Train Once, Use Flexibly: A Modular Framework for Multi-Aspect Neural News Recommendation

Andreea Iana[1], Goran Glavaš[2], Heiko Paulheim[1],
[1] Data and Web Science Group, University of Mannheim, Germany
[2] Center For Artificial Intelligence and Data Science, University of Würzburg, Germany
{andreea.iana, heiko.paulheim}@uni-mannheim.de, goran.glavas@uni-wuerzburg.de

Abstract

Recent neural news recommenders (NNRs) extend content-based recommendation (1) by aligning additional *aspects* (e.g., topic, sentiment) between candidate news and user history or (2) by diversifying recommendations w.r.t. these aspects. This customization is achieved by "hardcoding" additional constraints into the NNR's architecture and/or training objectives: any change in the desired recommendation behavior thus requires retraining the model with a modified objective. This impedes widespread adoption of multi-aspect news recommenders. In this work, we introduce MANNeR, a modular framework for *multi-aspect* neural news recommendation that supports on-the-fly customization over individual aspects at inference time. With metric-based learning as its backbone, MANNeR learns aspect-specialized news encoders and then *flexibly* and *linearly* combines the resulting aspect-specific similarity scores into different ranking functions, alleviating the need for ranking function-specific retraining of the model. Extensive experimental results show that MANNeR consistently outperforms state-of-the-art NNRs on both standard content-based recommendation and single- and multi-aspect customization. Lastly, we validate that MANNeR's aspect-customization module is robust to language and domain transfer.

1 Introduction

Neural content-based recommenders, trained to infer users' preferences from their click history, represent the state of the art in news recommendation (Li and Wang, 2019; Wu et al., 2023). While previously consumed content clearly indicates users' preferences, *aspects* other than content alone, namely categorical features of the news such as topical category, sentiment, news outlet, or stance, contribute to their news consumption decisions. Accordingly, some neural news recommenders (NNRs) leverage information on these aspects in addition to text content, be it (i) directly as model input (Wu et al., 2019a; Liu et al., 2020) or (ii) indirectly, as auxiliary training tasks (Wu et al., 2019c, 2020a).

Increased personalization is often at odds with *diversity* (Pariser, 2011). NNRs optimized to maximize congruity to users' preferences tend to produce suggestions highly similar in content to previously consumed news (Liu et al., 2021; Wu et al., 2020a; Sertkan and Neidhardt, 2023). Another strand of work thus focuses on increasing diversity of recommendations w.r.t. aspects other than content (e.g., sentiment). To this effect, prior work either (i) re-ranks content-based recommendations to decrease the aspectual similarity between them (Rao et al., 2013; Gharahighehi and Vens, 2023), or (ii) trains the NNR model by combining a content-based personalization objective with an aspect-based diversification objective (Wu et al., 2020a, 2022b; Shi et al., 2022; Choi et al., 2022).

Different users assign different importance to various news aspects (e.g., following developing events requires maximization of content-based overlap with the user's recent history; in another use-case, a user may prefer content-wise diversification of recommendations, but within the same topic of interest). Moreover, with personalization and diversification as mutually conflicting goals, users should be able to seamlessly define their own optimal trade-offs between the two. The existing body of work is ill-equipped for such multi-aspect customization, because each set of preferences – i.e., to personalize or diversify for each aspect – requires a different NNR model to be trained from scratch. Put differently, forcing global assumptions on personalization and diversification preferences (i.e., same for all users) into the model design and training prevents customization at inference time.

Contributions. We propose a *modular* framework for *Multi-Aspect* Neural News Recommendation

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 9555–9571

(MANNeR) to address this limitation. It leverages metric-based contrastive learning to induce a dedicated news encoder for each aspect, starting from a pretrained language model (PLM). This way, we obtain linearly-combinable aspect-specific similarity scores for pairs of news, allowing us to define ad-hoc at inference a custom ranking function for each user, reflecting their preferences across all aspects. MANNeR's modular design allows customization for any recommendation objective specified over (i) standard (i.e., content-based) personalization, (ii) aspect-based diversification, and (iii) aspect-based personalization. It also makes MANNeR easily extendable: to support personalization and diversification over a new aspect (e.g., news outlet), one only needs to train the aspect-specific news encoder for that aspect. Through extensive experiments with *topical categories* and *sentiment* as additional aspects next to content itself, we find that MANNeR outperforms state-of-the-art NNRs on standard content-based recommendation. Thanks to its module-specific outputs being *linearly composable* between objectives, we show – without training numerous models with different objectives – that depending on the recommendation goals, one can either (i) vastly increase aspect diversity (e.g., over topics and sentiment) of recommendations or (ii) improve aspect-based personalization, while retaining much of the content-based personalization performance. Finally, we demonstrate that MANNeR with a multilingual PLM is robust to the (cross-lingual) transfer of aspect-based encoders.

