Personalizing Diversity in News Recommendation: A User Study

Aishwarya Satwani[1], Robin Burke[2]

[1]Department of Computer Science, University of Colorado Boulder
[2]Department of Information Science, University of Colorado Boulder

Abstract

News recommendation systems must balance personalized content with diverse information exposure. Research has proposed various metrics and diversification algorithms to address this balance. However, the vast majority of this research has taken place with a limited number of datasets in offline settings that do not capture users' organic interactions with diverse content. In this research, we utilized the POPROX news recommendation platform to conduct a user study to compare generalized news diversification with a personalized model that increases diversity specifically for users who have indicated interest in diverse news topics. Our results indicate trends favorable to personalized diversification: subscribers in the personalized condition show higher intra-list diversity and maintained engagement relative to the non-personalized condition, though click sparsity in our naturalistic setting limits statistical power.

Keywords

recommender systems, diversity, news recommendation, user study, online study

1. Introduction

News recommendation systems increasingly influence public opinion and are therefore an important subject of study. This paper studies user responses to news diversity. Journalists understand diversity as operating across at least four dimensions: medium, topics, viewpoints, and source [1]. In this work, we focus on topic diversity.

A typical finding in offline studies is that there is a "diversity tax," where recommendation accuracy and click-through rates decline as diversification conflicts with individual user preferences. However, offline metrics may not align with user responses to off-policy recommendations. Hence the importance of user studies to explore user responses in context.

Many diversification methods generally assume that the preference or tolerance for diversity is constant across all users. For this reason, we seek to adapt the diversity of recommendations to users based on their propensity for diversity. This paper reports on a study conducted using the Platform for OPen Recommendation and Online eXperimentation (POPROX). We employed two treatments: one used reranking to increase news diversity in a uniform manner and another that adopted a personalized strategy.

With this study, we address the following research questions:

RQ1 How do POPROX readers respond to increased news diversity in the different treatments?
RQ2 Do users' explicit topic preferences provide a useful proxy for their interest in diverse news?
RQ3 To what extent does personalization help manage the tradeoff between engagement and the integration of diverse content?

FRAME'26: Methodology First - Rethinking Research Assessment in RecSys Workshop, September 28, 2026, Minneapolis, Minnesota, USA

aishwarya.satwani@colorado.edu (A. Satwani); robin.burke@colorado.edu (R. Burke)
0009-0007-7159-8133 (A. Satwani); 0000-0001-5766-6434 (R. Burke)

© 2026 Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-26

CEUR-WS.org/Vol-4341/paper9.pdf

2. Related Work

Recommender systems research has recognized the importance of news platforms, where recommenders take on an active role as gatekeepers [2]. At the same time, news is an area where normative concerns in the form of journalistic values, including diversity, may come into conflict with the consumer-oriented goals typically adopted in recommendation. Helberger [3] argues that democratic theories such as liberal, participatory, and deliberative form a logical normative framework for evaluating news recommenders, especially with respect to diversity.

Diversity in RSs is concerned with recommending items that appeal to users' broad set of interests [4]. In the context of news recommendation, however, Vrijenhoek et al. [5, 6] show that the choice of a diversity objective is, implicitly or explicitly, an injection of a journalistic value tied to an organizational mission. Tintarev et al. [7] broaden this discussion by developing a distance measure for diversity within a topic, enabling diverse recommendations while maintaining topic relevance through an adaptation of Maximal Marginal Relevance (MMR) [8], evaluated in both offline and real-user settings.

Recent work has emphasized that diversity should not be treated as a uniform objective across users. Eskandanian et al. [9] introduces the notion of user-specific diversity tolerance, arguing that individuals vary significantly in their preference for diverse recommendations.

Personalized reranking in the form of the PFAR algorithm was proposed for fairness-aware recommendation [10]. PFAR promotes items from underrepresented groups based on each user's diversity tolerance relative to those groups. Our work builds on this by adapting it for a diversity objective.

To understand the implications of news diversity, empirical user studies are needed, but such studies remain rare [11]. Heitz et al. [12] compares three recommendation conditions: diversity-based, personalization-based, and chronological ranking, using a controlled user study with both implicit behavioral signals and explicit ratings. The study found that recommendation design can actively nudge user exposure: while heavy news consumers across all conditions tend to engage with ideologically opposing viewpoints, diversity-oriented recommendation strategies encourage broader appreciation of cross-cutting content. Another user study [13] with 143 participants sought to operationalize the notion of a deliberative public sphere for news recommendations, showing the impact on news usage, and the influence on political knowledge, attitudes and voting behavior. In both of these studies, however, the users were paid participants in a limited time study.

