Simplifying Content-Based Neural News Recommendation: On User Modeling and Training Objectives

Andreea Iana
University of Mannheim
Mannheim, Germany
andreea.iana@uni-mannheim.de

Goran Glavaš
CAIDAS, University of Würzburg
Würzburg, Germany
goran.glavas@uni-wuerzburg.de

Heiko Paulheim
University of Mannheim
Mannheim, Germany
heiko.paulheim@uni-mannheim.de

ABSTRACT

The advent of personalized news recommendation has given rise to increasingly complex recommender architectures. Most neural news recommenders rely on user click behavior and typically introduce dedicated user encoders that aggregate the content of clicked news into user embeddings (*early fusion*). These models are predominantly trained with standard point-wise classification objectives. The existing body of work exhibits two main shortcomings: (1) despite general design homogeneity, direct comparisons between models are hindered by varying evaluation datasets and protocols; (2) it leaves alternative model designs and training objectives vastly unexplored. In this work, we present a unified framework for news recommendation, allowing for a systematic and fair comparison of news recommenders across several crucial design dimensions: (i) *candidate-awareness in user modeling*, (ii) *click behavior fusion*, and (iii) *training objectives*. Our findings challenge the status quo in neural news recommendation. We show that replacing sizable user encoders with parameter-efficient dot products between candidate and clicked news embeddings (*late fusion*) often yields substantial performance gains. Moreover, our results render contrastive training a viable alternative to point-wise classification objectives.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

neural news recommendation, user modeling, late fusion, training objectives, contrastive learning, evaluation

ACM Reference Format:
Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2023. 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 (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3539618.3592062

1 INTRODUCTION

In recent years, content-based news recommendation has seen increasingly complex neural recommender architectures that aim to customize suggestions to users' interests [13, 34]. Most neural news recommendation (NNR) models commonly comprise (i) a dedicated news encoder (NE) and (ii) a user encoder (UE) [34]. NEs – instantiated as a convolutional network [26, 29, 31], self-attention network [19, 32, 35], graph attention network [18], or pretrained transformer network [36, 41] – convert input features (e.g., titles, categories, entities) into the news embedding. UEs aggregate embeddings of clicked news into a user-level representation by means of sequential [1, 21, 27] or attentive [26, 29, 32] encoders that contextualize embeddings of clicked news based on patterns in clicking behavior [1, 15, 37]. We dub this predominant paradigm *early fusion* (EF) because it aggregates representations of clicked news (i.e., builds user representation) before comparison with the candidate.

Most NNR models encode users and candidate news separately, in a *candidate-agnostic* manner [1, 29, 32]. *Candidate-aware* models [20, 22, 25, 42], in contrast, acknowledge that not all clicked news are equally informative w.r.t. the relevance of the candidate (e.g., a candidate is often representative of only a subset of a user's preferences), and contextualize representations of clicked news with the embedding of the candidate in user-level aggregation with UE. Finally, the candidate's embedding (output of NE) is compared against the user embedding (output of UE): the candidate's recommendation score is computed directly as the dot product of the two embeddings [29] or with a feed-forward scorer [26]. NNR models are predominantly trained via standard classification objectives [26, 29, 32, 36] with negative sampling [7, 30].

The existing body of work has two main shortcomings. First, despite general design homogeneity, direct comparisons between recent NNRs are hindered by lack of transparency and adoption of ad-hoc evaluation protocols [8, 23]. In particular, a vast majority of personalized news recommenders are evaluated on proprietary datasets (e.g., MSN News [29, 32], Bing News [26], NewsApp [20]). Even the few models evaluated using the publicly available datasets such as Adressa [6] or MIND [38] cannot be directly compared due to different dataset splits and evaluation protocols (e.g., model selection strategy) [5, 27, 36, 42]. Secondly, simpler and arguably more intuitive design alternatives have largely been left unexplored. First, the existing work adopts EF as default architecture, proposing increasingly complex user encoding components [1, 20], often with little empirical justification for added complexity. Second, only a small fraction of NNRs leverage contrastive learning objectives [33, 41], despite such training criteria being proven highly effective in closely related retrieval and recommendation tasks [14, 28, 39, 40].

