists generated by the *ILD-BERT* variant contain 3 unique topics in the top 5.

of the ILD methods: they manage to combine diversification with relevance and avoid extreme cases.

Although *Interleaving* strongly outperforms the other methods in terms of topic diversity, it scores lower with respect to ILS, as it does not optimize for text similarity. Consequently, the items it recommends may still be highly, textually, similar. For example, the topics 'General', 'Sports', and 'Media & Culture' could all include articles about the same celebrity. Furthermore, the second article within a topic is likely still highly similar in content.

Fig. 3 shows how the different diversification methods distribute recommendations across topics at the first position. *NoDiversification* has a strong allocation to the topics 'Sports' and 'General', very little to 'Economy', and almost none to 'Interact with us' and 'Other'. All diversification methods shift the recommendations away from the dominant topics and give more recommendations of the less recommended topics. Due to the used text embeddings, *ILD-BERT*—and *ILD-TFIDF* to a lesser extent—is less likely to remove items from broad topics like 'General' or 'Economy'.

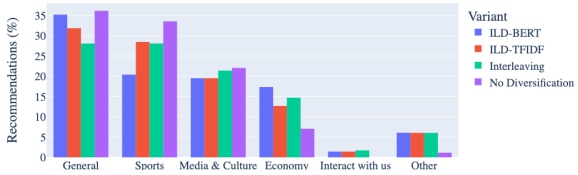

Figure 3: Frequency distribution of recommended topics by diversification method.

Fig. 4 provides a breakdown of how frequently each subtopic is recommended by each variant. The notable decline in 'Sports' recommendations, which we noted earlier, is primarily due to one subtopic: 'Football'—the largest 'Sports' subtopic. Here, we also observe that subtopics under the same topic also see different effects from ILD-based methods. For example, under the 'General' topic, 'World', 'Domestic', and 'Economy' increase in exposure, while the other subtopics receive lower exposure.

With *Interleaving*, the recommendation shifts are predictable—consistently reallocating recommendations from over-represented topics to smaller ones. In contrast, the embedding-based approaches encode more nuanced information than the topics or subtopics; with this, these methods are less likely to remove items from "broad" topics and subtopics during diversification.

We can assess how users interact with the diversified recommendations by comparing the click distribution to the recommendation

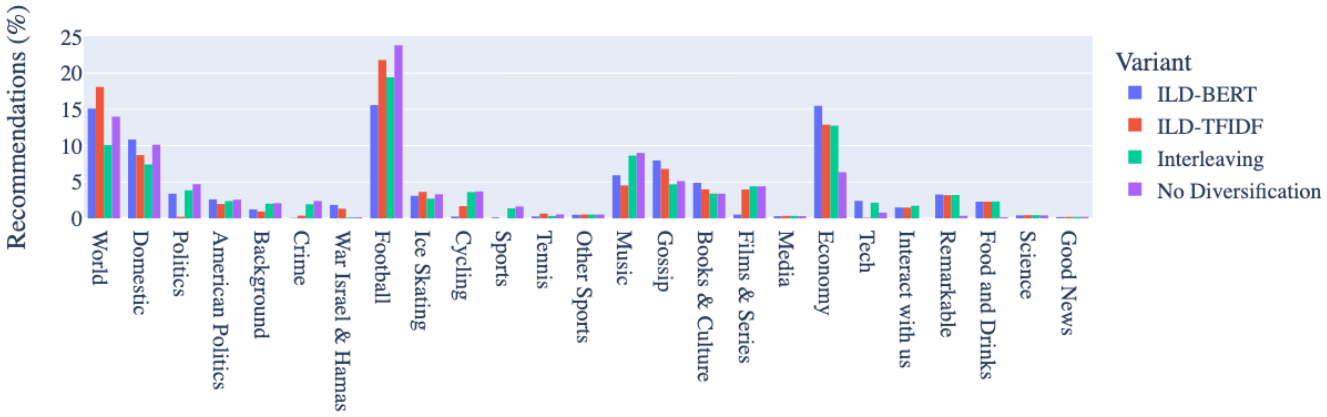

Figure 4: Frequency distribution of recommended subtopics by diversification method.