Basso et al. [14] finds that diverse recommendations may shift exposure toward cross-cutting content with a potential depolarizing effect, even without significant gains in subjective well-being.

Vercoutere et al. [15] describes a user study aimed at encouraging exploration of previously unseen topics, using graph-based enhancement of user profiles. Their online experiment ($n = 288$) significantly improves both recommendation and consumption diversity with minimal impact on accuracy, while also increasing user satisfaction.

Our study differs from this prior work in two ways. First, we have authentic long-time users of a news system and second, rather than a global approach to diversity enhancement, we include a treatment in which the degree of diversification is personalized.

3. POPROX Platform

POPROX is an online experimentation platform designed to support recommender systems research with real users in a naturalistic, longitudinal setting. The platform delivers daily newsletters comprising 10 news articles, selected from a candidate pool of approximately 100–150 articles retrieved via APIs from the Associated Press (AP), along with associated metadata and topic annotations.[1] During onboarding, users specify their preferences across 14 predefined news topics using a 5-point scale. Figure 1 shows a sketch of the system's operation.

[1]This experiment used v1.0 of the POPROX newsletter. The current v2.0 as of this writing offers a newsletter with 12 articles divided into topical sections. See poprox.ai for additional information.

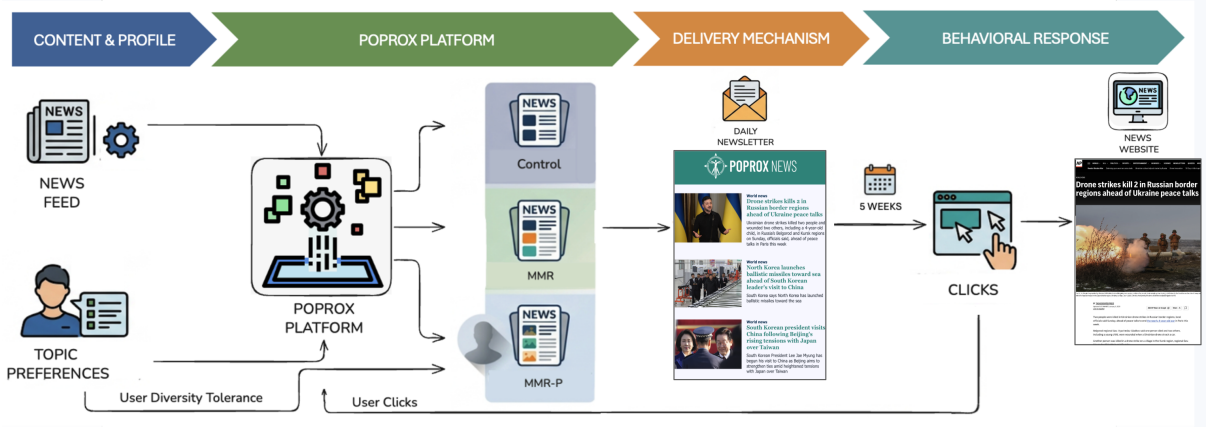


Figure 1: POPROX flowchart showing the four main stages of the system's operation, moving news content from our provider through newsletter generation to email delivery (a cropped example of the newsletter showing the first 3 (of 10) articles) and user interaction. Our experimental treatment was applied to the second stage.

A central feature of POPROX is its support for experimentation through its modular integration with external recommender systems. POPROX manages critical operational aspects, including subscriber interactions, human-subjects protections, and newsletter delivery.

Our code implemented only our experimental treatments[2]. The experiment can be reproduced by researchers who register as experimenters on the POPROX platform at poprox.ai.

4. Methodology

We conducted our experiment for a duration of five weeks. The experiment ran from December 2025 through January 2026, the slot allocated to us within the POPROX experimentation schedule. Participants were POPROX subscribers randomly assigned to one of three conditions. This study is exploratory; given the sample size, we treat all engagement differences as descriptive trends rather than confirmatory findings.

As shown in Figure 1, each day the platform ingests AP articles, applies condition-specific ranking, delivers a newsletter via a daily email, and records click-throughs to read articles on the AP website.