In this work, we remedy the above shortcomings of current NNRs and shed new light on user modeling and training objectives.[1] **1)** Concretely, we introduce a unified framework for neural news

[1] **Disclaimer:** In this work we focus exclusively on NNR models that do not resort to graph-based modeling of relations between users.

This work is licensed under a Creative Commons Attribution International 4.0 License.

SIGIR '23, July 23–27, 2023, Taipei, Taiwan
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9408-6/23/07.
https://doi.org/10.1145/3539618.3592062

recommendation, facilitating systematic and fair comparison of NNR models across three crucial design dimensions: (i) *candidate-awareness in user modeling*, (ii) *click behavior fusion*, and (iii) *training objectives*. **2)** We propose to replace user modeling with complex user encoders (i.e., *early fusion*) with simple pooling of dot-product scores between candidate and clicked news embeddings (i.e., *late fusion*). We show that, despite conceptual simplicity, LF brings substantial performance gains over EF-based NNR, rendering complex UEs empirically unjustified. **3)** Finally, we demonstrate the benefits of supervised contrastive training as a viable alternative to pointwise classification. Our work fundamentally challenges the status quo of NNR by introducing simpler and more effective alternatives to the established paradigm based on complex user modeling.

2 METHODOLOGY

Figure 1 depicts our unified evaluation framework for NNR, focusing on three critical dimensions of news recommendation. Given input data, comprising news and user behaviors, we analyze (i) candidate-agnostic (C-AG) vs. candidate-aware (C-AW) user modeling under (ii) two click behavior fusion strategies, namely EF and LF, where each model can be (iii) trained by minimizing either the standard cross-entropy loss (CE) or a supervised contrastive objective (SCL). Next, we describe the models selected for evaluation and formalize the concrete design choices.

2.1 User Modeling

Candidate-Agnostic (C-AG) Models. For these models, the UE produces the user embedding from embeddings of clicked news without contextualization against the candidate. We evaluate the following C-AG models, mutually differing in their NE component (i.e., how they embed the clicked news): (1) *NPA* [30] uses a personalized attention module to aggregate the representations of the users' clicked news, with projected embeddings of the users IDs as attention queries; (2) *NAML* [29] uses additive attention [2] to encode users' preferences; (3) *NRMS* [32] learns user representations with a two-layer encoder that consists of multi-head self-attention [24] and additive attention; (4) *LSTUR* [38] learns user representations with recurrent networks: a short-term user embedding is produced from the clicked news with a GRU [4], and combined with a long-term embedding, consisting of a randomly initialized and fine-tuned part; the final user embedding is then obtained either (i) as the final hidden state of the short-term GRU, initialized with the long-term embedding (*LSTUR$_{ini}$*), or (ii) by simply concatenating the short- and long-term user embeddings (*LSTUR$_{con}$*); (5) *CenNewsRec* [21] adopts a similar UE architecture as LSTUR, but learns long-term user vectors from clicked news using a sequence of multi-head self-attention and attentive pooling networks, as opposed to storing an explicit embedding per user; (6) *MINS* [27] encodes users through a combination of multi-head self-attention, multi-channel GRU-based recurrent network, and additive attention.

Candidate-Aware (C-AW) Models. UEs in candidate-agnostic models produce the same user embedding, regardless of the content of the candidate news. In contrast, UEs of candidate-aware models, two of which we include in our empirical analysis, produce user embeddings dependent on the candidate. (7) *DKN* [26] computes candidate-aware representations of users as the weighted sum of their clicked news embeddings, with weights being produced by an attention network that takes as input the embeddings of the candidate and of the clicked news, as produced by the NE. More recently, (8) *CAUM* [20] combines (i) a candidate-aware self-attention network to model long-range dependencies between clicked news, conditioned on the candidate, and (ii) a candidate-aware convolutional network (CNN) to capture short-term user interests from adjacent clicks, again conditioned with the candidate's content; the candidate-aware user embedding is finally obtained by attending over the long-range and short-term representations.