distribution. Fig. 5 shows the percentage of clicks for each topic relative to the total clicks within each variant. For *ILD-BERT*, while we observed a decrease in recommendations of the 'General' topic, the click percentage on this topic increased. This suggests that diversification allowed articles of the 'General' topic to appear higher in the recommendation list when, for example, sports content would be the user's closest interest. Interestingly, when comparing *ILD-TFIDF* and *Interleaving* on the 'Sports' topic, we observe that both variants expose users to about the same percentage of 'Sports' articles in the recommendations, but the click percentages differ substantially. This could indicate that while *Interleaving* forces 'Sports' articles onto every user, the ILD approaches are likely to recommend sports content only to people interested in this topic.

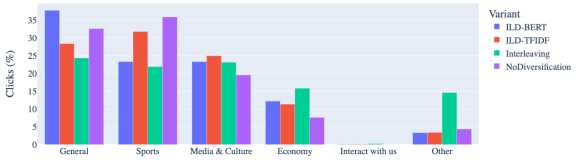

Figure 5: Topic click frequency by diversification method.

6.2 Results in Terms of Click-Through Rates

Pairwise relative comparisons of CTRs (with 95% CIs) show that *ILD-BERT* outperforms all other methods.

This result suggests that the *ILD-BERT* variant is more effective than the other methods at including diverse items in the recommendation lists that are also relevant for the respective user, increasing the likelihood that users find a relevant item in a diversified list.

Interestingly, the variant with the highest ILS improvement, *ILD-BERT*, also achieved the greatest improvement in terms of CTR. This suggests that optimizing for ILS using dense text embeddings can enhance user engagement, though it is no guarantee that optimal ILS leads to optimal CTR. *Interleaving*, while successful in diversifying topics, performed poorly in terms of CTR. This result showcases that maximizing topic diversity alone does not necessarily lead to more user interaction. In our setting, which includes only six topics and a mix of broad and narrow topics, we observe

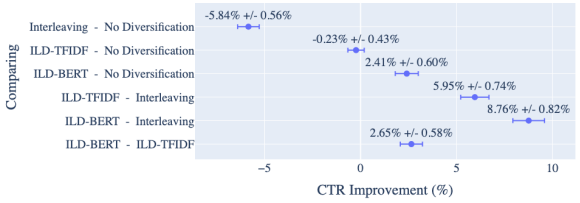

Figure 6: Pairwise relative comparisons of CTR with 95% CIs. Values indicate the relative change when comparing the left variant to the right. E.g., *Interleaving* has a 5.84% lower CTR than *NoDiversification*. A Wilcoxon signed-rank test ($p < 0.05$) confirmed the significance of the results, consistent with the CIs.

that using text embeddings instead of categorical variables gives a more nuanced approach to diversification, leading to better outcomes in terms of CTR. Additionally, we observe that the choice of embedding plays a critical role. In our case, the dense embedding (*ILD-BERT*) outperformed the sparse embedding (*ILD-TFIDF*).

6.3 User Perceptions

Fig. 7 illustrates the *perceived relevance* and *perceived variety* of the recommended items for the *ILD-BERT* and the *NoDiversification* groups. A Mann-Whitney U test indicates statistically significant differences in *perceived relevance* of the items between the two groups ($U = 19785.0$, $p = 0.0017$), with *ILD-BERT* achieving higher mean relevance scores (mean difference = 0.31, or approximately 10%). For *perceived variety*, a Mann–Whitney U test yielded a p-value of 0.059 ($U = 14918.5$), which is marginally above the conventional significance threshold ($\alpha = 0.05$). Interestingly, the *perceived variety* values are slightly lower for *ILD-BERT* compared to *NoDiversification* (mean difference = −0.18, or approximately −5%). Although not statistically significant, and the observed difference is small in magnitude, the observed trend warrants further investigation to understand its potential implications. In other words, users perceive the recommendations diversified through *ILD-BERT* as more relevant, even though they do not perceive a higher variety in the recommended items.