Our experimental treatments were implemented by extending the default ranking framework of the POPROX platform. The experiment involved creating three conditions: (1) Baseline (Control): the platform's standard ranking algorithm; (2) MMR (Generalized Diversification): a Maximal Marginal Relevance (MMR) reranker; (3) MMR-P (Personalized Diversification): a personalized version of MMR.

4.1. Condition 1: Baseline Recommender

The control condition consists of POPROX's standard recommender: a multi-signal neural news recommender built on NRMS [16], pretrained on the MIND large dataset [17]. The model integrates implicit click history, explicit topic preferences, and sparse article-level feedback into a unified relevance score.

The candidate set consists of Associated Press articles available on the platform at recommendation time (typically 100–150 articles per day). Each candidate article is represented using the pretrained NRMS encoder. Article titles are tokenized and truncated to 30 tokens before being passed through a multi-head self-attention encoder. The resulting article embedding $\mathbf{a}_i$ has dimensionality 768.

For each user u, up to 50 previously clicked articles are used to construct a user embedding. Let $H_u = \{a_1, \ldots, a_n\}$ denote the user's most recent clicks. Their embeddings are aggregated using the NRMS user encoder: $\mathbf{u}_{click} = \text{UserEncoder}(\mathbf{a}_1, \ldots, \mathbf{a}_n)$. If no click history is available, $\mathbf{u}_{click}$ is null.

During onboarding, users rate predefined editorial topics on a 5-point Likert scale. Each topic is associated with a synthetic definition article, which is encoded using the same NRMS news encoder to

[2]https://github.com/that-recsys-lab/poprox-rec-diverse

produce a topic embedding. User topic preferences are transformed into virtual click histories. For a topic with preference value p, the topic is repeated $|p-3|+1$ times when constructing positive or negative preference embeddings. These virtual clicks are passed through the NRMS user encoder and then normalized: $< \mathbf{u}^+_{topic}, \mathbf{u}^-_{topic} >$.

Users may provide article-level feedback as a thumbs-up or thumbs-down, but this optional feature was rarely used by subjects in our experiment. Positive and negative feedback are processed separately and aggregated using the NRMS user encoder to produce: $< \mathbf{u}^+_{fb}, \mathbf{u}^-_{fb} >$.

The final ranking score combines these values via dot-product similarity across embeddings (as shown below) with explicit preference signals s_{xp} weighted twice as strongly as implicit click history s_{click}.

1.	**Implicit click score**	$s_{click}(u, a_i) = \mathbf{a}_i^\top \mathbf{u}_{click}$
2.	**Topic score**	$s_{topic}(u, a_i) = \mathbf{a}_i^\top \mathbf{u}^+_{topic} - \mathbf{a}_i^\top \mathbf{u}^-_{topic}$
3.	**Feedback score**	$s_{fb}(u, a_i) = \mathbf{a}_i^\top \mathbf{u}^+_{fb} - \mathbf{a}_i^\top \mathbf{u}^-_{fb}$
4.	**Explicit preference score**	$s_{xp}(u, a_i) = (s_{topic}(u, a_i) + s_{fb}(u, a_i))/2$
5.	**Final fused score**	$s(u, a_i) = (s_{click}(u, a_i) + 2s_{xp}(u, a_i))/3$

Before ranking, a rule-based topic filter is applied using the user's stated topic preferences. For each article a_i, topics are extracted from entity labels provided by the news API and filtered to exclude low and retain high preference topics. The final recommendation slate consists of the top $k = 10$ articles ranked in descending order of $s(u, a_i)$. If filtering produces fewer than k candidates, remaining slots are filled from the unfiltered ranking. This baseline serves as the default recommender for all POPROX subscribers and as the Control condition in our experiment.

4.2. Condition 2: MMR Diversification

Condition 2 builds onto the baseline recommender a Maximum Marginal Relevance (MMR) [8] re-ranking step to promote intra-list diversity. Candidate articles are first scored using the same fused relevance function defined in the baseline: $s(u, a_i)$. Let $\mathcal{C}$ denote the filtered candidate set. For each article $a_i \in \mathcal{C}$, we retain the final relevance score $s(u, a_i)$ and the article embedding $\mathbf{a}_i$.