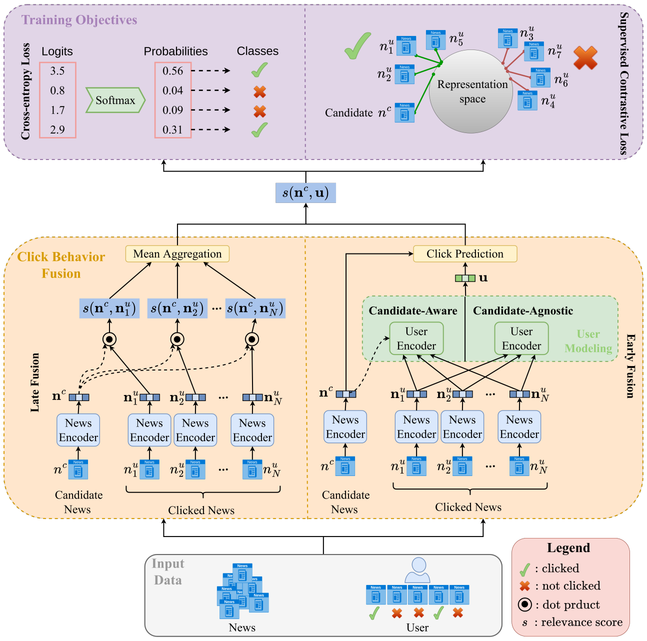

Figure 1: Illustration of the unified NNR framework, focusing on three crucial design dimensions: (i) candidate-awareness in user modeling (green box), (ii) click behavior fusion (orange box), and (iii) training objectives (purple box).

News Encoders. The NNR models included in our evaluation primarily use news titles as input, which they typically embed via pretrained word embeddings [17]. NAML, LSTUR, MINS, and CAUM additionally leverage category information, with categories embedded with a linear layer. CAUM additionally encodes title entities and DKN exploits knowledge graph embeddings [9]. The shallow word and entity embeddings are contextualized either using a combination of multi-head self-attention (in NRMS, MINS, CAUM), or a sequence of CNN [11] and additive attention networks (in NAML, LSTUR). NPA [30] also utilizes a CNN to contextualize word embeddings, followed by a personalized attention module, analogous to the one used in its user encoder, whereas DKN employs a word-entity-aligned knowledge-aware CNN [26]. CenNewsRec [21] combines the CNN network with multi-head self-attention and additive attention modules. Models with multiple feature vectors produce final news embeddings by simply concatenating them (LSTUR, CAUM), or by attending over them (NAML, MINS).

2.2 Click Behavior Fusion

We question whether the design and computational complexity of *early fusion* (EF), i.e., existence of dedicated user encoders in state-of-the-art NNR models, is justified. To this end, we propose, as a lightweight alternative, the *late fusion* (LF) approach that replaces the elaborate user encoders with mean-pooling of dot-product scores between the embedding of the candidate n^c and the embeddings of the clicked news n_i^u. Given a candidate news n^c and a sequence of news clicked by the user $H = n_1^u, ..., n_N^u$, we compute the relevance score of the candidate news with regards to the user u's history as $s(\mathbf{n}^c, u) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{n}^c \cdot \mathbf{n}_i^u$, where $\mathbf{n}$ denotes the embedding of a news learned by the news encoder and N the history length.

Although LF suggests that explicitly encoding user behavior may not be necessary for click prediction, user embeddings are still needed in collaborative-filtering models [13]. Note that the LF formulation above is equivalent to the dot product between the candidate embedding $\mathbf{n}^c$ and the mean of embeddings of the user's clicked news $\mathbf{n}_i^u$, $s(\mathbf{n}^c, u) = \mathbf{n}^c \cdot \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{n}_i^u\right)$. This means that LF can also seamlessly provide user embeddings (simply as averages of clicked news embeddings) if needed. LF can thus been seen as a *parameterless* user encoder, i.e., a computationally efficient alternative to complex parameterized UEs in existing EF models. Because (i) we produce embeddings of candidate and clicked news independently, and (ii) yield user embeddings as averages of clicked news embeddings, LF models are candidate agnostic (C-AG).

