

Peeling Back the Layers: An In-Depth Evaluation of Encoder Architectures in Neural News Recommenders

Andreea Iana¹, Goran Glavaš² and Heiko Paulheim¹

¹Data and Web Science Group, University of Mannheim, Germany

²Center for Artificial Intelligence and Data Science, University of Würzburg, Germany

Abstract

Encoder architectures play a pivotal role in neural news recommenders by embedding the semantic and contextual information of news and users. Thus, research has heavily focused on enhancing the representational capabilities of news and user encoders to improve recommender performance. Despite the significant impact of encoder architectures on the quality of news and user representations, existing analyses of encoder designs focus only on the overall downstream recommendation performance. This offers a one-sided assessment of the encoders' similarity, ignoring more nuanced differences in their behavior, and potentially resulting in sub-optimal model selection. In this work, we perform a comprehensive analysis of encoder architectures in neural news recommender systems. We systematically evaluate the most prominent news and user encoder architectures, focusing on their (i) representational similarity, measured with the Central Kernel Alignment, (ii) overlap of generated recommendation lists, quantified with the Jaccard similarity, and (iii) the overall recommendation performance. Our analysis reveals that the complexity of certain encoding techniques is often empirically unjustified, highlighting the potential for simpler, more efficient architectures. By isolating the effects of individual components, we provide valuable insights for researchers and practitioners to make better informed decisions about encoder selection and avoid unnecessary complexity in the design of news recommenders.

Keywords

neural news recommendation, evaluation, representational similarity, news encoder, user encoder, retrieval similarity

1. Introduction

Content-based neural models have become the state of the art in news recommendation. Neural news recommenders (NNRs) typically comprise a news encoder and a user encoder. The news encoder learns semantically meaningful representations of news articles, whereas the user encoder embeds the preferences of users based on their click history [1]. NNRs take the candidate news articles and a user's reading history as input. The relevance of the candidate to the user is determined by comparing, with a scoring function, the latent representations of the two inputs, generated with the corresponding encoders. Given the key role of encoders in NNRs, a significant body of research has focused on improving the quality of news encoding and user modeling to improve recommendation performance [2, 3, 1].

On the one hand, ablation studies of recommenders typically analyze individual model components in isolation, neglecting other architecturally comparable model designs [4, 5, 6]. At the same time, we see emerging evidence that widely used NNRs exhibit similar performance despite varying model complexities, and that the overall complexity of the recommenders' architecture could be reduced [7, 8]. This highlights the need for a more granular comparison of the individual building blocks to understand their behavior and impact on the overall system. While Möller and Padó [9] or Iana et al. [7] evaluated NNR components such as scoring functions and training objectives, a systematic analysis of encoder architectures is still lacking. Such insights would enable researchers and practitioners alike to make more informed choices about encoder selection in NNR design.

INRA 2024: 12th International Workshop on News Recommendation and Analytics, October 14–18, 2024, Bari, Italy

*Corresponding author.

✉ andreea.iana@uni-mannheim.de (A. Iana); goran.glavas@uni-wuerzburg.de (G. Glavaš); heiko.paulheim@uni-mannheim.de (H. Paulheim)

ORCID 0000-0002-7248-7503 (A. Iana); 0000-0002-1301-6314 (G. Glavaš); 0000-0003-4386-8195 (H. Paulheim)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

On the other hand, progress in the architectural design of news and user encoders is generally measured in terms of the recommender’s overall classification and ranking capability [4, 6, 10, 11, 7]. Nonetheless, the quality of the embeddings produced by the news and user encoders is equally crucial, given the reliance of the recommender on the dense retrieval paradigm. Therefore, evaluating NNRs and their components solely in terms of downstream recommendation performance provides a simplified perspective, potentially overlooking subtle differences in the encoders’ behavior. We thus argue that investigating the similarity of embeddings generated by various news and user encoders would offer a more nuanced understanding of their behavior, in turn benefiting the model selection process.

In this work, we perform a systematic analysis of the encoder architectures of NNRs. Unlike conventional evaluation studies, we isolate the effects of each core component to the largest possible extent. Concretely, we analyze the most prominent news and user encoder architectures in terms of (i) the similarity of learned news, and respectively, user representations, using the Central Kernel Alignment [12] metric, (ii) the similarity of the generated recommendation lists, quantified by means of the Jaccard coefficient, and (iii) the impact on the overall recommendation performance. Our findings provide a better understanding of news recommenders encoder architectures, not only from a recommendation performance perspective, but also in terms of their representational similarity. We demonstrate that the complexity of some encoding techniques is often empirically unjustified, emphasizing the potential benefits of simpler, more efficient architectures. These results fundamentally challenge the common practice of over-engineering NNR encoders. Consequently, we derive three key takeaways, arguing that (1) the semantic richness of news encoders is crucial for effective recommendation, that (2) user encoders can be significantly simplified without sacrificing performance, and lastly, (3) we advocate for more rigorous evaluation to guide better informed model selection.

2. Related Work

Neural news recommenders have significantly advanced in recent years, with encoder architectures playing a key role in capturing the semantic and contextual information of news articles and user profiles. Consequently, a large strand of work has focused on improving the representational capabilities of recommenders by developing ever more accurate, and often complex, news encoding and user modeling architectures. As such, these works have analyzed individual aspects of the NNR components, such as the use of different attention mechanisms in the news or user encoder [4, 5, 13], the impact of various user modeling [6, 10, 14, 13, 7] or news embedding [15, 4, 5, 13, 11, 16, 17] techniques, or the importance of modeling different news features [15, 4, 18, 19, 20, 21, 22] and user characteristics [5, 6, 23, 24]. Ablation studies in these cases are usually conducted in isolation for the component under consideration, without taking into account the broader architectural context.

In contrast, another strand of work has started evaluating the impact of NNR components or training strategies across an array of recommendation approaches. For example, Wu et al. [11] have investigated the usage of various pretrained language models as the backbone of widely used NNRs. Möller and Padó [9] have evaluated the impact of scoring functions, whereas Iana et al. [7] have analyzed different user modeling techniques and training objectives. The latter have highlighted the similar recommendation performance achieved by certain models despite differences in architectures and complexity, emphasizing the potential to simplify the design of news recommender systems. While these works shed new light on core components of the recommendation model, their evaluation is most often solely based on the downstream recommendation performance.

The similarity of encoders in NNRs can additionally be measured in terms of their generated representations. More generally, there exist numerous methods for quantifying the similarity of neural networks. Two main categories include (i) representational similarity, which assesses differences in the activations of intermediate layers of neural networks, and (ii) functional similarity, which compares the networks’ outputs in relation to their task [25]. Several works have focused on evaluating the representational similarity of (large) language models [26, 27, 28, 29] or of embedding models in Retrieval Augmented Generation systems [30], which are often employed as the news encoding component of NNRs.

Table 1

Abbreviations and their description.