Pairwise article similarity is computed using dot-product similarity in embedding space: $\text{sim}(a_i, a_j) = \sigma(\mathbf{a}_i^\top \mathbf{a}_j)$ where $\sigma(\cdot)$ denotes the sigmoid function. This produces a bounded similarity matrix: $\mathbf{S} = \sigma(\mathbf{A}\mathbf{A}^\top)$. Let $\mathcal{S}$ denote the set of already selected articles and $\mathcal{R}$ the remaining candidates. Articles $a_i \in \mathcal{R}$ are selected sequentially using the MMR criterion:

$$\text{MMR}(S, a_i) = \theta \cdot s(u, a_i) - (1-\theta) \cdot \max_{a_j \in \mathcal{S}} \text{sim}(a_i, a_j) \tag{1}$$

where $\theta \in [0, 1]$ controls the trade-off between relevance and diversity. In our experiments, we set $\theta = 0.6$, with the result that relevance receives 60% of the weight in the objective.

The final list ($K = 10$ articles) is built greedily: the highest-relevance article initializes $\mathcal{S}$, each subsequent position selects $a^* = \arg\max_{a_i \in \mathcal{R}} \text{MMR}(\mathcal{S}, a_i)$, removing a^* from $\mathcal{R}$ until K articles are selected.

4.3. Condition 3: Personalized MMR Re-ranking (MMR-P)

Condition 3 builds on the Maximal Marginal Relevance (MMR) algorithm from Condition 2 but adapts the relevance/diversity trade-off parameter individually for each user based on their topic preference distribution inspired by the approach used in [18].

As in Condition 2, we apply greedy sublist optimization. Instead of the MMR calculation in Equation 1, we apply a user-specific scoring calculation that takes into account the user's interest in diverse news.

We estimate a user-specific diversity tolerance parameter β based on the divergence between the user's topic preference distribution and a uniform distribution. Let $\mathbf{p}_u$ denote the normalized topic preference distribution for user u, and $\mathbf{q}$ the uniform topic availability distribution. We compute

$\beta_{\text{raw}} = \text{JS}(\mathbf{p}_u \parallel \mathbf{q})$, where $\text{JS}(\cdot)$ denotes the Jensen-Shannon divergence. The raw divergence is standardized using empirical mean μ and standard deviation σ estimated from the user population. Higher values of β indicate stronger concentration of preferences (i.e., lower tolerance for diversity).

For each recommendation position k, the selected item maximizes MMR-P$(i) = \theta_u s_i - (1 - \theta_u) \max_{j \in S} \text{sim}(i, j)$, where θ_u, a personalized trade-off parameter, replaces the original fixed θ from MMR. The user-specific parameter is defined as: $\theta_u = \text{clip}\left(\theta_0 \left(1 + \lambda\beta\right), 0, 1\right)$, where θ_0 is the base relevance weight (set to 0.6 as in Condition 2), and λ controls the strength of personalization (set to 0.5). The clipping ensures $\theta_u \in [0, 1]$. This treatment extends MMR by personalizing the relevance–diversity trade-off based on the user's topic preference distribution.

4.4. Offline Evaluation and Tuning

POPROX provides an integrated offline evaluation framework, which we used to tune θ and β prior to online deployment. Given a recommended list $R = (r_1, \ldots, r_k)$ and graded relevance signals, POPROX computes a suite of complementary metrics. Ranking quality is measured using Normalized Discounted Cumulative Gain (NDCG) and Mean Reciprocal Rank (MRR). We found that MRR was a complementary signal to NDCG in assessing early-precision behavior of ranking functions, helping us identify how the value of θ impacted the position of the first relevant item. List overlap is measured via Rank-Biased Overlap (RBO) [19]. Topical diversity is captured by Rank-Biased Entropy (RBE) which quantifies topical dispersion in the top-k by forming geometric rank weights: $w_i = (1 - d)d^{i-1}, \quad i = 1, \ldots, k$, summing these weights across articles to produce normalized topic probabilities q_t, and returning: RBE $= -\sum_t q_t \log_2 q_t$. Higher RBE indicates broader topical coverage; the implementation counts only mentions with `entity_type` "subject" or "topic". Finally, Intra-List Similarity (ILS) [20] measures the average pairwise cosine similarity between item embeddings. Our figures and tables report $ILD = 1 - ILS$, which measures diversity of the presented list (independent of click behavior).

1. **MMR** Candidate values of $\theta \in [0, 1]$ were evaluated offline across all metrics; $\theta = 0.6$ was selected as the best balance between relevance and diversity.
2. **MMR-P** We analyzed the empirical distribution of β across users and calibrated its integration to produce a stable, well-separated range of personalization effects while keeping mean ILD roughly equivalent to MMR, ensuring total diversity level would not confound the comparison between treatments.