2 Related Work

Personalized NNR. Neural content-based models have become the main vehicle of personalized news recommendation, replacing traditional recommenders relying on manual feature engineering (Wu et al., 2023). Most NNRs consist of a dedicated (i) news encoder (NE) and (ii) user encoder (UE) (Wu et al., 2023). The NE transforms input features into news embeddings (Wu et al., 2023, 2019d,b), whereas UEs create user-level representations by aggregating and contextualizing the embeddings of clicked news from the user's history (Okura et al., 2017; An et al., 2019; Wu et al., 2022c). The candidate's recommendation score is computed by comparing its embedding against the user embedding (Wang et al., 2018; Wu et al., 2019a). NNRs are primarily trained via point-wise classification objectives with negative sampling (Huang et al., 2013; Wu et al., 2021). Exploiting users' past behavior as NNR supervision leads to recommendations that are content-wise closest to previously consumed news, in contrast to methods based on non-personalized criteria (Son et al., 2013; Chen et al., 2017; Ludmann, 2017). More recent NNRs seek to augment content-based personalization by considering other aspects, such as categories, sentiment, emotions (Sertkan and Neidhardt, 2022), entities (Iana et al., 2024), outlets, or recency (Wu et al., 2023). These are incorporated in the NNR either as additional input to the NE (Wang et al., 2018; Gao et al., 2018; Wu et al., 2019a; Liu et al., 2020; Sheu and Li, 2020; Lu et al., 2020; Qi et al., 2021a; Xun et al., 2021), or in the form of an auxiliary training objective for the NE (Wu et al., 2019c, 2020a; Qi et al., 2021b).

Diversification. Personalized NNR reduces exposure to news dissimilar from those consumed in the past. Recommending "more of the same" constrains access to diverse viewpoints and information (Freedman and Sears, 1965; Heitz et al., 2022) and leads to homogeneous news diets and "filter bubbles" (Pariser, 2011), in turn reinforcing users' initial stances (Li and Wang, 2019). Consequently, a significant body of work attempts to diversify recommendations, either by re-ranking them to increase some measure of diversity (e.g. intra-list distance (Zhang and Hurley, 2008)) or by resorting to multi-task training (Gabriel De Souza et al., 2019; Wu et al., 2020a; Shi et al., 2022; Wu et al., 2022b; Choi et al., 2022; Raza, 2023), coupling the primary content-based personalization objective with auxiliary objectives that force aspect-based diversification.

Current NNR Limitations. Critically, existing approaches, by "hardcoding" aspectual requirements (i.e., personalization or diversification for an aspect) into the NNR's architecture and/or training objectives, cannot be easily adjusted for varying recommendation goals. Since even minor changes in the recommendation objective require retraining the NNR, current models are generally limited to fixed single-aspect recommendation scenarios (e.g., content-based personalization with topical diversification), and ill-equipped for multi-aspect customization. In this work, we rethink personalized news recommendation and propose a novel, modular multi-aspect recommendation framework that allows for ad-hoc creation of recommendation functions over aspects at inference time. This enables fundamentally different recommendation: one that lets each user define their own custom recommendation function, choosing the amount of personalization or diversification for each aspect.

3 Methodology

Personalized news recommendation produces for each candidate news n^c and user u with corresponding click history $H = \{n_1^u, n_2^u, ..., n_N^u\}$, a relevance score $s(n^c, u)$ that quantifies the candidate's relevance for the user. We define an *aspect* A_p as a categorical variable that encodes a news attribute (e.g. its category, stance, sentiment, provider), where each news n_i can belong only to one value of A_p (e.g. if A_p is the topic, then n_i may take exactly one value from {*politics*, *sports*, ...}). As discussed in §2, aspects are additional dimensions next to content over which to tailor recommendations, whether by (i) personalizing or (ii) diversifying over them. In line with earlier work, we define *aspect-based personalization* as the level of homogeneity between a user's recommendations and clicked news w.r.t. the distribution of aspect A_p. In contrast, we define *aspect-based diversity* as the level of uniformity of aspect A_p's distribution among the news in the recommendation list.

We next introduce our proposed *modular framework* MANNeR, illustrated in Fig. 1. Starting from a PLM, during (1) training, we reshape the PLM's representation space via contrastive learning, independently for each aspect; this results in one specialized NE for each aspect; at (2) inference, we can, depending on the recommendation task, aggregate the resulting aspect-specific similarity scores to produce a final ranking function.

3.1 News Encoder

We adopt a dual-component architecture for the NE coupling (i) a text and (ii) an entity encoder (Qi et al., 2021b,c). The former, a PLM, transforms the text input (i.e., concatenation of news title and abstract) into a text-based news embedding $\mathbf{n}_t$, given by the PLM's output `[CLS]` token representation. The latter learns an entity-level news embedding $\mathbf{n}_e$ by contextualizing pretrained embeddings of named entities (i.e., extracted from title and abstract) in a layer that combines multi-head self-attention (Vaswani et al., 2017) and additive attention (Bahdanau et al., 2014). The final news embedding $\mathbf{n}$ is the concatenation of $\mathbf{n}_t$ and $\mathbf{n}_e$.

3.2 Modular Training

MANNeR comprises two module types, each with a dedicated NE, responsible for content-based (`CR-Module`) and aspect-based (`A-Module`) recommendation relevance, respectively. We train both by minimizing the supervised contrastive loss (SCL, Eq. 1) which aims to reshape the NE's representation space so that embeddings of same-class instances become mutually closer (cf. a distance/similarity metric) than instances of different classes (Khosla et al., 2020; Gunel et al., 2020). To this end, we contrast the similarity score of a positive example (pair of same-class instances) against scores of corresponding negative examples (paired instances from different classes):

$$\mathcal{L} = -\sum_{i=1}^{N} \frac{1}{N_{y_i} - 1} \sum_{\substack{j \in [1,N] \\ i \neq j, y_i = y_j}} \log \frac{e^{(\mathbf{n}_i \cdot \mathbf{n}_j / \tau)}}{\sum_{\substack{k \in [1,N] \\ i \neq k}} e^{(\mathbf{n}_i \cdot \mathbf{n}_k / \tau)}} \quad (1)$$

with y_i as news n_i's label, N the batch size, N_{y_i} the number of batch instances with label y_i, and $\tau > 0$ the temperature hyperparameter controlling the extent of class separation. We use the dot product as the similarity metric for both module types.