2.3 Training Objectives

The vast majority of existing NNR work, regardless of the concrete user modeling architecture, tunes the parameters by minimizing the arguably most straightforward classification objective, cross-entropy loss (with negative sampling; see Figure 1), and largely fails to explore effective alternatives, foremost contrastive objectives [16, 33]. This prevents understanding of models effectiveness under different training regimes. We address this limitation by training all models (see §2.1) not only with (1) common cross-entropy loss (with negative sampling), but also via (2) a contrastive learning objective, in particular supervised contrastive loss [10].

3 EXPERIMENTAL SETUP

Data. We conduct experiments on the MINDsmall and MINDlarge datasets, introduced by Wu et al. [38]. Table 1 summarizes their main statistics. Since Wu et al. [38] do not release test set labels, we use the respective validation portions for testing, and split the respective training sets into temporally disjoint training (first four days of data) and validation portions (the last day).

Implementation and Optimization Details. We use 300-dimensional pretrained Glove embeddings [17] and 100-dimensional TransE embeddings [3] pretrained on Wikidata to initialize respectively the word and entity embeddings of the NNR models under comparison. We set the maximum history length to 50. Following Wu et al. [33], our negative sampling creates four negatives per positive example. We find the optimal temperature for SCL using the validation performance, sweeping the interval $[0.08, 0.3]$ with a 0.02 step. We train with batch size of 512 for all C-AG models, 256 for DKN and only 64 for CAUM (due to computational limitations). We set all other model-specific hyperparameters, to optimal values reported in the respective papers. We train all models with mixed precision, under a fixed computational budget: for 25 epochs on MINDsmall and 10 epochs on MINDlarge. We optimize with the Adam algorithm [12], with the learning rate set to $1e$-4. We repeat each experiment five times (with different random seeds) and report averages (and std. deviation) for common metrics: AUC, MRR, nDCG@5, and nDCG@10. Each model is trained on a single NVIDIA Tesla V100 GPU with 32GB memory. Our implementation is publicly available.[2]

Table 1: Statistics of the MINDsmall and MINDlarge datasets.

	MINDsmall		MINDlarge	
Statistic	Train	Test	Train	Test
# News	51,282	42,416	101,527	72,023
# Users	49,108	48,593	698,365	248,973
# Impressions	153,727	70,938	2,186,683	365,201
# Categories	17	17	18	17
# Subcategories	264	252	285	269

4 RESULTS AND DISCUSSION

Table 2 shows the performance on MINDsmall and MINDlarge for both C-AG (NPA, NAML, NRMS, LSTUR, CenNewsRec, and MINS) and C-AW models (DKN, CAUM), under four different configurations of our comparative evaluation framework: (i) user modeling with EF vs. LF, combined with (2) training with CE vs. SCL objective. We next dissect the results along the three axes of our framework (§2): user modeling, click behavior fusion, and training objectives.

Candidate-Agnostic vs. Candidate-Aware NNRs. We analyze C-AG vs. C-AW models under their default EF configuration, since with LF all models become candidate-agnostic. CAUM, with the most complex and candidate-aware UE, generally outperforms all other models under both training regimes (CE and SCL) and for most evaluation metrics. The gaps are particularly prominent on the large training dataset, MINDlarge, w.r.t. the AUC metric. This result alone could mislead to a conclusion that more complex, candidate-aware user modeling is necessary for better recommendation. The fact that (1) DKN, as the other C-AW model in our evaluation – generally performs much worse than C-AG models, as well as that (2) our LF models variants with trivial, parameterless UEs match or surpass the performance of CAUM with EF, undermine this conclusion. With the exception of DKN, all other models exhibit better performance when trained on the larger MINDlarge dataset. NAML and MINS are, however, competitive (except w.r.t. AUC metric) on MINDsmall, but fall behind CAUM on MINDlarge, suggesting that CAUM's elaborate UE benefits the most from more training data.