4.5. Evaluation Protocol

As part of a POPROX experiment, the platform collects:

1. **Newsletter delivery logs** articles sent, position, condition.
2. **Click/impression logs** which articles a user clicked, and when
3. **User interest logs** per-topic preference scores (1-5) with timestamps
4. **Article metadata** headline, topics, publish time, candidate pool
5. **Account-to-treatment mapping** condition assigned per user.

We use this data to evaluate engagement (CTR, participation rate), content diversity (ILD, Gini coefficients), and ranking overlap (RBO@10) across the three conditions.

5. Results

Note that the IRB protocol governing the POPROX platform prevents subscribers' demographic information from being shared with experimenters. Our analysis is therefore conducted without user attributes such as age, gender, or geographic location. The raw interaction data cannot be released under the terms of the POPROX researcher agreement[3] .

[3]For transparency, our results analysis code is public https://github.com/that-recsys-lab/poprox-rec-diverse-analysis

After the experiment was complete, we applied a set of data quality filters to mitigate noise and structural inconsistencies in the collected data. Two user accounts were excluded from the analysis exhibiting click volumes 29× and 18× above the sample median, likely reflecting automated or anomalous behavior. We removed instances where the code in one of the treatments failed due to transient errors in one of the other treatments.

Metric	Ctrl	MMR	MMR-P
Unique Users	88	85	84
Newsletters Delivered	3085	3029	2926
Total Clicks	123	72	119
Active Users (>0 clicks)	21	19	21
Clicked Newsletters	96	58	100
Engagement Metrics			
Newsletter PR (%)	3.11	1.91	3.42
User PR (%)	23.86	22.35	25.00
Avg. Clicks per Newsletter	3.99	2.38	4.07
Avg. CTR (%)	0.44	0.33	0.39
Avg. CTR Active Users (%)	1.86	1.46	1.58
Avg. Daily CTR (%)	0.40	0.24	0.41
CTR Std Dev	0.23	0.19	0.23

(a) Engagement metrics

Metric	Ctrl	MMR	MMR-P
Diversity & Concentration			
ILD	0.74	0.76	0.94
Article Gini	0.51	0.51	0.55
Entity Gini	0.51	0.52	0.62
Clicked Article Gini	0.024	0.027	0.062
Personalization Metrics			
Individual RBO@10	–	0.12	0.07
System RBO@10 (vs. Ctrl)	–	0.73	0.04

(b) Diversity and personalization metrics

Figure 2: Summary statistics for experimental conditions

After this data cleaning, we evaluated the three recommendation treatments—Control ($n = 88$), non-personalized diversification (MMR, $n = 85$), and personalized diversification (MMR-P, $n = 84$), with approximately 3K newsletters per condition. Our analysis focused on participation, click-through behavior, content diversity, and ranking overlap.

Contrary to the "diversity tax" expectation, descriptive trends suggest that personalized diversification (MMR-P) achieves the strongest participation and click performance, while still exhibiting the highest diversity across conditions. Non-personalized diversification (MMR) shows lower engagement despite modest diversity gains. Figure 2a summarizes the results. Clicks are sparse, so we focus here on active users, defined as those who clicked at least once during the experiment. Because of the sparsity of clicks and the relatively small number of active users, we do not find any trends reaching statistical significance[4]. We believe that these trends are suggestive of the value of personalized diversity, as established in prior offline work, but additional experimentation will be needed to establish this definitively.

MMR-P shows the strongest participation across both measures. The newsletter participation rate (PR), defined as the percentage of delivered newsletters that receive at least one click, reaches 3.42% under MMR-P vs. 3.11% (Control) and 1.91% (MMR). This shows that a larger share of newsletters elicit engagement under personalized diversification. Similarly, the user PR, defined as the proportion of users who generate at least one click during the experiment, follows the same pattern at 25%, 23.86%, and 22.35% respectively. Together, these results show that MMR-P increases both the share of engaging newsletters and matches Control in active user count, while outperforming MMR on both measures, suggesting that diversification without personalization may reduce engagement. Overall, MMR-P changes not only ranking quality but also what users choose to interact with.

