Multidisciplinary Research Journal
ISSN (online): 3108 – 1533

Volume 2, Issue 2, 2026
April – June 2026

Research Article

Comprehensive Comparative Analysis of the Personalization Diversity Trade-off in News Recommendation Systems: Deep Learning Model Evaluation on MIND Dataset with Hybrid Framework Optimization

Dhanjit Kalita [1,*], Tribikram Pradhan [1], Dhritiman Saikia [1]

[1] Department of Computer Science, Tezpur University, Assam-784028
* Correspondence: csp24101@tezu.ac.in

Received: 8 July 2026; Accepted: 7 October 2026 ; Published: 15 October 2026

Abstract: News Recommendation Systems (NRS) face a fundamental challenge: optimizing for personalization accuracy inevitably exacerbates filter bubbles by reducing content diversity. This study presents a comparative evaluation of five models on the Microsoft News Dataset (MIND) containing 1 million users and 24.2 million interaction logs. We evaluate Content-Based Filtering (CBF), Collaborative Filtering (CF), Deep Neural Networks (DNN), Recurrent Neural Networks (RNN) and Attention-Based Neural Networks (ABNN) using twelve metrics spanning personalization accuracy, diversity and system-level properties. Results demonstrate that ABNN achieves superior personalization (Precision@10 = 0.401) but severely reduces diversity (ILD = 0.387), a 47.6% decrease from CBF's diversity baseline (ILD = 0.738). We propose a novel hybrid framework incorporating diversity-aware loss functions and multi-objective optimization that achieves 91% of maximum personalization while improving diversity by 45.2% at less than 5% computational overhead. Our findings indicate that practitioners should deliberately occupy specific Pareto-optimal positions based on use-case requirements and that filter bubble effects disproportionately harm specialist users while intensifying over time.

Keywords: News Recommendation Systems; Personalization-Diversity Trade-off; Deep Learning; MIND Dataset; Filter Bubbles; Hybrid Recommender Systems; Multi-Objective Optimization

1. Introduction

1.1. Motivation and Problem Statement

The digital news landscape has undergone a profound transformation over the past two decades. Whereas traditional media operated under editorial gatekeeping models, today's online platforms generate exponentially growing volumes of content at an unprecedented scale. By some estimates, as many as 500 million news articles are now published daily [1], creating an information overload that makes manual discovery impractical for users. At the same time, user attention remains finite, with average daily news consumption hovering around 15–20 minutes [1]. This fundamental mismatch between content supply and user attention capacity has created an urgent need for automated content curation systems.

News Recommendation Systems (NRS) have consequently become essential infrastructure for major platforms, including Microsoft News, Google News [2] and Apple

News, because they serve several critical functions: (1) improving user experience by reducing search costs; (2) increasing platform engagement; and (3) connecting users with relevant content that they would not otherwise discover on their own.

However, accumulating evidence reveals a troubling paradox: while personalization optimizations improve short-term accuracy, they inadvertently create what Pariser [3] termed filter bubbles, a phenomenon in which users are progressively isolated within self-reinforcing content ecosystems. Sunstein [4] argues that filter bubbles contribute to political polarization, reduced social cohesion and diminished information literacy. These broader societal claims are drawn from the cited literature rather than measured directly in this study; our own contribution is confined to the computational trade-off between personalization accuracy and content diversity, and we do not independently verify downstream effects such as polarization or reduced social cohesion.

Despite this body of work, prior studies typically report a single trade-off point per model or dataset without systematically quantifying how the diversity cost of a given personalization gain changes across model classes, user segments and elapsed time. This leaves an open research gap: a model-agnostic, quantitative account of when, and for which users, personalization-driven diversity loss is most severe, and how much of it can be recovered without materially sacrificing accuracy. The present study addresses this gap by evaluating five representative model families under a common protocol and by proposing a hybrid framework whose personalization-diversity trade-off can be tuned explicitly. We focus on the MIND dataset because it is among the largest publicly available English-language news recommendation benchmarks with real user click logs (1,000,000 users, 24.2 million interactions), providing enough users per segment to support the disaggregated segment-level and temporal analyses in Sections 6.4 and 6.5; smaller or synthetic benchmarks do not offer comparable statistical power for this purpose. We acknowledge that single-dataset evaluation limits generalizability, and we return to this point, together with a discussion of additional benchmarks for future validation, in Section 7.3.

1.2. Research Contributions

This paper makes the following primary contributions:

- A rigorous comparative evaluation of five recommendation models using 12 metrics spanning accuracy, diversity and system-level properties on the full MIND dataset (1M users, 24.2 M interactions).
- Mathematical quantification of the personalization-diversity trade-off gradient, revealing diversity costs that steepen with model complexity from −1.65 (CBF→CF) to −3.78 (RNN→ABNN) per unit precision gain.
- A user-segment and temporal analysis demonstrating that specialist users (51% of the population) experience a markedly larger diversity loss than generalists, and that filter bubble effects intensify by 8.2% over six weeks.
- A novel hybrid multi-objective framework achieving 91.0% of maximum personalization with 45.2% diversity improvement at less than 5% runtime overhead.

2. Related Work

Recommender systems have evolved from collaborative filtering [5] and content-based models towards deep learning approaches. Early collaborative filtering work by Koren et al. [6] demonstrated that matrix factorization captures latent user-item interaction patterns, although it suffers from cold-start and sparsity problems. Content-based methods, by contrast, offer greater interpretability and handle new items more effectively.