CR-Module. Our `CR-Module` is a modification of the common content-based NNR architecture (Wu et al., 2023). Concretely, we encode both candidate and clicked news with a dedicated NE. However, following Iana et al. (2023b), we replace the widely used UEs (i.e., early fusion of clicked news representations) with the simpler (and non-parameterized) mean-pooling of dot-product scores between the candidate embedding $\mathbf{n}^c$ and clicked news embeddings $\mathbf{n}_i^u$: $s(\mathbf{n}^c, u) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{n}^c \cdot \mathbf{n}_i^u$ (i.e., late-fusion). We thus reduce the computational complexity of the standard approaches with elaborate parameterized UEs. We then update the `CR-Module`'s encoder (i.e., fine-tune the PLM) by minimizing SCL, with clicked candidates as positive and non-clicked news as negative examples for the user. As there are many more non-clicked news, we resort to negative sampling (Wu et al., 2022a).

A-Module. Each `A-Module` trains a specialized NE for one aspect other than content. Via the metric-based objective, we reshape the PLM's representation space to group news according to aspect classes. Given a multi-class aspect, we first construct the training set from the union of all news in the dataset. Sets of news with the same aspect label form the positive samples for SCL; we obtain the corresponding negatives by pairing the same news

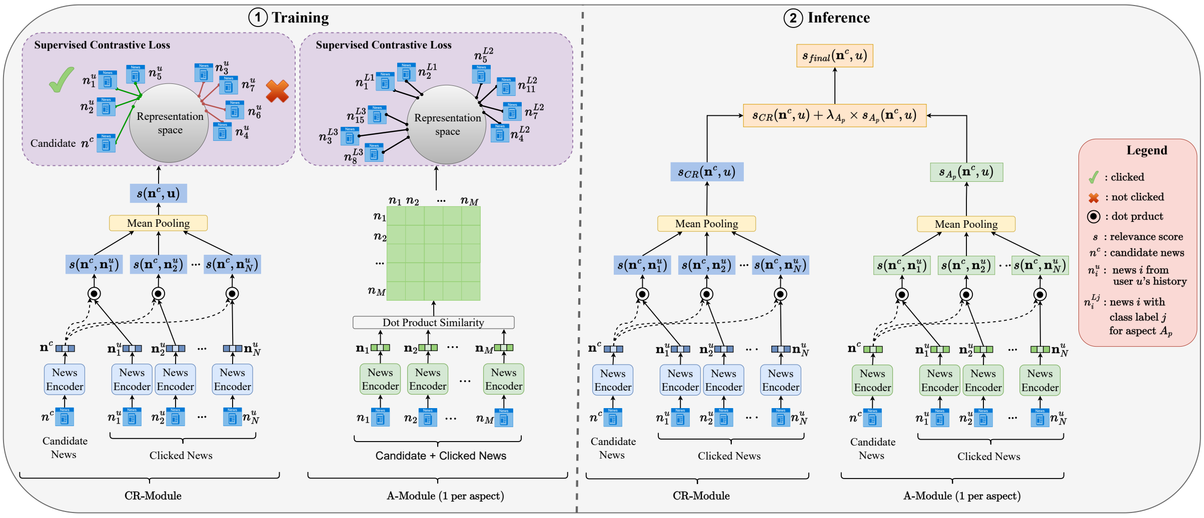


Figure 1: Illustration of the MANNeR framework. ① We train aspect-specialized NEs (i.e. `CR-Module` for content-based personalization, `A-Module` for aspect-based similarity) with metric-based contrastive learning. ② Inference: we linearly aggregate aspect-specific similarity scores between candidate and clicked news for final ranking.

from positive pairs with news from other aspect classes (e.g., for topical category as A_p, a news from *sports* is paired with the news from *politics* and/or *weather*). For each aspect, we independently fine-tune a separate copy of the same initial PLM. Note that the resulting aspect-specific NE encodes no information on user preferences: it only encodes the news similarity w.r.t. the aspect in question. Importantly, this implies that extending MANNeR to support a new aspect amounts to merely training an additional `A-Module` for that aspect.

3.3 Inference: Custom Ranking Functions

At inference time, the NEs of the `CR-Module` and of each of the `A-Modules` are leveraged identically: we encode the candidate news as well as the user's clicked news with the respective NE, obtaining their module-specific embeddings $\mathbf{n}^c$ and $\mathbf{n}_i^u$ – their dot product $s = \mathbf{n}^c \cdot \mathbf{n}_i^u$ quantifies their similarity according to the module's aspect (or content for `CR-Module`'s NE). As different NEs produce similarity scores of different magnitudes, we z-score normalize each module's scores per user. The final ranking score constitutes a *linear* aggregation of the content s_{CR} and aspect s_{A_p} similarity scores:

$$s_{final}(\mathbf{n}^c, \mathbf{u}) = s_{CR} + \sum_{A_p \in A} \lambda_{A_p} s_{A_p} \tag{2}$$

where λ_{A_p} is the scaling weight for the aspect score, and A the set of all aspects of interest. This linear composability of aspect-specific similarity scores allows not only generalization to multi-aspect recommendation objectives, but also different ad-hoc realizations of the ranking function that match custom recommendation goals: (i) with $\lambda_{A_p} = 0$, MANNeR performs standard content-based personalization, (ii) for $\lambda_{A_p} > 0$ it recommends based on both content- and aspect personalization, whereas (iii) for $\lambda_{A_p} < 0$ it simultaneously personalizes by content but diversifies for the aspect(s).