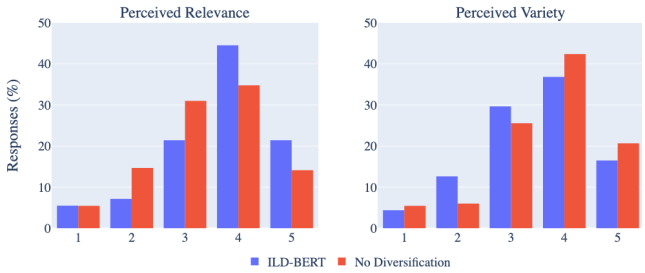

Figure 7: Distributions of Likert responses (1 = Strongly disagree ... 5 = Strongly agree) for *perceived relevance* and *perceived variety*, comparing *ILD-BERT* and *NoDiversification*.

7 Discussion

7.1 Discussion of Results

Our findings provide evidence that embedding-based diversification can improve engagement in a live NRS without sacrificing relevance. In our deployment, *ILD-BERT* achieved gains in CTR, higher objective diversity (lower ILS), and higher *perceived relevance* than the non-diversified *NoDiversification*. This suggests that semantic representations enable RSs to diversify content preserving alignment with users' interests, demonstrating that diversity and short-term engagement are not necessarily in conflict.

In contrast, topic-based *Interleaving* maximized the number of unique topics per list but resulted in the lowest CTR. This suggests that exposure to more topics alone is insufficient to create user value. As broad topic categories (e.g., 'General' or 'Sports') often contain highly heterogeneous content, diversification based on semantic embeddings therefore appear better suited to maintaining relevance than forcing categorical topic shifts. This effect is likely amplified by the production taxonomy used in our study, which comprised only six topics with uneven scope and relevance. Such coarse and imbalanced topic structures limit the potential of topic-level diversification, whereas embedding-based methods operate at the item-representation level and are therefore less dependent on the quality or granularity of editorial categories. This finding challenges the assumption that increasing topic coverage inherently benefits users. The click distribution analysis supports this interpretation. Under *ILD-BERT*, some topics were recommended less but received a higher share of clicks, indicating more efficient exposure. Diversification may thus help surface relevant items that would otherwise be overshadowed by highly similar content, increasing the chance that at least one item in the list matches the user's immediate interest.

A further notable result is the disconnect between objective and subjective diversity. Despite improvements in ILS, *ILD-BERT* did not increase *perceived variety* and even showed a slight (non-significant) decrease. This aligns with prior evidence showing that similarity-based diversity metrics capture aspects of similarity reduction that are not not directly reflected in user's perceptions of how varied a list feels. Semantically diverse items within the same broad topic may still appear similar from a user perspective, whereas explicit topical shifts are likely more salient. Thus, what RS optimize as diversity is not necessarily what users perceive as variety.

Finally, the comparison between *ILD-BERT* and *ILD-TFIDF* highlights the role of the representation space. In our setting, dense contextual embeddings were more effective at distinguishing semantically similar items than sparse lexical features. The choice of embedding therefore can materially influence diversification outcomes rather than being a minor implementation detail.

7.2 Limitations

Conducting the evaluation in a live production environment enhances external validity but introduces operational constraints that affect internal validity and generalizability. For instance, the user study followed a fixed week order, and participants could not be tracked across weeks. We therefore analyzed the two weekly samples as independent repeated cross-sections, acknowledging (yet, unlikely) potential participant overlap and time-related effects such as seasonality. Participation in the user study relied on a pop-up survey, which may introduce self-selection bias as respondents could differ systematically from non-respondents. We further acknowledge that results may depend on platform-specific aspects, including de-duplication on the homepage, the number and positions of personalized slots, and how recommendations are presented. Findings may unfold differently on platforms with different designs or interaction patterns. Similarly, the relative effectiveness of diversification strategies may depend on the underlying recommender and feature space. For instance, different underlying models, richer metadata, or other content types (e.g., images or videos) could yield different results. Additionally, the recency window was fixed to 24 hours due to production constraints. Different temporal settings could influence both diversity and engagement outcomes. Finally, as in any field study, we could not fully control for contemporaneous events or seasonal patterns that may have affected user behavior.

Despite these limitations, the study provides high face validity and practical insights into the deployment of diversification strategies in a real-world NRS.

8 Conclusion

This work presented a production-scale evaluation of diversification strategies in a large-scale NRS, combining an online experiment (A/B testing) and a complementary user study, which was also conducted on a large-scale production news platform.