Abbreviation	Description
CNN	convolutional neural network [35]
Att	attention network
AddAtt	additive attention [36]
MHSA	multi-head self-attention [37]
PLM	pre-trained language model
PLM _[CLS]	the PLM’s output [CLS] token representation
PLM _{tokenemb+Att}	PLM’s token embeddings pooled with an attention network [11]
SE	sentence encoder
Con	concatenation
Linear	linear layer
LF	late fusion [7]
GRU	gated recurrent unit [38]
CandAware	candidate-aware user encoder [10]

Nevertheless, to the best of our knowledge, no work so far compares neither user encoders nor news encoders with respect to representational and functional similarity. In this work, we fill this gap by comprehensively analyzing the primary components of NNR encoder architectures for both news and user inputs.

3. Methodology

We firstly introduce the building blocks of personalized NNRs. Afterwards, we discuss metrics to evaluate both the recommendation performance, as well as the representational similarity of the news and user encoders.

3.1. Encoders of Neural News Recommenders

Content-based neural news recommenders consists of a dedicated **(i) news encoder (NE)** and a **(ii) user encoder (UE)** [1]. The NE transforms different input features (e.g., title, abstract, categories, named entities, images) of a news article n into a latent news representation \mathbf{n} . The UE aggregates the embeddings of the clicked news \mathbf{n}_i^u from a user’s u history into a user-level representation \mathbf{u} . Finally, the embedding of a candidate news \mathbf{n}^c , outputted by the NE, is scored against the user representation \mathbf{u} produced by the UE, to determine the relevance of the candidate to the user $s(\mathbf{n}^c, \mathbf{u})$. The dot product of the two embeddings \mathbf{n}^c and \mathbf{u} is the most common scoring function [4]. NNRs are trained via conventional classification objectives [31] with negative sampling [32], or contrastive objectives [33, 34]. The building blocks of NNRs (i.e., NE, UE, scoring function, training objective) altogether drive the overall performance of the recommender. Since the NE and UE determine what information of the documents and users is embedded by the model, and ultimately, propagated through the recommendation pipeline, both types of encoders play a similarly important role in model selection. We introduce the abbreviations used for the remainder of the paper in Table 1.

News Encoder Architectures. The NE can generally be decomposed into a *text encoder*, which embeds the textual content of a news article, and several *feature-specific encoders* (e.g., category, sentiment, entity encoder), which learn to represent further input features different from text chunks. While the former represents a key component of all NNRs, the latter types of encoders are optional and only utilized whenever the textual content is enriched with additional features which might capture or emphasize other aspects of a news article. Lastly, the NE combines the intermediate embeddings produced by the text and feature-specific encoders into a news-level representation by means of a *multi-feature aggregation strategy*.

Table 2

Text encoder architectures.

Text Embedding Type	Text Encoder	References
word embeddings	CNN + AddAtt	[4, 6, 18, 23, 39, 40]
	MHSA + AddAtt	[41, 5, 19, 42, 43, 44, 21, 45, 46, 47, 10, 14]
	CNN + MHSA + AddAtt	[13]
language model	PLM _{tokenemb+Att}	[48, 11, 49, 50]
	PLM _[CLS]	[51, 52, 16, 53, 33]
	SE	[17]

Table 3

Multi-feature aggregation strategies for combining textual and categorical representations of news.

Multi-feature aggregation	References
AddAtt	[4, 40, 46, 54, 55, 14]
Linear	[44, 10]
Con	[6, 43, 56]

Table 4

User encoder architectures.

User Encoder	References
LF	[33, 17]
AddAtt	[41, 4, 23, 18, 57, 40, 56, 58]
MHSA+AddAtt	[5, 19, 48, 45, 20, 47]
GRU _{ini}	[6, 54]
GRU _{con}	[6, 44]
GRU+MHSA+AddAtt	[13]
CandAware (CNN+MHSA+AddAtt)	[10]

We list the most used types of text encoders that we consider in our analysis in Table 2, alongside examples of NNRs using them. We distinguish between text encoders that rely on pretrained word embeddings, contextualized by means of convolutional or self-attention networks, and the more recent architectures that employ pretrained language models.¹ We additionally consider the most common multi-feature aggregation approaches used to integrate text and other content feature (e.g., category) embeddings into the unified news representation, as shown in Table 3.

User Encoder Architectures. Parameterized UEs represent the most popular user modeling technique. They learn user representations by means of sequential or attentive networks that contextualize the embeddings of clicked news based on patterns in the user’s click behavior. UEs can be further differentiated into candidate-agnostic (i.e., users are encoded separately from candidate news) and candidate-aware (i.e., the user-level aggregation contextualizes the embeddings of clicked news against the embedding of each candidate) encoders [7]. More recently, Iana et al. [7] proposed the parameter-free late fusion (LF) approach. LF first averages the clicked news embeddings \mathbf{n}_i^u to a user embedding $\frac{1}{N}\sum_{i=1}^N \mathbf{n}_i^u = \mathbf{u}$. The inner product of the embedding of the candidate news \mathbf{n}^c and the user embedding \mathbf{u} then represents the relevancy score. Table 4 lists the main user encoder architectures that we evaluate in this work, together with examples of models using them.

¹Note that in this work we do not evaluate encoders which rely on news or user graphs, as such graphs are heavily dataset-dependent. We instead focus on the most used core components of encoders, and leave the analysis of graph-based techniques for future work.

3.2. Similarity Evaluation

We evaluate NEs and UEs on three dimensions: (i) downstream recommendation performance, (ii) similarity of generated recommendations, and (iii) similarity of learned news or user representations.

Downstream Recommendation Performance. NNRs are usually evaluated with regards to classification (e.g., AUC) and ranking (e.g., MRR, nDCG) performance. In this work, we focus on the ranking performance, which we quantify using nDCG@ k .

Similarity of Generated Recommendations. We analyze the retrieval similarity of recommenders that use different news or user encoder architectures by the similarity of their top- k recommended articles. Specifically, for the same set of users, we firstly generate the corresponding recommendation lists R and R' with models M and M' , respectively. We then measure the similarity of retrieved results with the Jaccard similarity coefficient:

$$Jaccard(R, R') = \frac{|R \cap R'|}{|R \cup R'|} \quad (1)$$

where $|R \cap R'|$ denotes the set of articles recommended by both models, and $|R \cup R'|$ the union of all unique news recommended by the two models. The Jaccard similarity score is bounded in the $[0, 1]$ interval, with 1 indicating that both models recommend an identical set of news. Note that the lengths of both recommendation lists will be equal to the full set of candidate news N_u^c for a given user u , namely $|R| = |R'| = |N_u^c|$, regardless of the recommendation model used. Thus, to differentiate the retrieval performance of two models, we compute the Jaccard similarity only for the top- k recommendations, ordered descendingly by the recommendation scores. Note that in comparison to nDCG@ k , the Jaccard similarity measures the overlap of the recommended news between two models without considering the order of the articles in the recommendation set.