4 Experimental Setup

We compare MANNeR against state-of-the-art NNRs on a range of single- and multi-aspect recommendation tasks. We experiment with two aspects: *topical categories* (*ctg*) and news *sentiment* (*snt*).

Baselines. We evaluate several NNRs trained on classification objectives. We follow Wu et al. (2021) and replace the original NEs of all baselines that do not use PLMs (instead, contextualizing word embeddings with convolution or self-attention layers) with the same PLM used in MANNeR.[1] We include two models optimized purely for content personalization: (1) NRMS (Wu et al., 2019d), and (2) MINER (Li et al., 2022). We further evaluate seven NNRs that inject aspect information. Thereof, five incorporate *topical categories*, i.e., (3) NAML (Wu et al., 2019a), (4) LSTUR (An et al., 2019), (5) MINS (Wang et al., 2022), (6) CAUM (Qi et al., 2022), (7) TANR (Wu et al., 2019c), and two the news *sentiment*: (8) SentiRec (Wu et al., 2020a), and (9) SentiDebias (Wu et al., 2022d).

[1] The only exception is the final text embedding, where Wu et al. (2021) pool tokens with an attention network.

Data. We carry out the evaluation on two prominent monolingual news recommendation benchmarks: MINDlarge (denoted MIND) (Wu et al., 2020b) with news in English and Adressa-1 week (Gulla et al., 2017) (denoted Adressa) with Norwegian news. Since Wu et al. (2020b) do not release test labels for MIND, we use the provided validation portion for testing, and split the respective training set into temporally disjoint training (first four days of data) and validation portions (the last day). Following established practices on splitting the Adressa dataset (Hu et al., 2020; Xu et al., 2023), we use the data of the first five days to construct user histories and the clicks of the sixth day to build the training dataset. We randomly sample 20% of the last day's clicks to create the validation set, and treat the remaining samples of the last day as the test set.[2] Since Adressa contains only positive samples (i.e., no data about users' seen but not clicked news), we randomly sample 20 news as negatives for each clicked article to build impressions following Yi et al. (2021).[3] As Adressa contains no disambiguated named entities, we use only the news title as input to MANNeR' NE, while on MIND we use all news features as NE input.

Regarding aspects, the topical category annotations are provided in both datasets. As no sentiment labels exist in neither MIND nor Adressa, we use a multilingual XLM-RoBERTa Base model (Conneau et al., 2020) trained on tweets and fine-tuned for sentiment analysis (Barbieri et al., 2022) to classify news into three classes: positive (pos), neutral, and negative (neg). We compute real-valued scores using the model's confidence scores s_i for class i, and the predicted sentiment class label $\hat{l}$ as follows:

$$s_{sent} = \begin{cases} (+1) \times s_{pos}, \text{ if } \hat{l} = pos \\ (-1) \times s_{neg}, \text{ if } \hat{l} = neg \\ (1 - s_{neutral}) \times (s_{pos} - s_{neg}), \text{ otherwise} \end{cases} \quad (3)$$

Evaluation Metrics. We report performance with AUC, MRR, nDCG@k ($k = \{5, 10\}$). We measure aspect-based diversity of recommendations at position k as the normalized entropy of the distribution of aspect A_p's values in the recommendation list:

$$D_{A_p}@k = -\sum_{j \in A_p} \frac{p(j) \log p(j)}{\log(|A_p|)} \quad (4)$$

where $A_p \in \{ctg, snt\}$, and $|A_p|$ is the number of A_p classes. If aspect-based personalization is successful, aspect A_p's distribution in the recommendations should be similar to its distribution in the user history. We evaluate personalization with the generalized Jaccard similarity (Bonnici, 2020):

$$PS_{A_p}@k = \frac{\sum_{j=1}^{|A_p|} \min(\mathcal{R}_j, \mathcal{H}_j)}{\sum_{j=1}^{|A_p|} \max(\mathcal{R}_j, \mathcal{H}_j)}, \quad (5)$$

where R_j and H_j represent the probability of a news with class j of A_p to be contained in the recommendations list R, and, respectively, in the user history H. As all metrics are bounded to $[0, 1]$, we measure the trade-off between content-based personalization (nDCG@k) and either aspect-based diversity $D_{A_p}@k$ or aspect-based personalization $PS_{A_p}@k$ with the harmonic mean. We denote this $\mathrm{T}_{\mathrm{A_p}}@k$ for single-aspect. For multi-aspect evaluation, i.e., when ranking for content-personalization by diversifying simultaneously over topics and sentiment, we adopt as evaluation metric the harmonic mean between nDCG@k, $\mathrm{D}_{\mathrm{ctg}}@k$ (topical category), and $\mathrm{D}_{\mathrm{snt}}@k$ (sentiment), denoted $\mathrm{T}_{\mathrm{all}}@k$.