One confounding factor that we do not control for, however, and which warrants a mindful comparison of the results, is that models differ not just in UE, but also in NE components, i.e., w.r.t. how they encode news and which features they use as input. For example, NAML and MINS, with an identical NE, achieve similar performance on MINDsmall. On MINDlarge, however, the more complex UE of MINS brings substantial gains over the simpler UE of NAML (but only under standard EF fusion and CE training).

[2] Code available at: https://github.com/andreeaiana/simplifying_nnr

Table 2: Recommendation performance of the compared models under combinations of click behavior fusion (CBF), and training objectives. We report averages and standard deviations across five different runs.

		MINDsmall								MINDlarge							
		AUC		MRR		nDCG@5		nDCG@10		AUC		MRR		nDCG@5		nDCG@10	
Model	CBF	CE	SCL	CE	SCL	CE	SCL	CE	SCL	CE	SCL	CE	SCL	CE	SCL	CE	SCL
NPA	EF	54.7±0.6	56.5±0.7	29.0±0.7	28.4±0.6	26.9±0.8	26.6±0.6	33.2±0.8	32.6±0.4	56.8±0.2	58.1±0.8	31.4±0.5	30.0±0.6	29.5±0.4	27.7±0.6	35.9±0.4	34.3±0.6
	LF	55.1±0.9	57.3±1.2	28.6±0.3	27.5±1.0	26.4±0.4	25.5±1.1	32.9±0.4	31.8±0.9	61.2±0.5	58.7±0.8	32.1±0.6	28.3±0.5	30.2±0.7	26.0±0.7	36.6±0.6	32.7±0.7
NAML	EF	50.1±0.0	57.1±1.1	33.6±0.5	32.2±0.7	31.6±0.6	30.4±0.7	38.0±0.5	36.8±0.7	50.1±0.0	60.4±0.6	33.2±0.4	34.2±0.4	31.3±0.5	32.5±0.3	37.9±0.4	38.9±0.3
	LF	50.0±0.0	62.7±0.5	33.7±0.8	32.0±0.7	31.8±0.9	30.3±0.7	38.1±0.8	36.6±0.6	50.0±0.0	65.4±0.5	32.7±0.5	33.5±0.5	31.0±0.5	31.7±0.5	37.6±0.4	38.2±0.4
NRMS	EF	52.6±1.3	59.9±0.6	27.6±0.8	29.2±0.7	25.7±0.5	27.2±0.9	32.3±0.5	33.7±0.7	54.6±1.4	62.8±0.7	31.9±1.0	32.4±0.5	30.0±1.1	30.5±0.7	36.6±1.0	36.9±0.6
	LF	58.9±1.0	60.2±1.1	31.8±0.7	30.7±0.6	29.9±0.7	28.7±0.6	36.3±0.6	35.1±0.6	56.1±2.1	63.6±1.1	32.9±0.7	32.4±0.6	31.7±1.1	30.6±0.8	37.8±0.4	37.1±0.7
LSTUR	EF_{ini}	53.5±1.2	55.4±0.5	29.6±0.5	28.1±0.7	28.0±0.5	26.5±0.7	34.4±0.4	32.9±0.6	50.0±0.1	56.9±1.5	32.5±2.4	31.3±1.5	30.9±2.4	29.8±1.8	37.4±2.4	36.2±1.6
	EF_{con}	50.2±0.0	59.8±1.4	31.8±0.7	31.3±0.8	30.1±0.8	30.3±1.3	36.4±0.7	36.2±0.7	51.4±0.4	54.3±0.4	27.7±0.4	26.5±0.2	25.9±0.5	24.6±0.2	32.3±0.5	31.1±0.2
	LF	50.0±0.0	50.0±0.0	33.8±0.6	33.9±0.6	31.9±0.7	32.0±0.7	38.0±0.6	38.1±0.6	50.0±0.0	50.0±0.0	34.7±0.6	33.1±0.2	33.2±0.6	31.6±0.4	39.2±0.5	37.7±0.3
CenNewsRec	EF	54.0±0.8	60.0±0.4	28.3±0.5	30.6±0.8	26.5±0.4	28.5±0.8	32.9±0.3	34.8±0.7	56.4±0.8	64.5±0.4	33.7±0.3	33.5±0.4	31.9±0.3	31.8±0.5	38.3±0.2	38.1±0.4
	LF	59.3±0.6	61.9±0.7	32.8±0.8	32.2±0.8	30.9±0.8	30.4±0.8	37.1±0.6	36.6±0.7	53.3±0.7	64.2±0.6	33.2±0.4	33.3±0.4	31.4±0.5	31.7±0.4	37.9±0.4	38.1±0.4
MINS	EF	50.6±0.3	62.9±1.7	33.7±1.0	32.4±0.3	31.9±1.1	30.7±0.4	38.3±0.9	37.1±0.3	51.7±0.2	65.8±0.5	34.3±0.2	34.4±0.5	32.5±0.4	32.6±0.5	39.1±0.4	39.1±0.5
	LF	59.1±1.2	64.2±0.7	35.0±0.5	34.1±0.6	33.2±0.6	32.3±0.7	39.4±0.6	38.5±0.6	53.8±0.6	66.7±0.8	34.9±0.2	34.8±0.7	33.0±0.2	33.1±0.7	39.5±0.2	39.6±0.6
DKN	EF	50.0±0.0	51.0±2.3	26.4±0.6	25.9±0.9	24.4±0.7	23.9±1.1	31.0±0.6	30.5±0.9	50.0±0.0	50.0±0.0	25.2±0.4	24.8±0.3	23.4±0.7	22.6±0.3	30.0±0.5	29.1±0.3
	LF	50.0±0.0	50.0±0.0	27.5±0.6	26.4±0.8	25.0±0.5	24.0±0.8	31.7±0.6	30.8±0.8	50.0±0.0	50.0±0.0	29.1±0.4	27.8±1.1	26.3±0.3	25.4±1.0	33.2±0.4	32.1±1.0
CAUM	EF	61.4±1.0	63.2±0.9	33.8±0.6	33.7±0.8	32.0±0.6	31.8±0.9	38.4±0.5	38.2±0.8	67.1±0.8	66.4±0.9	35.3±0.5	35.1±0.5	33.6±0.6	33.4±0.6	40.1±0.5	39.9±0.5
	LF	62.4±0.8	63.5±0.8	33.7±0.6	33.7±0.7	31.8±0.5	31.8±0.8	38.2±0.5	38.0±0.7	53.1±0.3	65.9±0.2	34.5±0.4	34.5±0.1	32.6±0.3	32.8±0.1	39.2±0.3	39.3±0.1