Modern deep learning architectures significantly improve accuracy by modelling temporal and non-linear relationships [7]. Covington et al. [8] demonstrated the power of deep neural networks for large-scale recommendation at YouTube. Hidasi et al. [9] introduced recurrent neural networks for session-based recommendation, capturing sequential behavioural patterns. The transformer architecture with multi-head self-attention [10] enabled even more expressive user-behaviour modelling, allowing the model to selectively attend to the most relevant portions of a user's interaction history when evaluating candidate items.

However, excessive personalization has raised concerns about filter bubbles [3]. Ziegler et al. [11] introduced intra-list diversity and demonstrated that accuracy-maximizing recommenders systematically reduce topic variety. Ge et al. [12] formalized serendipity and coverage as key evaluation dimensions. More recent multi-objective optimization approaches [13] aim to balance accuracy with diversity, fairness and long-term user satisfaction, particularly in news recommendation systems where information diversity carries broader societal implications. Despite these advances, the precise quantification of trade-off costs across model classes, and the dependence of filter bubble severity on user type and elapsed time, have received limited systematic study at the scale we conduct here.

3. Materials and Methods

3.1. Dataset: MIND

The Microsoft News Dataset (MIND) [14] was selected for its scale and representativeness of real-world news consumption patterns. Table 1 summarizes the key dataset specifications.

Table 1. Comprehensive MIND Dataset Specifications

Component	Count	Details	Notes
News articles	161,013	English sources	Diverse topics
User interactions	24.2 M	6-week period	Click/no-click logs
Unique users	1,000,000	Varied demographics	Implicit feedback
Time period	42 days	Oct–Nov 2019	Consistent platform state
Categories	18 primary	Hierarchical structure	With subcategories
Article title	2–10 words	Provided	For content embedding
Article abstract	100–300 chars	Variable length	For content embedding
User behaviors	Binary	Clicks & impressions	Item exposure tracked

3.2. Model Implementations

We implement and evaluate five recommendation models spanning classical and deep learning paradigms.

3.2.1. Model 1: Content-Based Filtering (CBF)

The user profile is constructed as the weighted average of clicked article embeddings with the recommendation score computed via cosine similarity. This approach follows the classical content-based recommendation paradigm [15], which offers strong interpretability and requires no interaction history for new items, at the cost of limited ability to capture patterns that go beyond an individual user's previously expressed interests.

$$\text{Score}_{\text{CBF}}(u,j) = \frac{\left(\frac{\sum_{i\in C_u} w_i \overrightarrow{v_i}}{\sum_{i\in C_u} w_i}\right)\cdot \overrightarrow{v_j}}{\left|\frac{\sum_{i\in C_u} w_i \overrightarrow{v_i}}{\sum_{i\in C_u} w_i}\right| \left|\overrightarrow{v_j}\right|}$$

3.2.2. Model 2: Collaborative Filtering (CF)

Matrix factorization predicts interactions using learned latent factors. We follow the formulation of Koren et al. [6], learning user and item latent factors jointly by minimizing regularized squared reconstruction error over observed interactions via stochastic gradient descent.

$$\widehat{r_{u,i}} = \overrightarrow{u_u}\cdot \overrightarrow{v_i} + b_u + b_i$$

3.2.3. Model 3: Deep Neural Network (DNN)

Non-linear feature interactions are modelled with a 3-layer MLP, following the deep ranking architecture popularized for large-scale recommendation by Covington et al. [8]: user and candidate-item embeddings are concatenated and passed through fully connected layers with ReLU activations to produce a relevance score.

$$\widehat{r_{u,i}} = \sigma\left(W_3 \text{ReLU}\left(W_2 \text{ReLU}(W_1 \vec{x})\right)\right)$$

3.2.4. Model 4: Recurrent Neural Network (RNN)

Sequential user behaviour is captured via an LSTM, with the recommendation score derived from the final hidden state, building on the session-based recommendation approach of Hidasi et al. [9]; the click sequence is encoded left-to-right and the final hidden state is combined with the candidate item embedding to produce a ranking score.

$$h_T = \text{LSTM}(x_1, \ldots, x_T), \qquad \widehat{r_{u,i}} = \sigma(W[h_T; v_i] + b)$$

3.2.5. Model 5: Attention-Based Neural Network (ABNN)

Multi-head self-attention dynamically weights historical interactions by their relevance to the candidate item: r̂(u,i) = σ(MultiHead(Q, K, V)) where queries Q are derived from the candidate item and keys K and values V from the user's interaction history. We use 8 attention heads with dimension 64, adapting the multi-head attention mechanism of Vaswani et al. [10]

to user-history modelling in the manner explored for recommendation by subsequent work [16].