Our contributions are threefold: (1) We provide evidence from a real-world setting on how diversification strategies affect objective diversity, user engagement, and user perceptions when deployed in a large-scale production NRS. (2) *ILD-BERT* improved CTRs and increased *perceived relevance*, showing that diversification can support user engagement without degrading users' perception of recommendation quality. (3) Despite clear improvements in ILS and number of topics recommended, users did not perceive higher variety. This highlights that the algorithms and models captured different aspects than what users perceive as variety in the news recommendation context.

Overall, our findings show that the effects of diversification depends on how diversity is operationalized and represented. For user modeling and personalization, this emphasizes the need to align algorithmic diversity objectives with users' perceptions, rather than relying solely on common similarity-based metrics.

Acknowledgments

This publication was supported by the Excellence in Digital Sciences and Interdisciplinary Technologies (EXDIGIT) project, funded by Land Salzburg under grant number 20204-WISS/263/6-6022.

References

[1] Himan Abdollahpouri, Edward C. Malthouse, Joseph A. Konstan, Bamshad Mobasher, and Jeremy Gilbert. 2021. Toward the Next Generation of News Recommender Systems. In *Companion Proceedings of the Web Conference 2021* (Ljubljana, Slovenia) *(WWW '21)*. ACM, New York, NY, USA, 402–406. doi:10.1145/3442442.3452327

[2] Christine Bauer, Chandni Bagchi, Olusanmi A Hundogan, and Karin van Es. 2024. Where are the Values? A Systematic Literature Review on News Recommender Systems. *ACM Transactions on Recommender Systems* 2, 3, Article 23 (2024), 40 pages. doi:10.1145/3654805

[3] Abraham Bernstein, Claes de Vreese, Natali Helberger, Wolfgang Schulz, Katharina Zweig, Christian Baden, Michael A. Beam, Marc P. Hauer, Lucien Heitz, Pascal Jürgens, Christian Katzenbach, Benjamin Kille, Beate Klimkiewicz, Wiebke Loosen, Judith Moeller, Goran Radanovic, Guy Shani, Nava Tintarev, Suzanne Tolmeijer, Wouter van Atteveldt, Sanne Vrijenhoek, and Theresa Zueger. 2021. Diversity in News Recommendation (Dagstuhl Perspectives Workshop 19482). *Dagstuhl Manifestos* 9, 1 (2021), 43–61. doi:10.4230/DagMan.9.1.43

[4] Balázs Bodó, Natali Helberger, Sarah Eskens, and Judith Möller. 2019. Interested in Diversity: The Role of User Attitudes, Algorithmic Feedback Loops, and Policy in News Personalization. *Digital Journalism* 7, 2 (2019), 206–229. doi:10.1080/21670811.2018.1521292

[5] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia) *(SIGIR '98)*. ACM, New York, NY, USA, 335–336. doi:10.1145/290941.291025

[6] Pablo Castells, Neil Hurley, and Saúl Vargas. 2022. Novelty and Diversity in Recommender Systems. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer US, New York, NY, USA, 603–646. doi:10.1007/978-1-0716-2197-4_16

[7] Lucas Colucci, Prachi Doshi, Kun-Lin Lee, Jiajie Liang, Yin Lin, Ishan Vashishtha, Jia Zhang, and Alvin Jude. 2016. Evaluating Item-Item Similarity Algorithms for Movies. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 2141–2147. doi:10.1145/2851581.2892362

[8] Patrik Dokoupil, Ludovico Boratto, and Ladislav Peska. 2024. User Perceptions of Diversity in Recommender Systems. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP '24)*. ACM, New York, NY, USA, 212–222. doi:10.1145/3627043.3659555

[9] Patrik Dokoupil, Ludovico Boratto, and Ladislav Peška. 2025. Accuracy and beyond-accuracy perspectives of controllable multi-objective recommender systems. *Information Processing & Management* 62, 6, Article 104267 (2025), 36 pages. doi:10.1016/j.ipm.2025.104267