Early vs. Late Click Behavior Fusion. Replacing complex EF-based UEs with the simple parameterless LF that we propose brings substantial performance gains across the board. Averaged across all models and both training objectives, LF brings massive gains of 5.58 and 4.63 MRR points on MINDsmall and MINDlarge, respectively. Equally importantly, with LF – i.e., with the same parameterless UE – models exhibit mutually much more similar performance than under EF, with other models generally closing the gap to CAUM. This suggests that LF makes differences in NE architectures across models less consequential, thus not only simplifying UE with parameterless averaging of clicked news embeddings, but also allowing for simpler news encoders.

Cross-Entropy vs. Supervised Contrastive Loss. Overall, we find SCL to be a viable alternative to the common cross-entropy based classification with negative sampling (compare columns CE and SCL across evaluation metrics in Table 2). SCL brings large gains over CE in terms of AUC (+8.26 points on MINDsmall and +12.14 points on MINDlarge, averaged across all models, in both EF and LF variants). This suggests that, SCL leads to better separation of clicked and not clicked news in the representation space. In contrast, SCL falls slightly behind CE according to ranking measures, MRR and nDCG (-1.54 and -1.78 MRR points on MINDsmall and MINDlarge, respectively). We hypothesize that this is because of hard negatives – news not clicked by the user that resemble user's clicked news – for which CE more directly signals irrelevance: these likely become highly-ranked false positives for SCL-trained models.