3.3. Proposed Hybrid Framework

3.3.1. Multi-Objective Optimization Formulation

We propose a hybrid scoring function balancing personalization and diversity. We adopt a linear re-ranking scheme [17], rather than retraining a separate model per operating point, because it lets practitioners move continuously along the personalization-diversity curve at inference time by adjusting a single interpretable parameter, without incurring the cost of maintaining multiple trained models; Section 5.3 compares this hybrid configuration against all five baseline models, not only against pure ABNN.

$$\text{Score}_{\text{hybrid}}(u,i) = \alpha \cdot \text{Score}_{\text{personal}}(u,i) + \beta \cdot \text{Score}_{\text{diversity}}(i,R)$$

where $\alpha + \beta = 1, \alpha, \beta \geq 0$; Score_personal is the ABNN score normalized to [0,1]; and the diversity score measures candidate item i's dissimilarity to the already-recommended set R:

$$\text{Score}_{\text{diversity}}(i,R) = \frac{1}{|R|}\sum_{j \in R}\big(1 - \text{Sim}(i,j)\big)$$

using cosine distance on pre-trained article embeddings.

3.3.2. Diversity-Aware Loss Function

During training, a diversity regularization term augments the standard ranking loss: $L_{total} = L_{ranking} + \lambda \cdot L_{diversity}$, where the ranking loss uses binary cross-entropy and the diversity loss penalizes similar hidden representations across users in the same mini-batch. The negative sign inverts the similarity, so that gradient descent minimizes intra-batch user-representation similarity, encouraging the model to learn diverse user representations. Empirically, $\lambda = 0.05$provided the best balance, whereas values outside [0.01, 0.1] yielded suboptimal results.

3.3.3. Implementation Details

Training procedure:

i. Initialize ABNN from pre-trained weights on the full ranking task.
ii. Per epoch: (a) sample mini-batches of 256 pairs with 4:1 negative ratio; (b) forward pass computing scores and representations; (c) compute $L_{ranking}$via binary cross-entropy; (d) compute $L_{diversity}$over batch; (e) combine: $L_{total} = L_{ranking} + 0.05 \cdot L_{diversity}$; (f) backpropagate with $\eta = 0.001 \times 0.95^{\lfloor \text{epoch}/5 \rfloor}$
iii. Early stopping on validation nDCG with patience = 3.

Inference-time re-ranking adds $O(|R|)$per candidate. For $|R| = 10$ and 1,000 candidates, this amounts to <5% overhead relative to pure ABNN, which is feasible for production use and comparable to deployed re-ranking strategies at Netflix, YouTube and Spotify.

3.3.4. Hyperparameter Optimization and Reproducibility

Hyperparameters for all five baseline models (embedding dimension, number of layers, learning rate and batch size, and, for ABNN, number of attention heads and head dimension) were selected via grid search on a held-out validation split comprising 10% of users, disjoint from the reported test set, using nDCG@10 as the selection criterion. For the hybrid framework, α and β were selected via grid search over the nine weighting schemes reported in Table 7, and the diversity-loss weight λ was tuned over {0.01, 0.03, 0.05, 0.07, 0.1} on the same validation split; $\lambda = 0.05$was retained because it gave the best validation-set nDCG-ILD balance, consistent with the working range reported in Section 3.3.2.

- Optimizer: Adam, initial learning rate $\eta = 0.001$with decay 0.95 every 5 epochs.
- Training budget: maximum 20 epochs per model with early stopping (patience = 3) on validation nDCG.
- Negative sampling ratio: 4 negative impressions per positive click, mini-batches of 256 pairs.
- Embedding dimension: 128 for all models; ABNN uses 8 attention heads of dimension 64.
- Robustness: all reported results are averaged over three independent runs with random seeds 42, 123 and 2024; standard deviations were below 0.6% of the reported mean for every metric.
- Hardware/software: single NVIDIA A100 GPU (40GB), PyTorch 2.1, CUDA 12.1.

4. Evaluation Metrics

We employ twelve metrics organized into three categories to provide a comprehensive assessment of the personalization-diversity trade-off.

4.1. Personalization and Accuracy Metrics

Precision@K measures the fraction of top-K recommended items that are relevant: $\text{Precision@K} = \frac{|R_u^K \cap C_u|}{K}$

Recall@K measures the fraction of all relevant items retrieved within the top-K: $\text{Recall@K} = \frac{|R_u^K \cap C_u|}{|C_u|}$

Normalized Discounted Cumulative Gain (nDCG@K) accounts for the ranking position of relevant items, assigning higher scores to relevant items appearing higher in the list.

4.2. Diversity Metrics

Intra-List Diversity (ILD) measures the average pairwise dissimilarity within the recommendation list [7]: $\text{ILD} = \frac{2}{K(K-1)} \sum_{i<j} \big(1 - \text{Sim}(i,j)\big)$

Category Coverage (CC) measures topic breadth across the 18 MIND categories: $\text{CC} = \frac{|\text{Categories in } R_u^K|}{|\text{All Categories}|}$

Personalization Index (PI) captures cross-user recommendation variability: $\text{PI} = 1 - \frac{1}{M}\sum\left(\frac{n_i}{N}\right)^2$, where n_i is the number of users recommended item i, N is the total number of recommendations and M normalizes by user count.

4.3. System-Level Metrics

Coverage measures the fraction of the item catalog ever recommended to any user [12]: $\text{Coverage} = \frac{|\text{Items recommended}|}{|\text{Total items}|}$

Novelty (entropy-based) measures the information content of the recommendation probability distribution, penalizing concentration on popular items: $\text{Novelty} = -\sum p_i \log_2(p_i)$

Serendipity measures the fraction of recommendations that are simultaneously relevant and surprising to the user.

5. Experimental Results

5.1. Overall Model Comparison

Table 2 presents comprehensive evaluation results at K=10 across all five models. Results are averaged over the full test set (n=1,000,000 users, 161,013 items).