Training Details. We use RoBERTa Base (Liu et al., 2019) and NB-BERT Base (Kummervold et al., 2021; Nielsen, 2023) in experiments on MIND and Adressa, respectively. We set the maximum history length to 50. We tune the main hyperparameters of all NNRs. We train all models with mixed precision, the Adam optimizer (Kingma and Ba, 2014), the learning rate of 1e-5 on MIND, 1e-6 on Adressa, and 1e-6 for the sentiment `A-Module` on both datasets. In `A-Module` training, we sample 20 instances per class,[4] while in `CR-Module` training we sample four negatives per positive example. We find the optimal temperature of 0.36 on MIND, and 0.14 on Adressa, for the `CR-Module`, and of 0.9 for all `A-Modules` on both datasets. We train all baselines and the `CR-Module` for 5 epochs on MIND and 20 epochs on Adressa, with a batch size of 8. We train each `A-Module` for 100 epochs, with the batch size of 60 for sentiment and 360 for topics. We repeat runs five times with different seeds and report averages and standard deviations for all metrics. We refer to Appendices B.1 - B.2 for further details about model sizes and hyperparameters.

5 Results and Discussion

We first discuss MANNeR's content personalization performance. We then analyze its capability

[2]Note that during validation and testing, we reconstruct user histories with all the samples of the first six days of data.

[3]Table 4 summarizes the datasets' statistics.

[4]For M class instances, we obtain $\frac{M^2-M}{2}$ positive pairs for that class for SCL.

Model	MIND AUC	MIND MRR	MIND nDCG@5	MIND nDCG@10	Adressa AUC	Adressa MRR	Adressa nDCG@5	Adressa nDCG@10
NRMS-PLM	63.0$_{\pm 1.5}$	35.5$_{\pm 0.6}$	33.4$_{\pm 0.7}$	39.9$_{\pm 0.6}$	72.3$_{\pm 3.3}$	43.0$_{\pm 2.7}$	44.3$_{\pm 2.8}$	51.3$_{\pm 2.3}$
MINER	63.1$_{\pm 1.2}$	35.5$_{\pm 1.1}$	33.7$_{\pm 1.1}$	40.0$_{\pm 1.0}$	70.1$_{\pm 4.9}$	37.3$_{\pm 4.1}$	38.5$_{\pm 5.1}$	46.3$_{\pm 4.1}$
NAML-PLM	60.6$_{\pm 3.4}$	37.6$_{\pm 0.4}$	35.9$_{\pm 0.4}$	42.2$_{\pm 0.4}$	50.0$_{\pm 0.0}$	45.0$_{\pm 5.0}$	47.2$_{\pm 5.5}$	52.5$_{\pm 4.1}$
LSTUR-PLM	54.6$_{\pm 3.0}$	33.3$_{\pm 1.5}$	31.7$_{\pm 1.8}$	38.3$_{\pm 1.7}$	65.0$_{\pm 7.2}$	43.1$_{\pm 1.7}$	44.8$_{\pm 2.6}$	51.2$_{\pm 2.0}$
MINS-PLM	61.3$_{\pm 2.7}$	36.2$_{\pm 0.3}$	34.5$_{\pm 0.4}$	40.8$_{\pm 0.3}$	74.3$_{\pm 3.2}$	44.2$_{\pm 2.9}$	47.3$_{\pm 3.3}$	53.0$_{\pm 3.4}$
CAUM$_{\text{no entities}}$-PLM	66.2$_{\pm 3.0}$	36.6$_{\pm 2.0}$	34.6$_{\pm 2.0}$	41.0$_{\pm 1.9}$	76.5$_{\pm 1.2}$	43.6$_{\pm 1.3}$	46.9$_{\pm 1.3}$	52.0$_{\pm 1.2}$
CAUM-PLM	66.4$_{\pm 3.1}$	36.2$_{\pm 1.2}$	34.3$_{\pm 1.3}$	40.8$_{\pm 1.3}$	–	–	–	–
TANR-PLM	63.3$_{\pm 1.1}$	37.0$_{\pm 1.0}$	35.2$_{\pm 1.0}$	41.6$_{\pm 0.9}$	50.0$_{\pm 0.0}$	43.8$_{\pm 1.0}$	45.6$_{\pm 1.3}$	51.4$_{\pm 0.6}$
SentiRec-PLM	62.2$_{\pm 0.7}$	35.7$_{\pm 0.4}$	33.9$_{\pm 0.4}$	40.5$_{\pm 0.4}$	67.6$_{\pm 2.7}$	33.1$_{\pm 2.4}$	32.9$_{\pm 3.8}$	40.8$_{\pm 2.4}$
SentiDebias-PLM	55.0$_{\pm 2.5}$	27.8$_{\pm 1.9}$	25.5$_{\pm 1.9}$	32.2$_{\pm 2.0}$	67.4$_{\pm 2.4}$	35.7$_{\pm 3.4}$	36.4$_{\pm 4.2}$	44.2$_{\pm 2.9}$
MANNeR (`CR-Module`)	**69.7**$_{\pm 0.9}$	**38.6**$_{\pm 0.6}$	**37.0**$_{\pm 0.6}$	**43.2**$_{\pm 0.6}$	**79.4**$_{\pm 1.7}$	**47.0**$_{\pm 2.4}$	**48.9**$_{\pm 2.8}$	**54.3**$_{\pm 2.5}$
Improvement (%)	+ 5.4	+ 2.8	+ 3.1	+ 2.3	+ 3.7	+ 4.6	+ 3.3	+ 2.5

Table 1: Content-based recommendation performance. We average results across five runs, and report the relative improvement over the best baseline. The best results per column are highlighted in bold, the second best underlined.