[10] Ryan Evans, Daniel Jackson, and Jaron Murphy. 2023. Google News and Machine Gatekeepers: Algorithmic Personalisation and News Diversity in Online News Search. *Digital Journalism* 11, 9 (2023), 1682–1700. doi:10.1080/21670811.2022.2055596

[11] Lucien Heitz, Juliane A. Lischka, Rana Abdullah, Laura Laugwitz, Hendrik Meyer, and Abraham Bernstein. 2023. Deliberative Diversity for News Recommendations: Operationalization and Experimental User Study. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore) *(RecSys '23)*. ACM, New York, NY, USA, 813–819. doi:10.1145/3604915.3608834

[12] Lucien Heitz, Juliane A. Lischka, Alena Birrer, Bibek Paudel, Suzanne Tolmeijer, Laura Laugwitz, and Abraham Bernstein. 2022. Benefits of Diverse News Recommendations for Democracy: A User Study. *Digital Journalism* 10, 10 (2022), 1710–1730. doi:10.1080/21670811.2021.2021804

[13] Karl Higley, Robin Burke, Michael D Ekstrand, and Bart P Knijnenburg. 2025. What News Recommendation Research Did (But Mostly Didn't) Teach Us About Building A News Recommender. In *Beyond Algorithms: Reclaiming the Interdisciplinary Roots of Recommender Systems Workshop (CEUR Workshop Proceedings, Vol. 4063)*, Eva Zangerle, Alan Said, and Christine Bauer (Eds.). CEUR-WS.org, Aachen, Germany, Article 1, 15 pages. https://ceur-ws.org/Vol-4063/paper1.pdf

[14] Mathias Jesse, Christine Bauer, and Dietmar Jannach. 2022. Intra-List Similarity and Human Diversity Perceptions of Recommendations: the Details Matter. *User Modeling and User-Adapted Interaction* 33, 4 (2022), 769–802. doi:10.1007/s11257-022-09351-w

[15] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Transactions on Interactive Intelligent Systems* 7, 1, Article 2 (2016), 42 pages. doi:10.1145/2926720

[16] Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News Recommender Systems: Survey and Roads Ahead. *Information Processing & Management* 54, 6 (2018), 1203–1227. doi:10.1016/j.ipm.2018.04.008

[17] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online Controlled Experiments at Large Scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*. ACM, 1168–1176. doi:10.1145/2487575.2488217

[18] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. 2008. Controlled Experiments on the Web: Survey and Practical Guide. *Data Mining and Knowledge Discovery* 18, 1 (2008), 140–181. doi:10.1007/s10618-008-0114-1

[19] Johannes Kruse, Kasper Lindskow, Saikishore Kalloori, Marco Polignano, Claudio Pomo, Abhishek Srivastava, Anshuk Uppal, Michael Riis Andersen, and Jes Frellsen. 2024. EB-NeRD a large-scale dataset for news recommendation. In *Proceedings of the Recommender Systems Challenge 2024* (Bari, Italy) *(RecSysChallenge '24)*. Association for Computing Machinery, New York, NY, USA, 11 pages. doi:10.1145/3687151.3687152

[20] Matevž Kunaver and Tomaž Požrl. 2017. Diversity in Recommender Systems – A Survey. *Knowledge-Based Systems* 123 (2017), 154–162. doi:10.1016/j.knosys.2017.02.009

[21] Jingang Liu, Chunhe Xia, Xiaojian Li, Haihua Yan, and Tengteng Liu. 2020. A BERT-Based Ensemble Model for Chinese News Topic Prediction. In *Proceedings of the 2020 2nd International Conference on Big Data Engineering* (Shanghai, China) *(BDE '20)*. ACM, New York, NY, USA, 18–23. doi:10.1145/3404512.3404524

[22] Mats Mulder, Oana Inel, Jasper Oosterman, and Nava Tintarev. 2021. Operationalizing Framing to Support Multiperspective Recommendations of Opinion Pieces. 11 pages. doi:10.1145/3442188.3445911

[23] Umberto Panniello, Alexander Tuzhilin, and Michele Gorgoglione. 2014. Comparing context-aware recommender systems in terms of accuracy and diversity. *User Modeling and User-Adapted Interaction* 24, 1 (2014), 35–65. doi:10.1007/s11257-012-9135-y