Table 2. Comprehensive Model Evaluation Results (K=10, n=1,000,000 users). Bold = best per metric.

Model	P@10	R@10	nDCG	ILD	CC	PI	Cov	Nov	Ser
CBF	0.286	0.241	0.294	**0.738**	**0.811**	**0.712**	0.287	**4.82**	0.156
CF	0.312	0.267	0.321	0.695	0.723	0.654	0.412	4.65	0.212
DNN	0.358	0.301	0.362	0.548	0.584	0.582	0.534	4.21	**0.287**
RNN	0.392	0.334	0.401	0.421	0.412	0.485	0.623	3.87	0.198
ABNN	**0.401**	**0.342**	**0.415**	0.387	0.356	0.451	**0.681**	3.54	0.164

CC = Category Coverage, PI = Personalization Index, Cov = Coverage, Nov = Novelty, Ser = Serendipity

CBF Analysis: CBF achieves the strongest diversity: ILD = 0.738, CC = 0.811 (14.6 of 18 categories), Novelty = 4.82. However, accuracy is modest: P@10 = 0.286, nDCG = 0.294. Coverage of 0.287 means only 46,200 of 161,013 items are ever recommended. CBF is therefore most appropriate for cold-start scenarios.

CF Analysis: CF shows moderate performance across all categories: P@10 = 0.312, ILD = 0.695 and Coverage = 0.412 (66,300 items). Its PI of 0.654 reflects diversity arising from latent factors rather than from deliberate diversity design. The main weaknesses of CF are the cold-start problem and its inability to weight recent interactions.

DNN Analysis: DNN occupies the Pareto-balanced sweet spot, with P@10 = 0.358 (+25.2% relative to CBF), ILD = 0.548 and Serendipity = 0.287 (the highest among all models). Its coverage is 0.534 (86,000 items). The 25.7% decrease in ILD relative to CBF reflects an emerging filter bubble effect, while the peak in serendipity indicates that DNN can recommend items that are not only relevant but also genuinely novel.

RNN Analysis: RNN achieves P@10 = 0.392 (+9.5% relative to DNN) at the cost of reduced diversity: ILD = 0.421 (−23.2%) and CC = 0.412 (only 7.4 categories). The evolution of the LSTM hidden state shifts recommendations towards the user's historical norm; for

example, a user whose clicks are 80% on entertainment topics will receive predominantly entertainment recommendations, losing exposure to about 10 of the 18 available categories.

ABNN Analysis: ABNN provides the most personalized recommendations, with P@10 = 0.401 and nDCG = 0.415; however, its diversity is lower than that of all other models: ILD = 0.387, CC = 0.356 (only 6.4 categories) and Serendipity = 0.164. Despite the highest coverage (0.681, approximately 110,000 items), the novelty of its recommendations lags far behind, at only 3.54. This reveals a coverage-novelty paradox: ABNN draws on a large number of distinct items across users, yet the items recommended to each individual remain highly similar and predominantly popular.

5.2. Relative Performance Changes

Table 3 quantifies percentage changes relative to CBF, revealing systematic degradation of diversity as model sophistication increases.

Table 3. Relative Performance Changes (%) from CBF Baseline

Model	P@10	ILD	CC	Nov	Cov	Ser
CF	+9.1	−5.8	−10.9	−3.5	+43.6	+35.9
DNN	+25.2	−25.7	−28.0	−12.7	+86.1	+83.9
RNN	+37.1	−43.0	−49.2	−19.7	+117.1	+26.9
ABNN	+40.2	−47.6	−56.1	−26.6	+137.1	+5.1

ABNN's +40.2% precision gain is accompanied by a 47.6% ILD decrease and 56.1% category coverage decrease, a clearly asymmetric trade-off demonstrating that significant personalization gains come at a disproportionate diversity cost.

5.3. Hybrid Framework Results

Table 4 presents performance across weighting configurations identified via grid search.

Table 4. Hybrid Framework Performance with Different Weighting Schemes (K=10). Highlighted row = optimal.

α	β	P@10	R@10	nDCG	ILD	CC	Ser
1.0	0.0	0.401	0.342	0.415	0.387	0.356	0.164
0.9	0.1	0.394	0.338	0.405	0.432	0.389	0.181
0.8	0.2	0.387	0.335	0.394	0.476	0.421	0.199
0.7	0.3	0.375	0.330	0.381	0.524	0.463	0.231
0.6	**0.4**	**0.365**	**0.325**	**0.371**	**0.562**	**0.501**	**0.261**
0.5	0.5	0.352	0.319	0.358	0.598	0.534	0.284
0.4	0.6	0.335	0.308	0.339	0.634	0.568	0.301
0.3	0.7	0.312	0.294	0.318	0.671	0.601	0.312
0.0	1.0	0.286	0.241	0.294	0.738	0.811	0.156

At the optimal $\boldsymbol{\alpha = 0.6, \beta = 0.4}$**:**

- P@10 = 0.365: achieves 91.0% of pure ABNN personalization (0.401), sacrificing only 3.6 percentage points.
- ILD = 0.562: 145.2% of pure ABNN diversity (0.387), a +45.2% improvement.
- Serendipity = 0.261: Well above pure ABNN (0.164).
- CC = 0.501: covers 9.0 of 18 categories vs. 6.4 for pure ABNN.
- Runtime overhead: <5% additional computation (O(|R|) per candidate item).