Model Size. Finally, we quantify the reduction in model parameters that LF brings w.r.t. EF. Figure 2 shows the number of trainable parameters in original EF configurations, on MINDsmall.[3] While the NE accounts for the majority of parameters in most models, the plot shows that the proportion of UE parameters is non-negligible for several models, and largest by a wide margin for LSTUR. With a parameterless UE, along with performance gains, LF brings a relative reduction of model size of 14.7%, 18.1%, and massive 82.3% for CenNewsRec, CAUM, and $\text{LSTUR}_{\text{ini}}$, respectively.

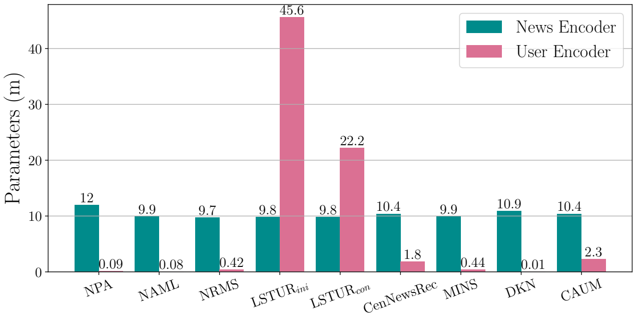

Figure 2: Number of model parameters (in millions).

5 CONCLUSION

Rapid development of personalized neural news recommenders hinders fair comparative model evaluations and systematic analyses of design choices. In this work we introduce a unified framework for neural news recommendation focusing on three crucial design dimensions of NNR: (i) candidate-awareness in user modeling, (ii) click behavior fusion, and (iii) training objectives. Extensive evaluation of a wide range of recent state-of-the-art models reveals that NNR can be drastically simplified: replacing complex user encoders with parameterless aggregation of clicked news embeddings brings substantial performance gains across the board, reducing at the same time model complexity. Further, we show that contrastive learning can be a viable alternative to standard classification-based (cross-entropy) loss. We hope that our findings will inspire more transparent NNR evaluation, including systematic model ablations to uncover the components that drive the performance.

ACKNOWLEDGMENTS

The authors acknowledge support by the state of Baden-Württemberg through bwHPC. Andreea Iana was supported by the ReNewRS project grant, which is funded by the Baden-Württemberg Stiftung in the Responsible Artificial Intelligence program.

[3] For some models, e.g., LSTUR with its user embedding matrix, the number of parameters depends on the size of the training data.

REFERENCES

[1] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural News Recommendation with Long- and Short-term User Representations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 336–345.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *ICLR* (2014).

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, Vol. 26.

[4] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. 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)*. 1724–1734.

[5] Shansan Gong and Kenny Q Zhu. 2022. Positive, Negative and Neutral: Modeling Implicit Feedback in Session-based News Recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1185–1195.

[6] Jon Atle Gulla, Lemei Zhang, Peng Liu, Özlem Özgöbek, and Xiaomeng Su. 2017. The Adressa Dataset for News Recommendation. In *Proceedings of the International Conference on Web Intelligence*. 1042–1048.

[7] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning Deep Structured Semantic Models for Web Search using Clickthrough Data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2333–2338.

[8] Andreea Iana, Mehwish Alam, and Heiko Paulheim. 2022. A Survey on Knowledge-aware News Recommender Systems. *Semantic Web* Preprint (2022), 1–62.

[9] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge Graph Embedding via Dynamic Mapping Matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 687–696.

[10] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vol. 33. 18661–18673.

[11] Yoon Kim. 2014. 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, 1746–1751. https://doi.org/10.3115/v1/D14-1181

[12] Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *ICLR* (2014).