Model	MIND nDCG@10	MIND D_{ctg}@10	MIND T_{ctg}@10	MIND D_{snt}@10	MIND T_{snt}@10	MIND T_{all}@10	Adressa nDCG@10	Adressa D_{ctg}@10	Adressa T_{ctg}@10	Adressa D_{snt}@10	Adressa T_{snt}@10	Adressa T_{all}@10
NRMS-PLM	39.9$_{\pm 0.6}$	50.0$_{\pm 1.1}$	44.3$_{\pm 0.4}$	66.4$_{\pm 0.5}$	49.8$_{\pm 0.5}$	49.9$_{\pm 0.3}$	51.3$_{\pm 2.3}$	31.8$_{\pm 1.0}$	39.2$_{\pm 0.5}$	61.5$_{\pm 0.5}$	55.9$_{\pm 1.2}$	44.6$_{\pm 0.5}$
MINER	40.0$_{\pm 1.0}$	49.4$_{\pm 1.2}$	44.2$_{\pm 0.4}$	65.7$_{\pm 0.9}$	49.7$_{\pm 1.0}$	49.6$_{\pm 0.5}$	46.3$_{\pm 4.1}$	31.1$_{\pm 0.6}$	37.1$_{\pm 1.6}$	60.9$_{\pm 0.5}$	52.5$_{\pm 2.8}$	42.7$_{\pm 1.5}$
NAML-PLM	42.2$_{\pm 0.4}$	47.3$_{\pm 0.3}$	44.6$_{\pm 0.3}$	65.1$_{\pm 0.4}$	51.2$_{\pm 0.3}$	49.9$_{\pm 0.3}$	52.5$_{\pm 4.1}$	30.6$_{\pm 2.4}$	38.6$_{\pm 2.1}$	61.6$_{\pm 0.6}$	56.7$_{\pm 2.6}$	44.0$_{\pm 1.9}$
LSTUR-PLM	38.3$_{\pm 1.7}$	50.0$_{\pm 1.2}$	43.4$_{\pm 0.7}$	65.6$_{\pm 0.3}$	48.4$_{\pm 1.3}$	48.9$_{\pm 0.5}$	51.2$_{\pm 2.0}$	29.9$_{\pm 4.6}$	37.7$_{\pm 5.2}$	61.4$_{\pm 0.5}$	55.8$_{\pm 1.2}$	43.2$_{\pm 3.8}$
MINS-PLM	40.8$_{\pm 0.3}$	49.1$_{\pm 1.0}$	44.6$_{\pm 0.3}$	66.3$_{\pm 0.9}$	50.5$_{\pm 0.1}$	50.0$_{\pm 0.4}$	53.0$_{\pm 3.4}$	33.6$_{\pm 1.7}$	41.0$_{\pm 1.0}$	61.8$_{\pm 0.6}$	57.0$_{\pm 1.8}$	46.2$_{\pm 0.9}$
CAUM$_{\text{no entities}}$-PLM	41.0$_{\pm 1.9}$	47.4$_{\pm 1.0}$	43.9$_{\pm 0.9}$	65.8$_{\pm 1.2}$	50.5$_{\pm 1.3}$	49.4$_{\pm 0.6}$	52.0$_{\pm 1.2}$	34.4$_{\pm 0.3}$	**41.4**$_{\pm 0.4}$	62.1$_{\pm 0.5}$	56.6$_{\pm 0.7}$	**46.6**$_{\pm 0.3}$
CAUM-PLM	40.8$_{\pm 1.3}$	47.8$_{\pm 0.9}$	44.0$_{\pm 1.0}$	66.1$_{\pm 0.5}$	50.6$_{\pm 1.0}$	49.6$_{\pm 0.9}$	–	–	–	–	–	–
TANR-PLM	41.6$_{\pm 0.9}$	48.9$_{\pm 0.9}$	45.0$_{\pm 0.3}$	66.1$_{\pm 0.8}$	51.1$_{\pm 0.7}$	50.3$_{\pm 0.3}$	51.4$_{\pm 0.6}$	32.9$_{\pm 1.7}$	40.1$_{\pm 1.1}$	61.8$_{\pm 0.7}$	56.1$_{\pm 0.2}$	45.4$_{\pm 1.0}$
SentiRec-PLM	40.5$_{\pm 0.4}$	49.4$_{\pm 0.4}$	44.5$_{\pm 0.1}$	67.0$_{\pm 0.6}$	50.4$_{\pm 0.4}$	50.1$_{\pm 0.2}$	40.8$_{\pm 2.4}$	**35.6**$_{\pm 0.6}$	38.0$_{\pm 1.1}$	**68.5**$_{\pm 0.2}$	51.1$_{\pm 1.9}$	44.6$_{\pm 1.0}$
SentiDebias-PLM	32.2$_{\pm 2.0}$	**52.0**$_{\pm 2.2}$	39.7$_{\pm 1.1}$	68.6$_{\pm 1.2}$	43.8$_{\pm 1.8}$	46.2$_{\pm 1.0}$	44.2$_{\pm 2.9}$	32.3$_{\pm 1.0}$	37.3$_{\pm 1.2}$	61.2$_{\pm 0.2}$	51.3$_{\pm 2.0}$	42.9$_{\pm 1.1}$
MANNeR (`CR-Module`)	**43.2**$_{\pm 0.6}$	49.3$_{\pm 0.3}$	46.0$_{\pm 0.3}$	65.4$_{\pm 0.6}$	52.0$_{\pm 0.4}$	51.1$_{\pm 0.2}$	**54.3**$_{\pm 2.5}$	31.7$_{\pm 0.2}$	40.0$_{\pm 0.7}$	61.4$_{\pm 0.3}$	57.6$_{\pm 1.5}$	45.3$_{\pm 0.6}$
MANNeR ($\lambda_{ctg} = -0.2/-0.3$, $\lambda_{snt} = 0$)	42.0$_{\pm 0.6}$	51.5$_{\pm 0.3}$	**46.2**$_{\pm 0.3}$	65.6$_{\pm 0.6}$	51.2$_{\pm 0.4}$	51.3$_{\pm 0.3}$	50.9$_{\pm 2.5}$	34.1$_{\pm 0.3}$	40.8$_{\pm 0.8}$	61.9$_{\pm 0.3}$	55.8$_{\pm 1.6}$	46.0$_{\pm 0.7}$
MANNeR ($\lambda_{ctg} = 0$, $\lambda_{snt} = -0.3/-0.2$)	42.8$_{\pm 0.7}$	49.8$_{\pm 0.2}$	46.0$_{\pm 0.4}$	**68.7**$_{\pm 0.3}$	**52.7**$_{\pm 0.4}$	**51.7**$_{\pm 0.3}$	53.8$_{\pm 2.5}$	32.4$_{\pm 0.2}$	40.4$_{\pm 0.7}$	63.0$_{\pm 0.3}$	**58.0**$_{\pm 1.5}$	45.9$_{\pm 0.6}$