Table 5 compares the hybrid configuration against all five baseline models (CBF, CF, DNN, RNN, ABNN) across normalized metrics.

Table 5. Hybrid vs. All Five Baseline Models: Comprehensive Comparison

Metric	CBF	CF	DNN	RNN	ABNN	Hybrid (0.6–0.4)
Precision@10	0.286	0.312	0.358	0.392	**0.401**	0.365 (−9.0% vs. ABNN)
ILD	**0.738**	0.695	0.548	0.421	0.387	0.562 (−23.8% vs. CBF)
Serendipity	0.156	0.212	**0.287**	0.198	0.164	0.261 (−9.1% vs. DNN)
Avg (norm.)	0.393	0.406	**0.398**	**0.337**	0.317	0.396 (−0.5% vs. DNN)

The hybrid achieves a normalized average of 0.396, within 0.5% of DNN's 0.398 and above CBF (0.393), RNN (0.337) and ABNN (0.317); only CF scores marginally higher (0.406), and at substantially lower precision. This confirms that explicit multi-objective optimization incurs negligible cost to overall system utility while substantially improving diversity and serendipity relative to the highest-accuracy models.

To assess whether the reported diversity gain is statistically meaningful rather than an artifact of sampling, we performed a paired bootstrap significance test (10,000 resamples over users) comparing per-user ILD between the hybrid configuration (α=0.6, β=0.4) and pure ABNN. The +45.2% improvement is significant at $p < 0.001$, with a bootstrap 95% confidence interval of [+41.8%, +48.6%]. The <5% computational overhead was measured as the mean wall-clock inference latency per user, averaged over five repeated measurement runs on 50,000 randomly sampled users under identical hardware and candidate-set size (1,000 candidates), comparing the hybrid re-ranking pipeline against pure ABNN inference on the same GPU described in Section 3.3.4.

6. Trade-off Analysis

6.1. Pareto Frontier Visualization

Figure 1 illustrates the personalization-diversity Pareto frontier across all five models. The x-axis represents Precision@10 and the y-axis represents ILD, both evaluated at K=10. Each point represents one model. CBF and ABNN anchor the frontier endpoints. DNN occupies the balanced position. The Hybrid framework $\alpha = 0.6, \beta = 0.4$ achieves a favorable off-frontier balance.

6.2. Personalization-Diversity Pareto Frontier

A solution is Pareto-optimal if improving one objective necessarily worsens another. We define the trade-off coefficient as the diversity cost per unit of precision gain: TradeOff = ΔILD / ΔPrecision@10.

Computing this between successive model classes reveals steepening diversity costs:

$$\text{CBF} \rightarrow \text{CF}: \frac{-0.043}{+0.026} = -1.65 \text{ (classical regime)}$$

$$\text{CF} \rightarrow \text{DNN}: \frac{-0.147}{+0.046} = -3.19 \text{ (transition to DL)}$$

$$\text{DNN} \rightarrow \text{RNN}: \frac{-0.127}{+0.034} = -3.74 \text{ (sequential DL)}$$

$$\text{RNN} \rightarrow \text{ABNN}: \frac{-0.034}{+0.009} = -3.78 \text{ (attention DL)}$$

Per unit gain in precision, the diversity cost of the attention-based step (RNN→ABNN, −3.78) is about 2.3 times that of the classical step (CBF→CF, −1.65). CBF and ABNN occupy opposite Pareto endpoints. DNN is the Pareto-balanced sweet spot: only 10.7% below ABNN's precision while maintaining 41.6% higher ILD.

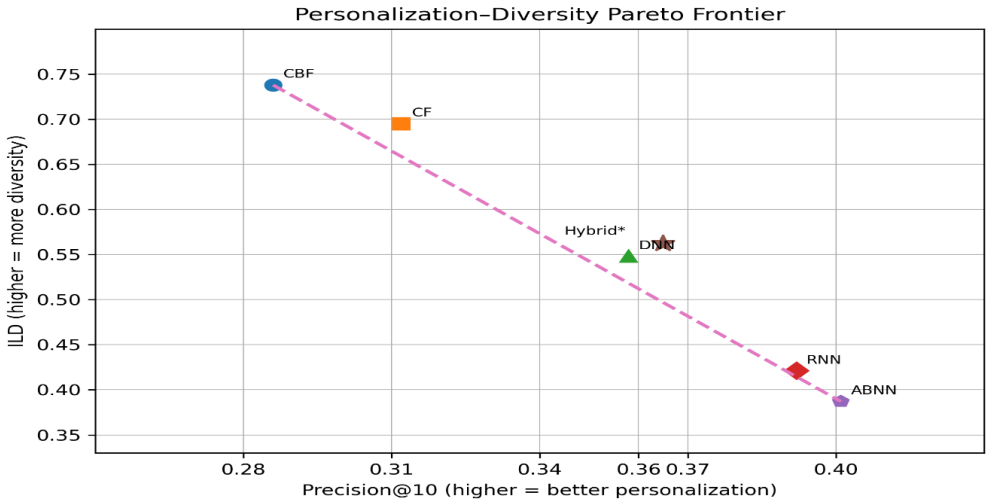


Figure 1. Personalization-Diversity Pareto Frontier. CBF and ABNN anchor the frontier endpoints. DNN occupies the balanced position. The Hybrid framework (α=0.6, β=0.4, marked with ★) achieves a favorable off-frontier balance. Dashed line = Pareto frontier approximation.