The ECDF plots in Figure 3a compare active-user participation in the three conditions. We see that MMR-P has fewer lower-clicking users, although MMR has a few more high-clicking users. This is encouraging because it reduces the risk that results are purely driven by a lucky split where MMR-

[4]Ekstrand et al., "Monte Carlo Power Analysis for Small-System A/B Trials: Detecting Differences in Newsletter Engagement," to appear.

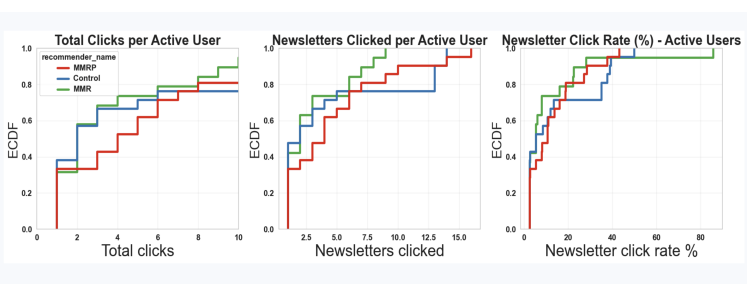
P received more naturally active users. In total clicks, MMR-P shows the strongest rightward shift, suggesting more active clicking per user. For clicked newsletters, MMR-P is also dominant overall, while MMR appears to top out around 9.

MMR-P records the highest clicks per 100 delivered newsletters (4.07, vs. 3.99 for Control and 2.38 for MMR). Since this exceeds the newsletter participation rate (3.42%), some newsletters receive more than one click. But this within-newsletter bundling effect is actually weaker for MMR-P (~1.19 clicks per engaged newsletter) than for Control (~1.28), suggesting MMR-P's advantage comes from reaching more distinct newsletters rather than deeper engagement within any one newsletter.

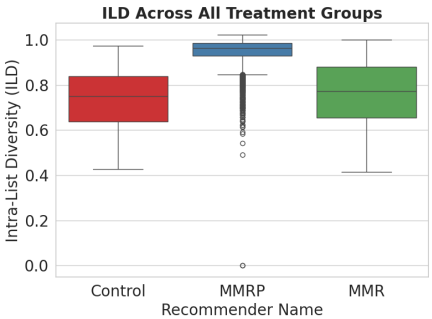
Control records the highest average CTR at 0.44%, while MMR-P remains close at 0.39% and outperforms MMR (0.33%). MMR-P further delivers the highest average daily CTR at 0.41%, compared to 0.40% for Control and 0.24% for MMR. For active users, Control shows the highest CTR (1.86%), followed by MMR-P (1.58%) and MMR (1.46%). Pairwise Mann–Whitney U tests found no statistically significant differences in per-user CTR across arms (Control vs. MMR: U=3812, p=.77; Control vs. MMR-P: U=3649, p=.85; MMR vs. MMR-P: U=3446.5, p=.60; all $|r|<0.04$), consistent with a study of this scale (n=84–88 each) rather than an absence of effect.

In terms of diversity and content distribution, MMR-P shows the strongest performance in intra-list diversity (ILD = 1-ILS). ILD is highest for MMR-P at 0.94, compared to 0.74 for Control and 0.76 for MMR, indicating that recommendations within each newsletter are substantially more semantically diverse under MMR-P. However, this increase in embedding-level diversity coexists with higher concentration in exposure at the topic and item level. The article Gini (0.55) and entity Gini (0.62), computed over topic-level distributions, are both higher than Control and MMR, suggesting that impressions are more unevenly distributed across topics and entities, with a smaller set receiving a larger share of exposure. This pattern is also reflected in the clicked article Gini (0.062), which is higher under MMR-P than under the other treatments, indicating that while individual lists are more diverse, attention is focused on a smaller subset of items overall.

The RBO results further highlight a clear distinction between individual-level similarity and system-level RBO. At the individual level, MMR shows higher overlap with Control (RBO = 0.12) than MMR-P (0.07), indicating that MMR's recommendations remain closer to the baseline on a per-user, per-timestamp basis, while MMR-P introduces greater deviation and exploration in individual lists. System-level RBO reveals a different pattern: MMR's exposure ranking remains highly similar to Control (RBO@10 = 0.73), indicating that although individual users receive different specific lists, MMR still surfaces largely the same set of popular articles in aggregate. MMR-P, by contrast, shows low overlap with Control at both the individual (0.07) and system level (0.04), indicating that personalization reshapes not only which articles individual users see but also the overall distribution of exposure across the article pool. This is consistent with the higher concentration (Gini) reported above.