Embedding Similarity. Numerous measures quantify the representational similarity of neural networks [25]. Many of these methods require an identical dimensionality of the compared embeddings or an alignment of the latent representation spaces across models. Since these constraints are not straightforwardly met by the embeddings produced with different news and user encoder architectures, we choose to measure the similarity of embeddings using the Centered Kernel Alignment (CKA) with a linear kernel [12]. Concretely, for a given representation \mathbf{E} , we firstly mean-center it column-wise. Afterwards, we compute the pair-wise similarity of the representation of each instance i to all other instances in \mathbf{E} . Each row i in the resulting similarity matrix \mathbf{S} thus comprises the similarity between instance's i embedding and all other embeddings, including itself. For two different models with the same number of embeddings \mathbf{E} and \mathbf{E}' , the resulting representational similarity matrices \mathbf{S} and \mathbf{S}' , respectively, can be directly compared using the Hilbert-Schmidt Independence Criterion (HSIC) [59] as follows:

$$CKA(\mathbf{E}, \mathbf{E}') = \frac{HSIC(\mathbf{S}, \mathbf{S}')}{\sqrt{HSIC(\mathbf{S}, \mathbf{S})HSIC(\mathbf{S}', \mathbf{S}')}} \quad (2)$$

The CKA similarity scores are bounded to the interval $[0, 1]$, with a score of 1 denoting equivalent representations.

4. Experimental Setup

Data. We conduct experiments on the MINDsmall [60] dataset. Since Wu et al. [60] do not release the test set labels, we use the validation portion for testing, and split the respective training set into temporarily disjoint training (the first four days of data) and validation (the last day of data) subsets.

Evaluation Setup. We separately evaluate the encoder architectures of NNRs. In all experiments, we consider both mono-feature (e.g., title) and multi-feature (e.g., title and categories) inputs for the NE. In the latter case, we learn category representations by means of a linear encoder that combines a category

ID embedding layer with a dense layer [4, 6, 10, 14]. Moreover, in our analysis of NE architectures, we adopt the *late fusion* approach [7] instead of the traditional parameterized UEs. This evaluation setup allows us to isolate the effects of NEs, and to avoid additional confounding factors stemming from the UE, which also influence the output of the NNR. Similarly, when evaluating the similarity of UE architectures, we keep the underlying NE of the recommender fixed, i.e., we analyze different UEs for the same base NE.

Implementation and Optimization Details. We train all models with the standard cross-entropy loss, using dot product as the scoring function. We use 300-dimensional pretrained Glove embeddings [61] to initialize the word embeddings of the word embedding-based text encoders. Additionally, we use RoBERTa-base [62] and the news-specialized multilingual sentence encoder NaSE [17] for the PLM-based and SE-based text encoders, respectively. We fine-tune only the last four layers of the language models. Following prior work [32], we sample four negatives per positive example during training. We set the maximum history length to 50, and train all models with mixed precision, the Adam optimizer [63], and a batch size of 8. We train all NNRs with word embedding-based NEs for 20 epochs, and those with language model-based NEs for 10 epochs. We tune the main hyperparameters of all NNRs using grid search. Concretely, we search for the optimal learning rate in $\{1e-3, 1e-4, 1e-5\}$. We optimize the number of heads in the multi-head self-attention networks in [8, 12, 16, 20, 24, 32], and the query vector dimensionality by sweeping the interval [50, 200] with a step of 50. We run all experiments using the implementations available in the NewsRecLib library [64], on a cluster with virtual machines, training each model on a single NVIDIA A100 40GB GPU.²

5. Results and Discussion

We begin by analyzing the similarity of core NE architectures, followed by an evaluation of UE similarity using the same base news encoding approach. In both cases, we first compare the architectures in terms of ranking performance and retrieval similarity, as these are standard evaluation approaches in the recommender systems field. We then assess the architectures from the perspective of pair-wise embedding similarity.

5.1. News Encoder Architectures

Figure 1 shows the ranking performance, in terms of nDCG@10, of NNRs for different news encoders and input features. For the same input type, e.g. mono-feature, we find a high similarity between the performance of recommenders based on the same family of text encoders. Specifically, text encoders using pretrained static word embeddings are outperformed by those based on PLMs. Moreover, MHSA+AddAtt and CNN+MHSA+AddAtt appear to have nearly identical performance, despite the increased complexity of the latter architecture. Similarly, simply using the [CLS] token representation produced by the PLM instead of pooling tokens with an attention network as proposed by Wu et al. [11] leads to slightly better performance while maintaining a lighter text encoder.

Our findings show that among the three multi-feature aggregation strategies, the Linear and AddAtt approaches always outperform the Con technique. This is intuitive, as the concatenation of vectors with varying dimensionality from non-aligned representation spaces will be sub-optimal. In contrast, both other aggregation strategies project the intermediate text and category embeddings in the same latent representation space. Most importantly, we find that leveraging categories in addition to textual news content as input features is most beneficial for word embedding-based text encoders, and becomes irrelevant or slightly detrimental for the domain-adapted sentence encoder. This can be explained, on the one hand, by the better representational capabilities of the much larger language models which acquire contextual understanding during pretraining compared to static word embeddings. On the other hand, sentence encoders, especially domain-specialized models such as NaSE [17], better capture

²<https://github.com/andreeaiana/newsreclib>

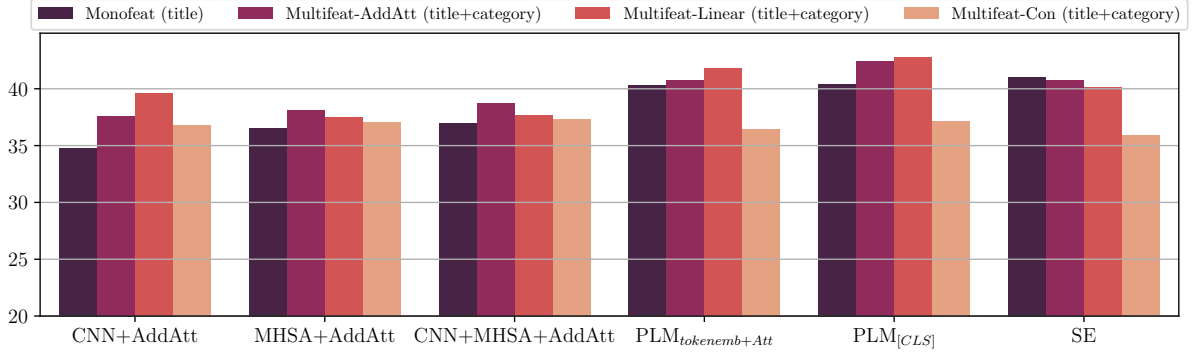
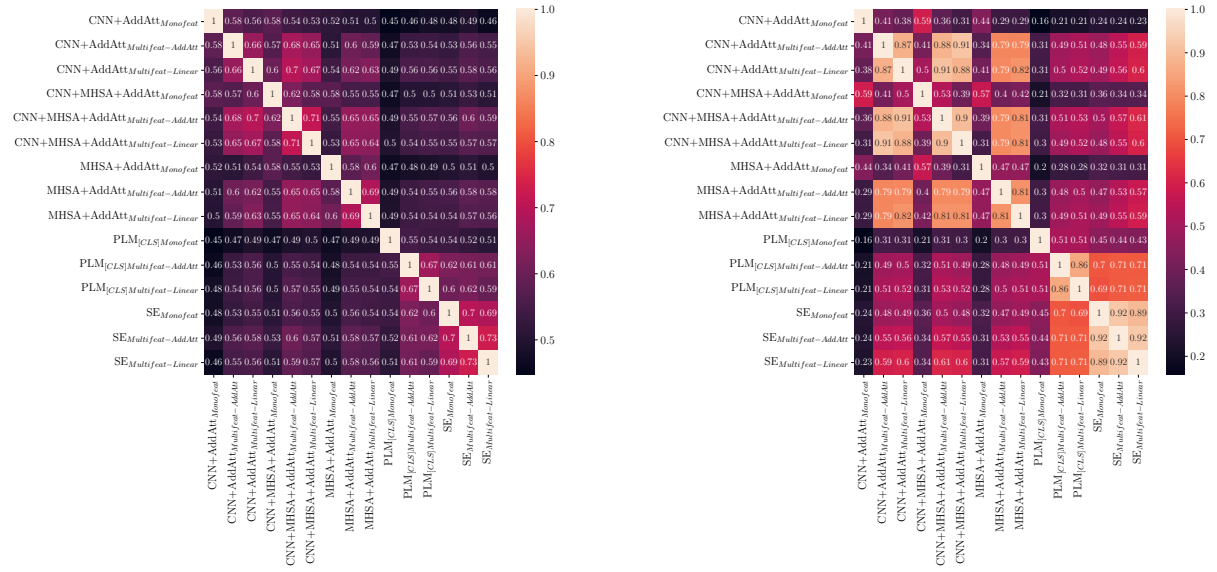