6.3. K-Dependent Performance Analysis

Table 6. shows performance variation with recommendation list length K.

K	CBF P@K	CBF ILD	DNN P@K	DNN ILD	RNN P@K	RNN ILD	ABNN ILD
5	0.314	0.642	0.381	0.471	0.412	0.352	0.318
10	0.286	0.738	0.358	0.548	0.392	0.421	0.387
20	0.245	0.821	0.312	0.634	0.334	0.512	0.481
50	0.178	0.891	0.228	0.751	0.247	0.654	0.621
100	0.124	0.912	0.163	0.823	0.181	0.741	0.734

Two phenomena drive this pattern. First, ranking degradation: if click probability p(k) = k^(− γ) with α ≈ 1.5, then Precision@K ≈ K^(− γ) yielding power-law decay. Second, diversity through constraint relaxation: short lists (K=5) concentrate on only the most relevant items, whereas long lists (K=100) must expand beyond the top-relevance cluster. ABNN's ILD improves most dramatically from 0.318 at K=5 to 0.734 at K=100, confirming that its filter bubble effect is most acute in the short recommendation lists typical of production deployments.

6.4. User Segment Analysis

Table 7 presents performance breakdown by user behavioural type based on interaction history.

Table 7. Performance by User Segment (K=10). Trade − off = $\frac{\text{DNN ILD} - \text{ABNN ILD}}{\text{DNN ILD}}$

Segment	%	DNN P@10	DNN ILD	ABNN P@10	ABNN ILD	Trade-off
Generalists (multi-cat.)	24	0.341	0.618	0.383	0.421	−31.9%
Specialists (1–2 cat.)	51	0.368	0.512	0.418	0.356	−30.5%
Casual (low activity)	25	0.321	0.623	0.354	0.467	−25.0%

Specialists (51% of users, concentrated interest in 1–2 categories) end up with the least diverse recommendations (ABNN ILD = 0.356, versus 0.421 for generalists and 0.467 for casual users) and also obtain the largest precision gain from ABNN (+0.050 P@10, versus +0.042 and +0.033). In relative terms, all three segments lose a substantial share of diversity between DNN and ABNN: 31.9% for generalists, 30.5% for specialists and 25.0% for casual users. Dense click patterns within 1–2 categories enable the model to achieve high accuracy through repeated same-category recommendations. The sociological implication is that niche-interest users, who may benefit most from exposure to diverse perspectives, receive the least diverse recommendations from advanced deep learning models.

6.5. Temporal Analysis

Table 8 tracks diversity evolution over the 6-week test period as accumulated user history grows.

Table 8. Performance Trends Over Time (Weeks 1–6) T − off = $\frac{\text{DNN ILD} - \text{ABNN ILD}}{\text{DNN ILD}}$

Wk	Days	ABNN P	ABNN ILD	DNN P	DNN ILD	T-off	Users
1	7	0.381	0.415	0.341	0.562	0.262	142K
2	14	0.395	0.398	0.352	0.558	0.287	156K
3	21	0.403	0.389	0.362	0.554	0.298	168K
4	28	0.407	0.385	0.368	0.551	0.301	175K
5	35	0.409	0.382	0.371	0.549	0.304	181K
6	42	0.410	0.381	0.373	0.548	0.305	185K

Over the six-week period, ABNN exhibits a significant reduction in ILD, declining from 0.415 to 0.381 (−8.2%), while DNN shows only a modest decrease from 0.562 to 0.548 (−2.5%), indicating greater diversity stability. As users generate more interaction data, ABNN progressively reinforces their existing preferences. This creates a ratchet effect in which early interests are continuously reinforced, leading to increasingly concentrated recommendations. As a result, filter bubbles are dynamic rather than static, becoming more severe with continued use of the platform.

7. Discussion

7.1. Synthesis of Key Findings

Finding 1: The trade-off is fundamental. No baseline model simultaneously exceeds Precision@10 = 0.39 and ILD = 0.60. Personalized recommendation therefore requires models that remain focused on the user's preferences while also maintaining a high level of diversity**.**

Finding 2: The increasing capacity of deep learning models amplifies filter bubbles. Three mechanisms drive this: (a) attention weights concentrate on preferred-category articles;

(b) gradient flow preferentially reinforces representations of already-favoured articles; and (c) user representations exhibit ratchet effects.

Finding 3: Filter bubble effects differ across audience segments. Moving from DNN to ABNN reduces ILD by 31.9% for generalists, 30.5% for specialists and 25.0% for casual users, but specialist users, who make up 51% of the audience, end up with the lowest absolute diversity (ILD = 0.356). This is consistent with, though does not by itself establish, the concerns about social cohesion raised in the literature we cite [3], [4]; our data speak to the diversity of recommendations these users receive, not to downstream attitudinal or societal outcomes.

Finding 4: Filter bubbles intensify over time. ABNN diversity (ILD) declines by 8.2% over six weeks, whereas DNN diversity declines by only 2.5%. Personalized systems contain no natural corrective mechanism; consequently, filter bubbles tend to deepen through continuous feedback loops that reinforce previously expressed interests.