(a) Empirical Cumulative Distribution plots of activity statistics per active user: total clicks, newsletters clicked, and newsletter click rate.



(b) Distribution of intra-list diversity (ILD) across the treatments.

Figure 3: Engagement and Diversity distributions

6. Discussion

For RQ1, MMR-P subscribers show higher participation rates than MMR and descriptively comparable engagement to Control, suggesting that readers respond more positively to personalized diversity than to uniform diversification. For RQ3, these results support the view that personalization helps manage the engagement-diversity tradeoff, i.e., MMR-P maintains CTR close to Control while achieving substantially higher ILD. RQ2 remains difficult to answer directly: click sparsity prevented a robust comparison of engagement patterns between high- and low-beta users, and we leave this for future work with larger samples.

The findings highlight both the potential and the constraints of combining personalization with diversification. The sparse clicks in our dataset limit statistical power, making it difficult to draw strong conclusions about true differences in engagement among conditions.

A further limitation arises from the dependency on the daily news catalog. Since article exposure, and therefore topical coverage, is determined by the content from the AP feed on a given day, the balance between personalization and diversity is constrained by supply. This makes it difficult to disentangle whether observed effects are driven by the recommendation strategy or by variations in the underlying content pool. We note that due to the needs of the POPROX experiment scheduling process (not under our control), the experiment ended up being conducted over December–January, a period that introduced distinct seasonal variation in news content including year-end summary coverage that may have affected both user engagement patterns and the calibration of our diversity condition.

The experiment is conducted in a fully naturalistic setting: users are neither compensated nor aware of the conditions[5]. Engagement is therefore entirely voluntary, limited to observable clicks. We cannot capture passive consumption of the newsletter that does not produce clicks, such as headline scanning. This introduces measurement bias that likely underestimates true engagement and impact.

One place where we see the impact of the naturalistic setting is in the failure of our diversity calibration. Recall that our θ_0 distribution in Condition 3 was tuned to have roughly the same mean ILD as the fixed θ in Condition 2. Figure 3b shows that this is clearly not the case during the experiment period. Condition 3 had higher average diversity although its range is larger, as expected. We believe that this is due to differences in the topic distributions found in the news cycles of our tuning data (Summer 2025) with our experimental period some months later. This suggests that a strategy aimed at a fixed degree of diversification may need to adapt to the particulars of the news environment. Importantly, because MMR-P ended up with higher mean ILD than MMR, the engagement advantage of MMR-P cannot be attributed solely to personalization, some portion may reflect the higher diversity level itself, making the two treatments harder to compare cleanly.

A key limitation is the absence of pre-condition behavioral data: without prior click patterns, we cannot distinguish historically inactive users from those disengaged by the intervention, making it difficult to attribute engagement changes directly to the conditions.

Additionally, we do not have access to the full ranked list (e.g., top 100 items) after the baseline model, which limits our ability to analyze how far reranking shifts items. As a result, we cannot compute reranking metrics such as Least Item Promoted [21] in the different conditions.

7. Conclusion

We study the role of diversification in a naturalistic news recommendation setting, comparing a baseline system with non-personalized and personalized diversification strategies. Our results show that increasing diversity meaningfully changes the structure of recommendations and user interaction patterns, but can reduce engagement when applied without regard to user preferences. Incorporating personalization helps mitigate this drop, improving participation and click volume compared to uniform diversity while preserving high levels of intra-list diversity.

[5]Users see an encoded treatment ID at the bottom of each newsletter but do not know the specifics of any experimental conditions.

These findings suggest that diversity should be treated as one of the primary objectives in news recommendation, given its role in supporting broader information exposure. At the same time, personalization can serve as an effective mechanism to maintain user engagement, enabling systems to deliver diverse content without substantially disengaging users. Importantly, the observed trade-offs and limitations highlight the need for more calibrated approaches that balance diversity, relevance, and user behavior in dynamic, real-world settings.

Personalization introduces a fundamental tension in the delivery of diverse news. If we view diversity as a user-oriented objective, calibrating news delivery to user preferences, then personalized diversification is exactly what is needed and our experiments support the idea that this approach is beneficial. On the other hand, diversity is not uniformly delivered but unevenly distributed: some users receive highly diverse content, while others are exposed to even more concentrated feeds. If we view diversity as an essential journalistic value, then the personalized re-ranking implemented here risks sidelining this value for some users. This raises a fundamental normative question for real-world news recommender systems: should personalization determine who is exposed to diverse news, or should diversity be guaranteed as a journalistic obligation central to the democratic role of news? Our study, demonstrating the limits of generic diversification, makes this question all the more pointed.