Figure 1: Ranking performance (nDCG@10) of recommenders depending on the news encoder architecture and input features.



(a) Jaccard similarity of top-10 recommended news.

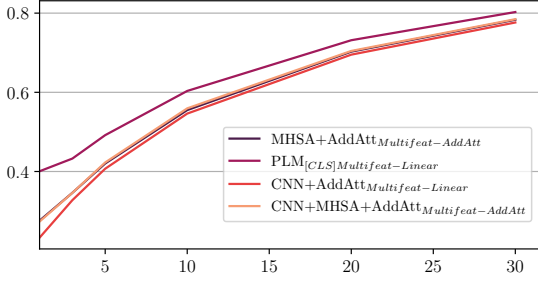
(b) CKA similarity of news embeddings.

Figure 2: Retrieval and representational similarity of models with different news encoder architectures and input features. Each model’s subscript indicates the type of input, and the multi-feature aggregation strategy, if used.

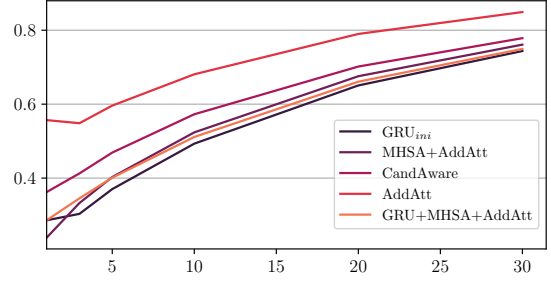
nuances and topics from text due to their pretraining objectives that focus on the overall sentence-level semantics.

We find these similarities in ranking performance between the various news encoding architectures to be reflected in the similarity of retrieved articles. Figure 2a illustrates the pair-wise Jaccard similarity scores between the top-10 recommended news per model. Note that we exclude $PLM_{tokenemb+Att}$, as well as the Con multi-feature aggregation strategy from further analysis for the sake of brevity, and due to their poorer performance. As expected, models from the same family of text encoders show higher similarity scores. The lower Jaccard similarities across word embedding and PLM-based intra-family models using mono-feature versus multi-feature input supports our previous observation regarding the low relevance of categorical input for the domain-adapted SE.

The overall pair-wise Jaccard similarities could initially suggest that most NEs result in little overlap in their recommendation lists. However, a Jaccard similarity score of 0.54 between two models for a list of $k = 10$ recommended items means that, in practice, the two models output 7 identical articles. Analogously, a score of 0.66 indicates an overlap of 8 out of 10 recommendations. As Figure 2a shows, the recommendations generated by the various NE architectures differ by more than 3 articles in a list of length 10 only in rare cases. In other words, regardless of the architectural differences and complexities,



(a) SE_{Monofeat} against the best performing architectures from the other news encoder families.



(b) LF against other user encoder architectures evaluated, with CNN+AddAtt as the base news encoder.

Figure 3: Evolution of Jaccard similarity for different values of k .

the encoders retrieve, on average, the same articles in over 70% of the time.

Taking a look at the CKA similarity of the test set news embeddings produced with the different NEs, shown in Figure 2b, corroborates our hypothesis: intra-family NEs tend to produce similar embeddings when using the same type of input features. The news-adapted SE constitutes the only exception, as its embeddings are not significantly influenced by leveraging categories as additional input features. Additionally, we observe a higher representational similarity between the CNN+AddAtt, MHA+AddAtt, and CNN+MHA+AddAtt models with multi-feature input, and a slightly lower similarity between PLM[CLS] and SE-based models. Overall, the high similarity of representations, of recommendation performance, and the large overlap of generated recommendations by the CNN+AddAtt, MHA+AddAtt, and CNN+MHA+AddAtt multi-feature NEs contest the empirical contribution of incremental architectural changes in the NE architecture of some NNRs.

Lastly, we contrast the representational similarity of models against their retrieval similarity. Figure 3a illustrates the evolution of Jaccard similarity scores between the SE_{Monofeat} encoder and the best performing architecture from each remaining NE family for different values of k . For small values of k , we observe a lower similarity of retrieved news for inter-family text encoders, with scores converging toward 1 for larger k . An important insight here is that for low values of k (e.g., $k < 10$), the news articles retrieved by different NEs tend to be identical, on average, in more than half of the recommended items (e.g., a Jaccard of 0.42 for $k = 5$ translates into an overlap of 3 out of 5 items). We observe this behavior even for models with lower representational similarity scores, e.g., word embedding-based NEs versus language model-based NEs. This is relevant from a practical perspective, where retrieval similarity is of most interest for small values of k . It would imply, on the one hand, that the representational similarity of NEs might not directly correlate with the retrieval performance for low k . On the other hand, this evidence re-affirms our earlier hypothesis that small differences in the architecture and complexity of news encoders do not result in large differences in the actual recommended items.