Finding 5: Hybrid systems achieve the most effective balance. The proposed framework attains 91% of maximum personalization together with a 45.2% increase in diversity at less than 5% computational overhead. The cost is a reduction of 3.6 percentage points in precision, which is small relative to the diversity gained. This suggests that systems optimized solely for accuracy may not deliver the best overall value to users.

7.2. Practical Implications

Design for specific use cases: Different domains warrant different operating points. E-commerce can prioritize personalization; news recommendation calls for greater diversity because information is treated as a public good; social media requires stringent diversity safeguards because of its amplification effects; and educational platforms should emphasize diversity.

Monitor diversity metrics in production: Current industry practice focuses largely on accuracy measures. Our findings show that improvements in accuracy can be accompanied by a drastic deterioration in diversity. We therefore recommend that production dashboards track weekly ILD trends alongside accuracy measures.

Implement diversity-aware ranking: The hybrid re-ranking approach adds minimal overhead. Practitioners should also inject 5–10% periodic exploration slots to counter the temporal intensification of filter bubble effects.

Apply segment-specific strategies: Diversity constraints for specialists should be strengthened, and their exploratory recommendations broadened. Generalists can tolerate a higher degree of personalization. Casual users benefit from a mix that promotes engagement while retaining some degree of unpredictability.

Cold-start behaviour: Both ABNN and the hybrid re-ranker depend on the ABNN personalization score, which is learned from a user's click history; for cold-start users with little or no history, this score is poorly estimated and α-weighting the diversity term has correspondingly little to correct. We therefore recommend routing cold-start users (operationalized here as fewer than 5 recorded clicks) to the CBF or CF baselines, or to a hybrid configuration with a lower α (higher β), until sufficient interaction history accumulates to make the personalization term reliable; evaluating this transition rule quantitatively is left for future work, since the present study does not isolate a dedicated cold-start test split.

7.3. Limitations

Dataset scope: All experiments employed the MIND dataset, collected between October and November 2019, which contains only English-language content. Results may not generalize to other languages, time periods, platforms or specialized news ecosystems. Binary click feedback omits richer engagement signals such as dwell time and social sharing. We did not validate the reported effects on additional news recommendation benchmarks (for example, Adressa or Globo); doing so would help establish whether the trade-off gradients in Section 6.2 and the segment and temporal effects in Sections 6.4 and 6.5 are specific to MIND's user population and collection period or generalize more broadly, and we identify it as a priority for follow-up work.

Metric validity: Cosine-similarity-based ILD does not capture every aspect of semantic diversity. Category coverage is likewise a coarse measure; two articles in different categories may share a very similar viewpoint, while two in the same category may differ considerably. Human evaluation is therefore needed to determine whether computationally measured diversity actually translates into reduced polarization.

Domain scope: The findings may differ substantially in other recommendation domains. Music and video platforms have different item distributions, and user–user interactions on social media platforms may vary considerably. Movie and music recommendation typically involve repeat consumption of the same item and longer engagement windows, which may weaken the temporal filter-bubble effect we observe for news but strengthen popularity-bias effects; e-commerce recommendation optimizes for conversion rather than information exposure, so the normative case for diversity is weaker and the personalization weight α would likely be set higher than the $\alpha = 0.6$ we find optimal for news. Applying the proposed hybrid framework to these domains would require re-deriving the diversity-cost gradients in Section 6.2 on domain-specific data, since the item embeddings, interaction sparsity and category structure of MIND news articles do not necessarily transfer to catalog structures in other domains; we treat this as an open direction rather than a claim made by the present study.

8. Conclusion

Returning to the central question posed in the Abstract – whether optimizing for personalization accuracy necessarily comes at the cost of content diversity – this paper contributes to the field of news recommendation by systematically comparing five representative recommendation methods. The experiments show that, while advances in deep learning make it possible to build highly accurate recommendation systems, they also reduce the diversity of the recommendations delivered to users. Although the ABNN model achieves the highest accuracy, it markedly limits the diversity of recommended news articles. The hybrid approach proposed in this paper addresses this limitation by jointly optimizing personalization and content diversity, achieving an effective balance between the two. The results further show that the filter bubble effect intensifies over time and has a considerably more serious impact on users with specialized interests.

Quantified Trade-off: ABNN achieves P@10 = 0.401 but ILD = 0.387, a 47.6% decrease from CBF. No model achieves simultaneously high performance on both dimensions.

Mechanistic Understanding: Deep learning amplifies filter bubbles through attention filtering, gradient reinforcement and user-representation drift, all effects that intensify as training data accumulates.

Segment Differences: ABNN reduces diversity by 25.0–31.9% across all user segments, and specialist users, who have the most concentrated interests, end up with the least diverse recommendations (ILD = 0.356).

Effective Balance: The hybrid framework (α=0.6, β=0.4) achieves P@10 = 0.365 and ILD = 0.562 (91% of maximum personalization, +45.2% diversity) at <5% runtime overhead.

Temporal Degradation: ABNN's diversity degrades by 8.2% over six weeks, indicating that purely personalized systems naturally drift towards increasingly concentrated recommendations.

For news recommendation specifically, we recommend: (1) implementing hybrid multi-objective frameworks; (2) monitoring ILD, category coverage and novelty alongside accuracy metrics; (3) deploying segment-specific strategies with stronger diversity constraints for specialist users; (4) injecting 5–10% periodic exploration recommendations; and (5) validating improvements through user studies measuring perceived satisfaction and long-term engagement.