Declaration on Generative AI

The authors used Claude Opus 5.5 (Anthropic) for grammar and spelling checks, formatting assistance, and to help fix a code bug during the experiment. The authors reviewed all outputs and take full responsibility for the publication's content.

References

[1] K. Karppinen, Journalism, pluralism, and diversity, Journalism 19 (2018).

[2] L. A. Møller, Recommended for you: How newspapers normalise algorithmic news recommendation to fit their gatekeeping role, Journalism Studies 23 (2022) 800–817.

[3] N. Helberger, On the democratic role of news recommenders, in: Algorithms, automation, and news, Routledge, 2021, pp. 14–33.

[4] 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, WWW '05, Association for Computing Machinery, New York, NY, USA, 2005, p. 22–32. URL: https://doi.org/10.1145/1060745.1060754. doi:10.1145/1060745.1060754.

[5] S. Vrijenhoek, M. Kaya, N. Metoui, J. Möller, D. Odijk, N. Helberger, Recommenders with a mission: assessing diversity in news recommendations, in: Proceedings of the 2021 conference on human information interaction and retrieval, 2021, pp. 173–183.

[6] S. Vrijenhoek, G. Bénédict, M. Gutierrez Granada, D. Odijk, M. De Rijke, Radio–rank-aware divergence metrics to measure normative diversity in news recommendations, in: Proceedings of the 16th ACM conference on recommender systems, 2022, pp. 208–219.

[7] N. Tintarev, E. Sullivan, D. Guldin, S. Qiu, D. Odjik, Same, same, but different: algorithmic diversification of viewpoints in news, in: Adjunct publication of the 26th conference on user modeling, adaptation and personalization, 2018, pp. 7–13.

[8] J. Carbonell, J. Goldstein, 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, 1998, pp. 335–336.

[9] F. Eskandanian, B. Mobasher, R. Burke, A clustering approach for personalizing diversity in collaborative recommender systems, in: Proceedings of the 25th conference on user modeling, adaptation and personalization, 2017, pp. 280–284.

[10] W. Liu, R. Burke, Personalizing fairness-aware re-ranking, arXiv preprint arXiv:1809.02921 (2018).

[11] C. Bauer, C. Bagchi, O. A. Hundogan, K. Van Es, Where are the values? a systematic literature review on news recommender systems, ACM Transactions on Recommender Systems 2 (2024) 1–40.

[12] L. Heitz, J. A. Lischka, A. Birrer, B. Paudel, S. Tolmeijer, L. Laugwitz, A. Bernstein, Benefits of diverse news recommendations for democracy: A user study, Digital Journalism 10 (2022) 1710–1730.

[13] L. Heitz, J. A. Lischka, R. Abdullah, L. Laugwitz, H. Meyer, A. Bernstein, Deliberative diversity for news recommendations: Operationalization and experimental user study, in: Proceedings of the 17th ACM Conference on Recommender Systems, 2023, pp. 813–819.

[14] L. Basso, I. Nalis, J. Neidhardt, News diversity and well-being–an experimental exploration of diversity-aware recommender systems, 2023.

[15] S. Vercoutere, G. Joris, T. De Pessemier, L. Martens, Improving consumption diversity via graph-based topic nudging, User Modeling and User-Adapted Interaction 35 (2025) 8.

[16] 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.

[17] 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.

[18] W. Liu, J. Guo, N. Sonboli, R. Burke, S. Zhang, Personalized fairness-aware re-ranking for microlending, in: Proceedings of the 13th ACM conference on recommender systems, 2019, pp. 467–471.

[19] W. Webber, A. Moffat, J. Zobel, A similarity measure for indefinite rankings, ACM Transactions on Information Systems (TOIS) 28 (2010) 1–38.

[20] 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.

[21] A. Aird, E. Štefancová, A. Buhayh, C. All, M. Homola, N. Mattei, R. Burke, Integrating individual and group fairness for recommender systems through social choice, in: Proceedings of the Nineteenth ACM Conference on Recommender Systems, 2025, pp. 177–186.