5.2. User Encoder Architectures

We next investigate the ranking performance, with regards to $nDCG@10$, for different UE architectures for the same base NE. Figure 4 displays the corresponding results, for both mono-feature, and well as multi-feature input. We find that the LF, AddAtt, and CandAware encoders perform the best across all families of NEs. More specifically, the much simpler LF and AddAtt encoders outperform the complex CandAware modeling technique in the case of language model-based NEs, and perform similarly with CandAware for word embedding-based NEs, as previously suggested by Iana et al. [7]. Surprisingly, these two approaches also consistently achieve better ranking than sequential-based UEs (i.e., GRU+MHA+AddAtt, GRU_{ini}, GRU_{con}). Once again, we see that using categorical information alongside the textual content as input to the NE benefits all recommenders regardless of the UE family. The only exception, as previously discussed, are SE-based NNRs. Interestingly, we see that multi-feature inputs close the gap (i) in between inter-family UEs for the same base NE, and (ii) across intra-family UEs for

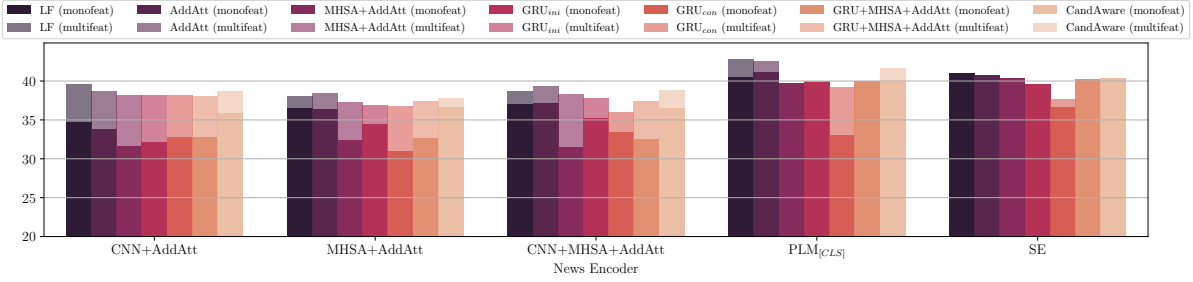
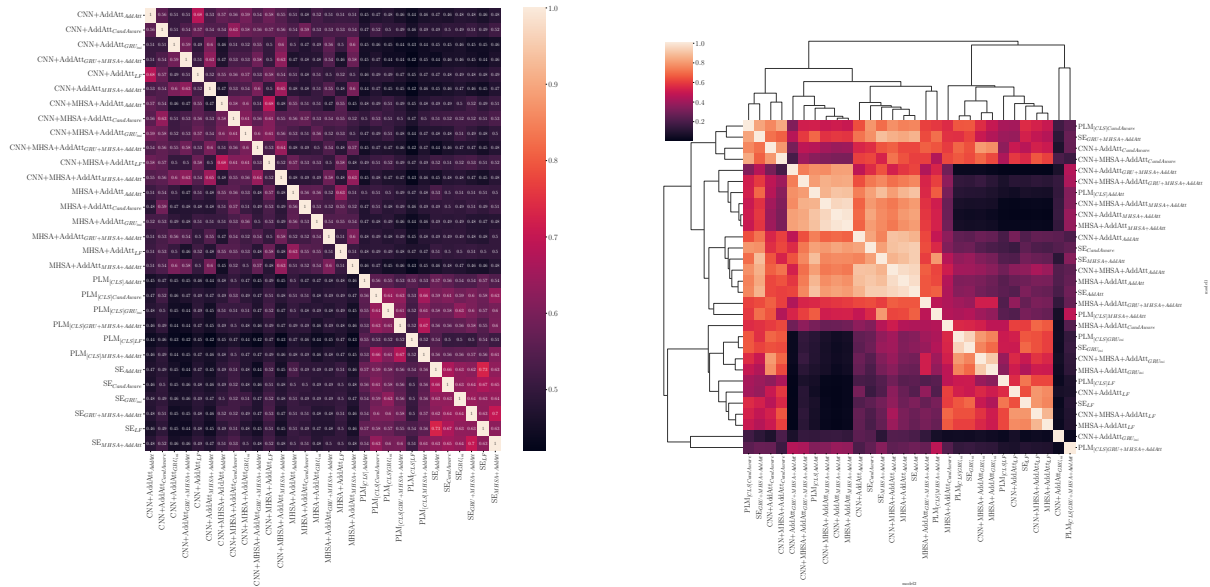


Figure 4: Ranking performance of different recommenders (nDCG@10) depending on the user encoder architecture, for different base news encoder families. The dark bars denote the ranking obtained when using a mono-feature input (i.e., title) in the news encoder, whereas the lighter bars indicate the (generally higher) scores gained with a multi-feature input (i.e., title and category), and the best multi-feature aggregation strategy per news encoder family.



(a) Jaccard similarity of top-10 recommended news.

(b) CKA similarity of user embeddings.

Figure 5: Retrieval and representational similarity of models with different user encoder architectures. Each model name denotes the base news encoder, with the user encoder architecture indicated by the subscript.

different underlying NEs. Most importantly, our findings corroborate earlier results from Iana et al. [7] and Möller and Padó [8] that the complexity of user encoders can be simplified, particularly when the bi-encoder NNR leverages language models pretrained, or even domain-specialized, on large-scale corpora, to obtain news representations.

The heatmap in Figure 5a shows the Jaccard similarity scores for the top-10 recommendations, for the different UE families, when using only the title as input to the NE.³ We exclude GRU_{con} from further analysis as it underperforms the counterpart variant GRU_{ini}. We observe that in terms of retrieval similarity, the NNRs are clustered based on the underlying NE family, regardless of the UE used. Once again, the results indicate a large overlap of recommended news (i.e., on average, of at least 7 out of 10 recommendations) for the UEs within these clusters. Moreover, we observe comparable similarity patterns across inter-family UEs for the same NE family; different NEs change only the absolute magnitude of the Jaccard similarity scores. Within intra-family clusters of NEs, the findings re-affirm that LF and AddAtt have the highest overlap in terms of the top-10 recommended articles; their generated recommendations usually differ in at most 2 or 3 items, on average. This is intuitive, as LF represents a special case of AddAtt, where the attention weights are all equal, and set to the inverse

³The results with multi-feature input are similar, and we omit them for the sake of brevity.

of the user’s history length.

We delve deeper into the retrieval similarity of UE architectures. Figure 3b shows the Jaccard similarity of LF against the other user modeling approaches for a recommender with a CNN+AddAtt-based NE, for different values of k . As in Section 5.1, the Jaccard similarity of recommended news is sensitive to the value of k , with scores converging toward 1 for larger values of k . On the one hand, the scores of sequential UEs (GRU_{ini}, GRU+MHSA+AddAtt) are clustered closely together, which can be explained by their shared sequential component. However, the retrieved articles appear to be more similar between sequential and non-sequential UEs (e.g., higher Jaccard similarity between GRU+MHSA+AddAtt and MHSA+AddAtt) across intra-family NEs, than between sequential UEs. This could be attributed to the architectural differences of the two models, among which GRU+MHSA+AddAtt employs an attention network similar to that of MHSA+AddAtt. These mixed results, combined with the better performing non-sequential UEs, call into question the efficiency of modeling the news recommendation task as a sequential recommendation problem [52].