Table 2: Single-aspect diversification. For MANNeR, we list the best results (cf. T_{A_p}) of single-aspect diversification as λ_{A_p} (MIND/Adressa). The best results per column are highlighted in bold, the second best underlined.

for single- and multi-aspect (i) diversification and (ii) personalization. In the aspect customization setups, we treat MANNeR's `CR-Module` as a baseline. Lastly, we evaluate its ability to re-use pretrained aspect-specific modules in cross-lingual transfer.

5.1 Content Personalization

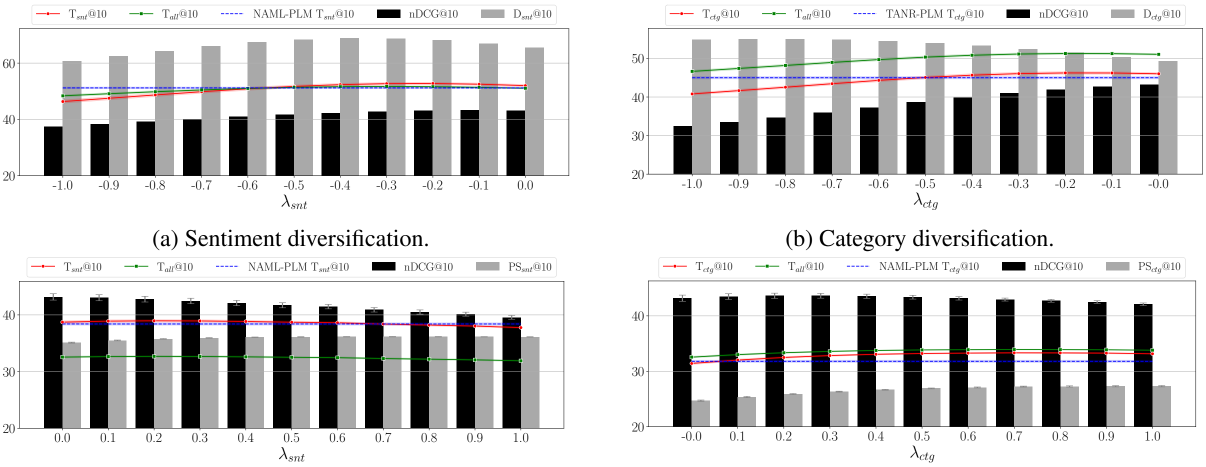
Table 1 summarizes the results on content personalization. Since the task does not require any aspect-based customization, we evaluate the MANNeR variant that uses only its CR-Module at inference time (i.e., $\lambda = 0$). MANNeR consistently outperforms all state-of-the-art NNRs in terms of both classification and ranking metrics on both datasets. Given that MANNeR's `CR-Module` derives the user embedding by merely averaging clicked news embeddings, these results question the need for complex parameterized UEs, present in all the baselines, in line with the findings of Iana et al. (2023b).

We ablate the `CR-Module`'s content personalization performance for (i) different inputs to the NE and (ii) alternative architecture designs and training objectives. We find that all groups of features (e.g., abstract, named entities) contribute to the overall performance (cf. Fig. 6a). Moreover, we confirm the findings of Iana et al. (2023b) that (i) late fusion outperforms a parameterized UE (i.e., early fusion), and that (ii) SCL better separates classes than cross-entropy loss, in line with other similarity-oriented NLP tasks (Reimers and Gurevych, 2019).

5.2 Single-Aspect Customization

Diversification. Table 2 summarizes the results on aspect diversification tasks. Most baselines (including MANNeR's `CR-Module` without aspect diversification) obtain similar diversification scores (D_{ctg} and D_{snt}). The sentiment-aware SentiRec-PLM, with an explicit auxiliary sentiment diversification objective, yields the highest sentiment diversity on Adressa; this comes at the expense of content personalization quality (lowest nDCG). On MIND, the sentiment-specific SentiDebias-PLM achieves the highest sentiment diversity, but also exhibits lower content personalization performance. Overall, these results point to a trade-off between content personalization and aspectual diversity: models with higher D_{A_p} tend to have a lower nDCG.