[24] Javier Parapar and Filip Radlinski. 2021. Towards Unified Metrics for Accuracy and Diversity for Recommender Systems. In *Proceedings of the 15th ACM Conference on Recommender Systems* (Amsterdam, The Netherlands) *(RecSys '21)*. ACM, New York, NY, USA, 75–84. doi:10.1145/3460231.3474234

[25] Lorenzo Porcaro, Emilia Gómez, and Carlos Castillo. 2022. Perceptions of Diversity in Electronic Music: the Impact of Listener, Artist, and Track Characteristics. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1, Article 109 (April 2022), 26 pages. doi:10.1145/3512956

[26] Shaina Raza and Chen Ding. 2020. A Regularized Model to Trade-off between Accuracy and Diversity in a News Recommender System. In *2020 IEEE International Conference on Big Data (Big Data 2020)*. IEEE, 551–560. doi:10.1109/BigData50022.2020.9378340

[27] Shaina Raza and Chen Ding. 2021. Deep Neural Network to Tradeoff between Accuracy and Diversity in a News Recommender System. In *2021 IEEE International Conference on Big Data (Big Data 2021)*. IEEE, 5246–5256. doi:10.1109/BigData52589.2021.9671467

[28] Shaina Raza and Chen Ding. 2022. News Recommender System: a Review of Recent Progress, Challenges, and Opportunities. *Artificial Intelligence Review* 55, 1 (2022), 749–800. doi:10.1007/s10462-021-10043-x

[29] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. doi:10.18653/v1/D19-1410

[30] Alexander Semenov, Maciej Rysz, Gaurav Pandey, and Guanglin Xu. 2022. Diversity in News Recommendations Using Contextual Bandits. *Expert Systems with Applications* 195, Article 116478 (2022), 8 pages. doi:10.1016/j.eswa.2021.116478

[31] Pedro Silva, Bhawna Juneja, Shloka Desai, Ashudeep Singh, and Nadia Fawaz. 2023. Representation Online Matters: Practical End-to-End Diversification in Search and Recommender Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. ACM, New York, NY, USA, 1735–1746. doi:10.1145/3593013.3594112

[32] Manel Slokom, Savvina Daniil, and Laura Hollink. 2025. How to Diversify any Personalized Recommender?. In *Advances in Information Retrieval*, Claudia Hauff, Craig Macdonald, Dietmar Jannach, Gabriella Kazai, Franco Maria Nardini, Fabio Pinelli, Fabrizio Silvestri, and Nicola Tonellotto (Eds.). Springer Nature Switzerland, Cham, 307–323. doi:10.1007/978-3-031-88717-8_23

[33] Alain D. Starke, Sebastian Øverhaug, and Christoph Trattner. 2021. Predicting Feature-Based Similarity in the News Domain Using Human Judgments. In *Proceedings of the 9th International Workshop on News Recommendation and Analytics (INRA 2021)* (Amsterdam, The Netherlands) *(CEUR Workshop Proceedings)*. Aachen, Germany, 1–18. https://ceur-ws.org/Vol-3143/paper1.pdf

[34] Alain D. Starke, Vegard R. Solberg, Sebastian Øverhaug, and Christoph Trattner. 2024. Examining the Merits of Feature-Specific Similarity Functions in the News Domain Using Human Judgments. *User Modeling and User-Adapted Interaction* 34, 4 (2024), 995–1042. doi:10.1007/s11257-024-09412-2

[35] Christoph Trattner and Dietmar Jannach. 2020. Learning to Recommend Similar Items from Human Judgements. *User Modeling and User-Adapted Interaction* 30, 1 (2020), 49 pages. doi:10.1007/s11257-019-09245-4

[36] Célina Treuillier, Sylvain Castagnos, Evan Dufraisse, Özlem Özgöbek, and Armelle Brun. 2026. Shift your Focus for the Greater Good: Improving Fairness at no cost for Accuracy and Diversity in News Recommender Systems. *ACM Transactions on Recommender Systems* (Jan. 2026). doi:10.1145/3790097

[37] Célina Treuillier, Sylvain Castagnos, Özlem Özgöbek, and Armelle Brun. 2024. Beyond Trade-offs: Unveiling Fairness-Constrained Diversity in News Recommender Systems. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP '24)*. ACM, New York, NY, USA, 143–148. doi:10.1145/3627043.3659571