We shift our attention to the pair-wise similarity of user embeddings generated by the different types of UEs for the users in the test set, illustrated in the heatmap of Figure 5b. We additionally perform a hierarchical clustering on the heatmap to identify clusters of similar UEs [65]. In contrast to the retrieval results, we find that the architecturally comparable families of UEs dictate the similarity of embeddings, regardless of the underlying NE used. Most surprisingly, we find that although the top-recommended news by GRU_{ini} and GRU+MHSA+AddAtt moderately overlap, their user representations are highly dissimilar. Moreover, the latent representations of AddAtt appear more similar to other attention-based UEs than with LF. This could be explained by the fact that as a particular case of AddAtt, the parameterless LF does not reshape the embedding space, as it simply computes an average of the user’s clicked news. Nonetheless, these differences in the representational similarities of UEs also do not appear to directly correlate with more dissimilar retrieval performance. This suggests that in real-world applications, the lightweight and conceptually simple LF constitutes an equally effective and more efficient alternative to AddAtt, and especially, to more complex architectures.

5.3. Key Takeaways

Following the results of our in-depth analysis of the embedding and retrieval similarity of the most prominent news and user encoder architectures, we highlight several key takeaways.

Semantic Richness is Key. Our analysis demonstrates that the semantic richness of news encoders, achieved either through multi-feature input or contextualized language models, significantly outweighs the impact of UEs. This is particularly the case when initializing news representations with large-scale PLMs. Additionally, contextualized language models can effectively capture semantic nuances, such as topical information, without heavily relying on categorical annotations. From a practical standpoint, this reduces the need for manual or automatic feature engineering, streamlining the NNR design process. We hence argue that research on news encoding should focus more on leveraging and adapting existing semantically informed, contextualized language models for the task of news recommendation, rather than on incrementally modifying existing architectures.

User Encoders Can be Considerably Simplified. Our findings show that retrieval similarity is primarily influenced by the underlying NE family, rather than the specific UE used. At the same time, simpler approaches, such as LF and AddAtt, not only result in significantly better ranked results, but their retrieved items largely overlap with those recommended by more complex UE architectures. These findings thus render simpler architectures as better and more lightweight user modeling alternatives. Additionally, the high retrieval similarity between parameter-free (i.e., LF) and parameterized (e.g., AddAtt) encoders heavily indicates that, in practice, there is little empirical justification for an additional parameterized component in the news recommender system. Furthermore, the similarity of sequential and non-sequential encoders indicates that treating news recommendation as a sequential problem might be sub-optimal. We speculate that the high item churn characteristic of news, combined with short user histories, limit the benefits of differentiating between long and short-term user preferences,

in contrast to other domains, such as movie or book recommendation. In conclusion, in line with Möller and Padó [8], we posit that user modeling should not focus exclusively on the architectural component, but instead, should pay closer attention to the users' motivations to consume certain news, on the one hand, and to collecting richer and more accurate user (relevance) feedback, on the other hand.

More Rigorous Evaluation is Needed for Better Model Selection. Our findings, along with recent research [9, 7, 8], highlight the limitations of current evaluation practices in news recommendation. By focusing solely on performance metrics, we risk overlooking critical aspects of model behavior, leading to sub-optimal component selection and incremental model advancement. Therefore, we advocate for a more comprehensive and rigorous evaluation approach. Ablation studies should consider the broader architectural context, and together with model comparisons, should extend beyond performance-based evaluation to include a more granular behavioral and representational analysis. This would provide a more nuanced understanding of model similarities and differences, guiding researchers and practitioners toward better informed model selection decisions.

6. Conclusion

Despite the central role played by encoder architectures in neural news recommenders, their advancement and understanding is generally limited to one-sided evaluation in terms of recommendation performance. In this work, we conducted a comprehensive evaluation of encoder architectures in neural news recommenders, by systematically analyzing their (i) representation similarity, (ii) overlap of generated recommendations, and (iii) overall recommendation performance. Evaluations of recommenders on standard benchmarks often reveal insignificant performance differences between compared models or among their ablated components. Consequently, our analysis of differences in representational similarity and retrieval overlap of neural news recommenders serves as a complementary evaluation tool for understanding the relationship between the architectural design, behavior, and downstream performance of models.

Our findings offer more nuanced insights into the interplay of news and user encoders, and challenge the assumption that complex encoding techniques are essential for accurate news recommendation. We demonstrate that simpler, yet equally effective architectures can yield comparable results. This underscores the importance of understanding recommenders' behavior from multiple perspectives, and of balancing model complexity with performance. Specifically, we emphasize three key takeaways: (1) the crucial role of semantic richness in news encoders, (2) the potential for simplifying user encoders without sacrificing accuracy, and (3) the need for more rigorous evaluation and ablation studies to inform architectural design choices. By fostering a more transparent and nuanced understanding of encoder architectures in neural news recommenders, we hope to guide researchers and practitioners toward more efficient and effective model designs.

Acknowledgments

The authors acknowledge support by the state of Baden-Württemberg through bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG. We also thank Fabian David Schmidt for proof-reading.

References

- [1] C. Wu, F. Wu, Y. Huang, X. Xie, Personalized news recommendation: Methods and challenges, *ACM Transactions on Information Systems* 41 (2023) 1–50. doi:<https://doi.org/10.1145/3530257>.
- [2] M. Karimi, D. Jannach, M. Jugovac, News recommender systems—survey and roads ahead, *Information Processing & Management* 54 (2018) 1203–1227. doi:<https://doi.org/10.1016/j.ipm.2018.04.008>.