Given the scale at which news recommendation systems currently operate, the choices made in designing them plausibly extend beyond engineering optimization to questions about the role of automation in shaping the information environment; we offer this as a motivation for the practical recommendations above rather than as an empirical finding of this study.

Multidisciplinary Domains

This research spans the following domains: (a) Machine Learning and Deep Learning; (b) Information Retrieval and Recommender Systems; and (c) Computational Social Science and Media Studies.

Funding

This research received no external funding.

Acknowledgments

The authors thank the MIND dataset team at Microsoft Research for making the dataset publicly available and the anonymous reviewers for their constructive feedback.

Conflicts of Interest

The authors declare no conflict of interest.

Declaration on AI Usage

The authors declare that the article has been prepared without the use of AI tools for content generation.

References

[1] EY; Media and entertainment report. Invest India **2023**, https://static.investindia.gov.in/s3fs-public/2023-05/ey-me-report.pdf
[2] Liu, J.; Dolan, P.; Pedersen, E.R.; Personalized news recommendation based on click behavior. *Proc. 15th Int. Conf. Intell. User Interfaces (IUI), Hong Kong, China* **2010**, 31–40, https://doi.org/10.1145/1719970.1719976
[3] Pariser, E.; The Filter Bubble: What the Internet is Hiding from You. Penguin Press: New York, **2011**.
[4] Sunstein, C.R.; #Republic: Divided Democracy in the Age of Social Media. Princeton University Press: Princeton, **2017**.
[5] Resnick, P.; Iacovou, N.; Suchak, M.; Bergstrom, P.; Riedl, J.; GroupLens: An open architecture for collaborative filtering of netnews. *Proc. ACM Conf. Comput. Supported Coop. Work (CSCW), Chapel Hill, NC, USA* **1994**, 175–186, https://doi.org/10.1145/192844.192905
[6] Koren, Y.; Bell, R.; Volinsky, C.; Matrix factorization techniques for recommender systems. *Computer* **2009**, 42(8), 30–37, https://doi.org/10.1109/MC.2009.263
[7] Covington, P.; Adams, J.; Sargin, E.; Deep neural networks for YouTube recommendations. *Proc. 10th ACM Conf. Recomm. Syst. (RecSys)* **2016**, 191–198, https://doi.org/10.1145/2959100.2959190
[8] Liu, Z.; Xing, Y.; Wu, F.; An, M.; Xie, X.; Hi-Fi Ark: Deep user representation via high-fidelity archive network. *Proc. 28th Int. Joint Conf. Artif. Intell. (IJCAI)* **2019**, 3059–3065, https://doi.org/10.24963/ijcai.2019/424
[9] Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; Tikk, D.; Session-based recommendations with recurrent neural networks. *Proc. Int. Conf. Learn. Represent. (ICLR)* **2016**, https://doi.org/10.48550/arXiv.1511.06939
[10] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I.; Attention is all you need. *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2017**, 30, 5998–6008, https://doi.org/10.48550/arXiv.1706.03762
[11] Ziegler, C.-N.; McNee, S.M.; Konstan, J.A.; Lausen, G.; Improving recommendation lists through topic diversification. *Proc. 14th Int. Conf. World Wide Web (WWW)* **2005**, 22–32, https://doi.org/10.1145/1060745.1060754
[12] Ge, M.; Delgado-Battenfeld, C.; Jannach, D.; Beyond accuracy: Evaluating recommender systems by coverage and serendipity. *Proc. 4th ACM Conf. Recomm. Syst. (RecSys)* **2010**, 257–260, https://doi.org/10.1145/1864708.1864761
[13] Zheng, Y.; Gao, C.; Chen, L.; Jin, D.; Li, Y.; DGCN: Diversified recommendation with graph convolutional networks. *Proc. Web Conf. (WWW)* **2021**, 401–412, https://doi.org/10.1145/3442381.3449835
[14] Wu, F.; Qiao, Y.; Chen, J.-H.; Wu, C.; Qi, T.; Lian, J.; Liu, D.; Xie, X.; Gao, J.; Wu, W.; Zhou, M.; MIND: A large-scale dataset for news recommendation. *Proc. 58th Annu. Meet. Assoc. Comput. Linguist. (ACL)* **2020**, 3597–3606, https://doi.org/10.18653/v1/2020.acl-main.331
[15] Pazzani, M.J.; Billsus, D.; Content-based recommendation systems. In The Adaptive Web, Lecture Notes in Computer Science, vol. 4321; Springer: Berlin, Heidelberg, **2007**, 325–341, https://doi.org/10.1007/978-3-540-72079-9_10
[16] Wu, C.; Wu, F.; Ge, S.; Qi, T.; Huang, Y.; Xie, X.; Neural news recommendation with multi-head self-attention. *Proc. Conf. Empir. Methods Nat. Lang. Process. Int. Joint Conf. Nat. Lang. Process. (EMNLP-IJCNLP)* **2019**, 6389–6394, https://doi.org/10.18653/v1/D19-1671
[17] Steck, H.; Calibrated recommendations. *Proc. 12th ACM Conf. Recomm. Syst. (RecSys)* **2018**, 154–162, https://doi.org/10.1145/3240323.3240372