[13] Miaomiao Li and Licheng Wang. 2019. A Survey on Personalized News Recommendation Technology. *IEEE Access* 7 (2019), 145861–145879.

[14] Yizhi Li, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2021. More Robust Dense Retrieval with Contrastive Dual Learning. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 287–296.

[15] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based News Recommendation for Millions of Users. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1933–1942.

[16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748* (2018).

[17] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[18] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021. Personalized News Recommendation with Knowledge-aware Interactive Matching. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 61–70.

[19] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021. 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)*. 5457–5467.

[20] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2022. News Recommendation with Candidate-aware User Modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1917–1921.

[21] Tao Qi, Fangzhao Wu, Chuhan Wu, Yongfeng Huang, and Xing Xie. 2020. Privacy-Preserving News Recommendation Model Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1423–1432.

[22] Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. 2021. HieRec: Hierarchical User Interest Modeling for Personalized News Recommendation. 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)*. 5446–5456.

[23] Shaina Raza and Chen Ding. 2022. News Recommender System: A Review of Recent Progress, Challenges, and Opportunities. *Artificial Intelligence Review* (2022), 1–52.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems* 30.

[25] Heyuan Wang, Fangzhao Wu, Zheng Liu, and Xing Xie. 2020. Fine-grained Interest Matching for Neural News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 836–845.

[26] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep Knowledge-aware Network for News Recommendation. In *Proceedings of the 2018 World Wide Web Conference*. 1835–1844.

[27] Rongyao Wang, Shoujin Wang, Wenpeng Lu, and Xueping Peng. 2022. News Recommendation via Multi-interest News Sequence Modelling. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7942–7946.

[28] Yinwei Wei, Xiang Wang, Qi Li, Liqiang Nie, Yan Li, Xuanping Li, and Tat-Seng Chua. 2021. Contrastive Learning for Cold-start Recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*. 5382–5390.

[29] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Attentive Multi-view Learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*. 3863–3869.

[30] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. NPA: Neural News Recommendation with Personalized Attention. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2576–2584.

[31] Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019. Neural News Recommendation with Topic-aware News Representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 1154–1159.

[32] Chuhan Wu, Fangzhao Wu, Suyu Ge, Tao Qi, Yongfeng Huang, and Xing Xie. 2019. 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)*. 6389–6394.

[33] Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. 2022. Rethinking InfoNCE: How Many Negative Samples Do You Need?. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, Lud De Raedt (Ed.). International Joint Conferences on Artificial Intelligence Organization, 2509–2515.

[34] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized News Recommendation: Methods and Challenges. *ACM Transactions on Information Systems* 41, 1 (2023), 1–50.

[35] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020. 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*. 44–53.

[36] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Empowering News Recommendation with Pre-trained Language Models. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1652–1656.

[37] Chuhan Wu, Fangzhao Wu, Tao Qi, Chenliang Li, and Yongfeng Huang. 2022. Is News Recommendation a Sequential Recommendation Task?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2382–2386.

[38] Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Ming Zhou. 2020. Mind: A Large-scale Dataset for News Recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3597–3606.

[39] Xu Xie, Fei Sun, Zhaoyang Liu, Shiwen Wu, Jinyang Gao, Jiandong Zhang, Bolin Ding, and Bin Cui. 2022. Contrastive Learning for Sequential Recommendation. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. IEEE, 1259–1273.

[40] Chun Yang, Jianxiao Zou, JianHua Wu, Hongbing Xu, and Shicai Fan. 2022. Supervised Contrastive Learning for Recommendation. *Knowledge-Based Systems* 258 (2022), 109973.

[41] Yang Yu, Fangzhao Wu, Chuhan Wu, Jingwei Yi, and Qi Liu. 2022. Tiny-NewsRec: Effective and Efficient PLM-based News Recommendation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 5478–5489.

[42] Qi Zhang, Qinglin Jia, Chuyuan Wang, Jingjie Li, Zhaowei Wang, and Xiuqiang He. 2021. Amm: Attentive Multi-field Matching for News Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1588–1592.