- [3] S. Raza, C. Ding, News recommender system: a review of recent progress, challenges, and opportunities, *Artificial Intelligence Review* (2022) 1–52. doi:<https://doi.org/10.1007/s10462-021-10043-x>.
- [4] C. Wu, F. Wu, M. An, J. Huang, Y. Huang, X. Xie, Neural news recommendation with attentive multi-view learning, in: *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, 2019, pp. 3863–3869. doi:[10.24963/ijcai.2019/536](https://doi.org/10.24963/ijcai.2019/536).
- [5] 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. doi:[10.18653/v1/D19-1671](https://doi.org/10.18653/v1/D19-1671).
- [6] M. An, F. Wu, C. Wu, K. Zhang, Z. Liu, X. Xie, Neural news recommendation with long-and short-term user representations, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 336–345. doi:<https://doi.org/10.18653/v1/P19-1033>.
- [7] A. Iana, G. Glavas, H. Paulheim, Simplifying content-based neural news recommendation: On user modeling and training objectives, in: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 2384–2388. doi:<https://doi.org/10.1145/3539618.3592062>.
- [8] L. Möller, S. Padó, Explaining neural news recommendation with attributions onto reading histories, *ACM Transactions on Intelligent Systems and Technology* (2024). doi:<https://doi.org/10.1145/3673233>.
- [9] L. Möller, S. Padó, Understanding the relation of user and news representations in content-based neural news recommendation, *Joint Proceedings of 10th International Workshop on News Recommendation and Analytics (INRA'22) and the Third International Workshop on Investigating Learning During Web Search (IWILDS'22) co-located with the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'22) (2022)*. URL: <https://ceur-ws.org/Vol-3411/INRA-paper2.pdf>.
- [10] T. Qi, F. Wu, C. Wu, Y. Huang, News recommendation with candidate-aware user modeling, in: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 1917–1921. doi:<https://doi.org/10.1145/3477495.3531778>.
- [11] C. Wu, F. Wu, T. Qi, Y. Huang, Empowering news recommendation with pre-trained language models, in: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 1652–1656. doi:<https://doi.org/10.1145/3404835.3463069>.
- [12] S. Kornblith, M. Norouzi, H. Lee, G. Hinton, Similarity of neural network representations revisited, in: *International Conference on Machine Learning*, PMLR, 2019, pp. 3519–3529.
- [13] T. Qi, F. Wu, C. Wu, Y. Huang, X. Xie, Privacy-preserving news recommendation model learning, in: *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1423–1432. doi:[10.18653/v1/2020.findings-emnlp.128](https://doi.org/10.18653/v1/2020.findings-emnlp.128).
- [14] R. Wang, S. Wang, W. Lu, X. Peng, News recommendation via multi-interest news sequence modelling, in: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 7942–7946. doi:<https://doi.org/10.1109/ICASSP43922.2022.9747149>.
- [15] H. Wang, F. Zhang, X. Xie, M. Guo, DKN: Deep knowledge-aware network for news recommendation, in: *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 1835–1844. doi:[10.1145/3178876.3186175](https://doi.org/10.1145/3178876.3186175).
- [16] J. Li, J. Zhu, Q. Bi, G. Cai, L. Shang, Z. Dong, X. Jiang, Q. Liu, Miner: Multi-interest matching network for news recommendation, in: *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 343–352. doi:<https://doi.org/10.18653/v1/2022.findings-acl.29>.
- [17] A. Iana, F. D. Schmidt, G. Glavaš, H. Paulheim, News without borders: Domain adaptation of multilingual sentence embeddings for cross-lingual news recommendation, *arXiv preprint arXiv:2406.12634* (2024). doi:<https://doi.org/10.48550/arXiv.2406.12634>.
- [18] C. Wu, F. Wu, M. An, Y. Huang, X. Xie, Neural news recommendation with topic-aware news

- representation, in: Proceedings of the 57th Annual meeting of the association for computational linguistics, 2019, pp. 1154–1159. doi:10.18653/v1/P19-1110.
- [19] C. Wu, F. Wu, T. Qi, Y. Huang, SentiRec: Sentiment diversity-aware neural news recommendation, in: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, 2020, pp. 44–53. URL: <https://aclanthology.org/2020.aacl-main.6>.
- [20] J. Xun, S. Zhang, Z. Zhao, J. Zhu, Q. Zhang, J. Li, X. He, X. He, T.-S. Chua, F. Wu, Why do we click: visual impression-aware news recommendation, in: Proceedings of the 29th ACM international conference on multimedia, 2021, pp. 3881–3890. doi:<https://doi.org/10.1145/3474085.3475514>.
- [21] T. Qi, F. Wu, C. Wu, Y. Huang, PP-Rec: News recommendation with personalized user interest and time-aware news popularity, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 5457–5467. doi:10.18653/v1/2021.acl-long.424.
- [22] C. Wu, F. Wu, T. Qi, Y. Huang, Mm-rec: multimodal news recommendation, arXiv preprint arXiv:2104.07407 (2021). doi:<https://doi.org/10.48550/arXiv.2104.07407>.
- [23] C. Wu, F. Wu, M. An, T. Qi, J. Huang, Y. Huang, X. Xie, Neural news recommendation with heterogeneous user behavior, 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. 4874–4883. doi:<https://doi.org/10.18653/v1/D19-1493>.
- [24] C. Wu, F. Wu, T. Qi, Q. Liu, X. Tian, J. Li, W. He, Y. Huang, X. Xie, Feedrec: News feed recommendation with various user feedbacks, in: Proceedings of the ACM Web Conference 2022, 2022, pp. 2088–2097. doi:10.1145/3485447.3512082.
- [25] M. Klabunde, T. Schumacher, M. Strohmaier, F. Lemmerich, Similarity of neural network models: A survey of functional and representational measures, arXiv preprint arXiv:2305.06329 (2023). doi:<https://doi.org/10.48550/arXiv.2305.06329>.
- [26] J. Wu, Y. Belinkov, H. Sajjad, N. Durrani, F. Dalvi, J. Glass, Similarity analysis of contextual word representation models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 4638–4655. doi:<https://doi.org/10.18653/v1/2020.acl-main.422>.
- [27] M. Klabunde, M. B. Amor, M. Granitzer, F. Lemmerich, Towards measuring representational similarity of large language models, in: UniReps: the First Workshop on Unifying Representations in Neural Models, 2023.
- [28] D. Brown, C. Godfrey, N. Konz, J. Tu, H. Kvinge, Understanding the inner-workings of language models through representation dissimilarity, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 6543–6558. doi:<https://doi.org/10.18653/v1/2023.emnlp-main.403>.
- [29] M. Freestone, S. K. K. Santu, Word embeddings revisited: Do llms offer something new?, arXiv preprint arXiv:2402.11094 (2024). doi:<https://doi.org/10.48550/arXiv.2402.11094>.
- [30] L. Caspari, K. G. Dastidar, S. Zerhoubi, J. Mitrovic, M. Granitzer, Beyond benchmarks: Evaluating embedding model similarity for retrieval augmented generation systems, arXiv preprint arXiv:2407.08275 (2024). doi:<https://doi.org/10.48550/arXiv.2407.08275>.
- [31] P.-S. Huang, X. He, J. Gao, L. Deng, A. Acero, L. Heck, Learning deep structured semantic models for web search using clickthrough data, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, 2013, pp. 2333–2338. doi:<https://doi.org/10.1145/2505515.2505665>.
- [32] C. Wu, F. Wu, Y. Huang, Rethinking InfoNCE: How many negative samples do you need?, in: L. D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22, International Joint Conferences on Artificial Intelligence Organization, 2022, pp. 2509–2515. doi:10.24963/ijcai.2022/348.
- [33] A. Iana, G. Glavaš, H. Paulheim, Train once, use flexibly: A modular framework for multi-aspect neural news recommendation, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the As-