[38] Karin van Es and Dennis Nguyen. 2024. Balancing Needs and Values: A Multi-Stakeholder Examination of Algorithmic News Recommenders in the Netherlands. *Journalism Practice* (2024), 1–19. doi:10.1080/17512786.2024.2392654

[39] Saúl Vargas, Linas Baltrunas, Alexandros Karatzoglou, and Pablo Castells. 2014. Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems* (Foster City, Silicon Valley, CA, USA) *(RecSys '14)*. ACM, New York, NY, USA, 209–216. doi:10.1145/2645710.2645743

[40] Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, Daan Odijk, and Maarten De Rijke. 2022. RADio – Rank-Aware Divergence Metrics to Measure Normative Diversity in News Recommendations. In *Proceedings of the 16th ACM Conference on Recommender Systems* (Seattle, WA, USA) *(RecSys '22)*. ACM, New York, NY, USA, 208–219. doi:10.1145/3523227.3546780

[41] Sanne Vrijenhoek, Savvina Daniil, Jorden Sandel, and Laura Hollink. 2024. Diversity of What? On the Different Conceptualizations of Diversity in Recommender Systems. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (Rio de Janeiro, Brazil) *(FAccT '24)*. ACM, New York, NY, USA, 573–584. doi:10.1145/3630106.3658926

[42] Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a Mission: Assessing Diversity in News Recommendations. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. ACM, New York, NY, USA, 173–183. doi:10.1145/3406522.3446019

[43] Charlie Wang, Arpita Agrawal, Xiaojun Li, Tanima Makkad, Ejaz Veljee, Ole Mengshoel, and Alvin Jude. 2017. Content-Based top-N Recommendations with Perceived Similarity. In *Proceedings of the 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2017)*. IEEE, 1052–1057. doi:10.1109/SMC.2017.8122750

[44] Yuyan Wang, Cheenar Banerjee, Samer Chucri, Fabio Soldo, Sriraj Badam, Ed H. Chi, and Minmin Chen. 2025. Beyond Item Dissimilarities: Diversifying by Intent in Recommender Systems. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1* (Toronto, ON, Canada) *(KDD '25)*. ACM, New York, NY, USA, 2672–2681. doi:10.1145/3690624.3709429

[45] Mark Wilhelm, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. 2018. Practical Diversified Recommendations on YouTube with Determinantal Point Processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*. ACM, New York, NY, USA, 2165–2173. doi:10.1145/3269206.3272018

[46] Martijn C. Willemsen, Mark P. Graus, and Bart P. Knijnenburg. 2016. Understanding the role of latent feature diversification on choice difficulty and satisfaction. *User Modeling and User-Adapted Interaction* 26, 4 (2016), 347–389. doi:10.1007/s11257-016-9178-6

[47] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. MIND: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, 3597–3606. doi:10.18653/v1/2020.acl-main.331

[48] Eva Zangerle and Christine Bauer. 2022. Evaluating Recommender Systems: Survey and Framework. *Comput. Surveys* 55, 8, Article 170 (2022), 38 pages. doi:10.1145/3556536

[49] Yuying Zhao, Yu Wang, Yunchao Liu, Xueqi Cheng, Charu C. Aggarwal, and Tyler Derr. 2025. Fairness and Diversity in Recommender Systems: A Survey. *ACM Transactions on Intelligent Systems and Technology* 16, 1, Article 2 (Jan. 2025), 28 pages. doi:10.1145/3664928

[50] Yongsen Zheng, Guohua Wang, Yang Liu, and Liang Lin. 2024. Diversity Matters: User-Centric Multi-Interest Learning for Conversational Movie Recommendation. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. ACM, New York, NY, USA, 9515–9524. doi:10.1145/3664647.3680909

[51] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107, 10 (2010), 4511–4515. doi:10.1073/pnas.1000488107

[52] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving Recommendation Lists through Topic Diversification. In *Proceedings of the 14th International Conference on World Wide Web* (Chiba, Japan) *(WWW '05)*. ACM, New York, NY, USA, 22–32. doi:10.1145/1060745.1060754