- sociation for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 9555–9571. URL: <https://aclanthology.org/2024.findings-emnlp.558>.
- [34] R. Liu, B. Yin, Z. Cao, Q. Xia, Y. Chen, D. Zhang, Perconet: News recommendation with explicit persona and contrastive learning, arXiv preprint arXiv:2304.07923 (2023). doi:<https://doi.org/10.48550/arXiv.2304.07923>.
- [35] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1746–1751. doi:[10.3115/v1/D14-1181](https://doi.org/10.3115/v1/D14-1181).
- [36] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, ICLR (2014).
- [37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 6000–6010. URL: <https://dl.acm.org/doi/abs/10.5555/3295222.3295349>.
- [38] K. Cho, B. van Merriënboer, Ç. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using rnn encoder–decoder for statistical machine translation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724–1734. doi:[10.3115/v1/D14-1179](https://doi.org/10.3115/v1/D14-1179).
- [39] H.-S. Sheu, S. Li, Context-aware graph embedding for session-based news recommendation, in: Proceedings of the 14th ACM Conference on Recommender Systems, 2020, pp. 657–662. doi:<https://doi.org/10.1145/3383313.3418477>.
- [40] T. Santosh, A. Saha, N. Ganguly, Mvl: Multi-view learning for news recommendation, in: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2020, pp. 1873–1876. doi:<https://doi.org/10.1145/3397271.3401294>.
- [41] J. Gao, X. Xin, J. Liu, R. Wang, J. Lu, B. Li, X. Fan, P. Guo, Fine-grained deep knowledge-aware network for news recommendation with self-attention, in: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, 2018, pp. 81–88. doi:<https://doi.org/10.1109/WI.2018.0-104>.
- [42] C. Wu, F. Wu, T. Qi, Y. Huang, User modeling with click preference and reading satisfaction for news recommendation, in: IJCAI, 2020, pp. 3023–3029. doi:<https://doi.org/10.24963/ijcai.2020/418>.
- [43] S. Ge, C. Wu, F. Wu, T. Qi, Y. Huang, Graph enhanced representation learning for news recommendation, in: Proceedings of the Web Conference 2020, 2020, pp. 2863–2869. doi:<https://doi.org/10.1145/3366423.3380050>.
- [44] D. H. Tran, S. Hamad, M. Zaib, A. Aljubairy, Q. Z. Sheng, W. E. Zhang, N. H. Tran, N. L. D. Khoa, Deep news recommendation with contextual user profiling and multifaceted article representation, in: Web Information Systems Engineering–WISE 2021: 22nd International Conference on Web Information Systems Engineering, WISE 2021, Melbourne, VIC, Australia, October 26–29, 2021, Proceedings, Part II 22, Springer, 2021, pp. 237–251. doi:https://doi.org/10.1007/978-3-030-91560-5_17.
- [45] C. Wu, F. Wu, T. Qi, Y. Huang, Two birds with one stone: Unified model learning for both recall and ranking in news recommendation, in: Findings of the Association for Computational Linguistics: ACL 2022, 2022, pp. 3474–3480. doi:<https://doi.org/10.18653/v1/2022.findings-acl.274>.
- [46] C. Wu, F. Wu, Y. Huang, X. Xie, User-as-graph: User modeling with heterogeneous graph pooling for news recommendation, in: IJCAI, 2021, pp. 1624–1630. doi:<https://doi.org/10.24963/ijcai.2021/224>.
- [47] C. Wu, F. Wu, X. Wang, Y. Huang, X. Xie, Fairness-aware news recommendation with decomposed adversarial learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 4462–4469. doi:<https://doi.org/10.1609/aaai.v35i5.16573>.
- [48] C. Wu, F. Wu, Y. Huang, X. Xie, Neural news recommendation with negative feedback, CCF Transactions on Pervasive Computing and Interaction 2 (2020) 178–188. doi:<https://doi.org/10.1007/s42486-020-00044-0>.
- [49] Q. Zhang, Q. Jia, C. Wang, J. Li, Z. Wang, X. He, Amm: Attentive multi-field matching for news

- recommendation, in: Proceedings of the 44th international ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1588–1592. doi:<https://doi.org/10.1145/3404835.3463232>.
- [50] Q. Zhang, J. Li, Q. Jia, C. Wang, J. Zhu, Z. Wang, X. He, Unbert: User-news matching bert for news recommendation., in: IJCAI, volume 21, 2021, pp. 3356–3362. doi:<https://doi.org/10.24963/ijcai.2021/462>.
- [51] Q. Jia, J. Li, Q. Zhang, X. He, J. Zhu, Rmbert: News recommendation via recurrent reasoning memory network over bert, in: Proceedings of the 44th international ACM SIGIR Conference on Research and Development in Information Retrieval, 2021, pp. 1773–1777. doi:<https://doi.org/10.1145/3404835.3463234>.
- [52] C. Wu, F. Wu, T. Qi, C. Li, Y. Huang, Is news recommendation a sequential recommendation task?, in: Proceedings of the 45th international ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 2382–2386. doi:<https://doi.org/10.1145/3477495.3531862>.
- [53] K. Shivaram, P. Liu, M. Shapiro, M. Bilgic, A. Culotta, Reducing cross-topic political homogenization in content-based news recommendation, in: Proceedings of the 16th ACM Conference on Recommender Systems, 2022, pp. 220–228. doi:<https://doi.org/10.1145/3523227.3546782>.
- [54] Y. Sun, F. Yi, C. Zeng, B. Li, P. He, J. Qiao, Y. Zhou, A hybrid approach to news recommendation based on knowledge graph and long short-term user preferences, in: 2021 IEEE International Conference on Services Computing (SCC), IEEE, 2021, pp. 165–173. doi:<https://doi.org/10.1109/SCC53864.2021.00029>.
- [55] S. Raza, C. Ding, Deep dynamic neural network to trade-off between accuracy and diversity in a news recommender system, arXiv preprint arXiv:2103.08458 (2021). doi:<https://doi.org/10.48550/arXiv.2103.08458>.
- [56] S. Han, H. Huang, J. Liu, Neural news recommendation with event extraction, arXiv preprint arXiv:2111.05068 (2021). doi:<https://doi.org/10.48550/arXiv.2111.05068>.
- [57] D. Liu, J. Lian, S. Wang, Y. Qiao, J.-H. Chen, G. Sun, X. Xie, KRED: Knowledge-aware document representation for news recommendations, in: Proceedings of the 14th ACM Conference on Recommender Systems, 2020, pp. 200–209. doi:[10.1145/3383313.3412237](https://doi.org/10.1145/3383313.3412237).
- [58] X. Zhang, Q. Yang, D. Xu, Combining explicit entity graph with implicit text information for news recommendation, in: Companion Proceedings of the Web Conference 2021, 2021, pp. 412–416. doi:<https://doi.org/10.1145/3442442.3452329>.
- [59] A. Gretton, O. Bousquet, A. Smola, B. Schölkopf, Measuring statistical dependence with hilbert-schmidt norms, in: International Conference on Algorithmic Learning Theory, Springer, 2005, pp. 63–77.
- [60] 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. doi:<https://doi.org/10.18653/v1/2020.acl-main.331>.
- [61] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543. doi:[10.3115/v1/D14-1162](https://doi.org/10.3115/v1/D14-1162).
- [62] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: <https://arxiv.org/abs/1907.11692>. arXiv:1907.11692.
- [63] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, ICLR (2014).
- [64] A. Iana, G. Glavaš, H. Paulheim, Newsreclib: A pytorch-lightning library for neural news recommendation, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2023, pp. 296–310. doi:<https://doi.org/10.18653/v1/2023.emnlp-demo.26>.
- [65] M. L. Waskom, Seaborn: statistical data visualization, Journal of Open Source Software 6 (2021) 3021. doi:<https://doi.org/10.21105/joss.03021>.