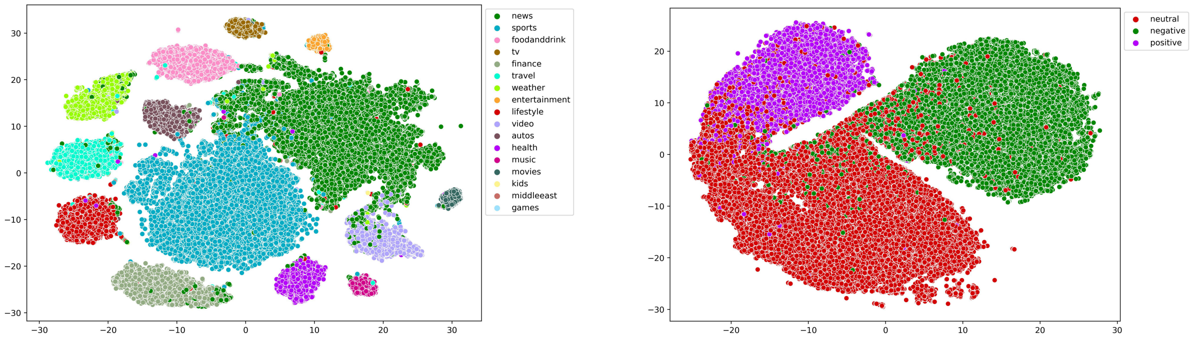
Unlike all other models, MANNeR can trade content personalization for diversity (and vice-versa) with different values of the aspect coeffi-



(a) Sentiment diversification. (b) Category diversification.

(c) Sentiment personalization. (d) Category personalization.

Figure 2: Results of single-aspect customization for MANNeR and the best baseline on MIND.



(a) Category-shaped embedding space. (b) Sentiment-shaped embedding space.

Figure 3: t-SNE plots of the news embeddings in the test set of MIND.

cients λ_{A_p}. Figs. 2a-2b illustrate its performance in single-aspect sentiment and category diversification tasks for different values of λ_{snt}, and λ_{ctg}, respectively, on MIND. The steady drop in nDCG together with the steady increase in $\text{D}_{\text{A}_\text{p}}$ indeed indicate the existence of a trade-off between content personalization and aspect diversification. For topical categories we observe a steeper decline in content personalization quality with improved diversification than for sentiment. Sentiment diversity reaches peak performance for $\lambda_{snt} = -0.4$, whereas category diversity continues to increase up to $\lambda_{ctg} = -0.9$. Intuitively, content-based recommendation is more aligned with the topical than with the sentiment consistency of recommendations. The best trade-off (i.e., maximal performance w.r.t. $\text{T}_{\text{A}_\text{p}}$@10) is achieved for $\lambda_{snt} = -0.3$ for sentiment, and $\lambda_{ctg} = -0.2$ for topics.[5] We attribute these effects to the representation spaces of the A-Modules. Fig. 3 shows the 2-dimensional t-SNE visualizations (Van der Maaten and Hinton, 2008) of the news embeddings produced with category-specialized, and respectively, sentiment-specialized NEs trained on MIND. The results confirm that the encoder's latent representation space was reshaped to group same-class instances. The separation of classes, however, is less prominent for the representation spaces of the encoders trained on Adressa (cf. Fig. 8, e.g., the effect is stronger on the category-shaped embedding space).[6]

Personalization. Table 3 displays the results on aspect personalization tasks. TANR, trained with an auxiliary topic classification task, underperforms NAML, which uses topical categories as NE input features, in category personalization on both datasets. MANNeR's CR-Module alone (i.e., without any aspect customization) yields competitive category personalization performance. We believe that this is because (i) the CR-Module is best in content personalization and (ii) category personalization is well-aligned with content personalization

[5]We report analogous results on Adressa in Figs. 7a-7b.

[6]We believe that this is because Adressa has 10 times fewer news than MIND, with over half of the topical categories in Adressa being represented with fewer than 100 examples.

Model	MIND						Adreesa					
	nDCG@10	PS_{ctg}@10	T_{ctg}@10	PS_{snt}@10	T_{snt}@10	T_{all}@10	nDCG@10	PS_{ctg}@10	T_{ctg}@10	PS_{snt}@10	T_{snt}@10	T_{all}@10
NRMS-PLM	$39.9_{\pm0.6}$	$23.9_{\pm0.2}$	$29.9_{\pm0.3}$	$35.1_{\pm0.1}$	$37.3_{\pm0.3}$	$31.5_{\pm0.2}$	$51.3_{\pm2.3}$	$34.3_{\pm0.4}$	$41.1_{\pm1.0}$	$41.8_{\pm0.1}$	$46.1_{\pm1.0}$	$41.3_{\pm0.7}$
MINER	$40.0_{\pm1.0}$	$23.9_{\pm0.4}$	$29.9_{\pm0.5}$	$35.0_{\pm0.2}$	$37.3_{\pm0.4}$	$31.5_{\pm0.4}$	$46.3_{\pm4.1}$	$34.4_{\pm0.2}$	$39.4_{\pm1.5}$	$42.0_{\pm0.0}$	$43.9_{\pm1.8}$	$40.2_{\pm1.0}$
NAML-PLM	$42.2_{\pm0.4}$	$25.5_{\pm0.2}$	$31.8_{\pm0.2}$	$35.1_{\pm0.2}$	$38.4_{\pm0.2}$	$32.8_{\pm0.2}$	$52.5_{\pm4.1}$	$36.1_{\pm0.8}$	$42.7_{\pm1.7}$	$41.8_{\pm0.1}$	$46.5_{\pm1.7}$	$42.4_{\pm1.1}$
LSTUR-PLM	$38.3_{\pm1.7}$	$24.0_{\pm1.0}$	$29.5_{\pm1.2}$	$34.8_{\pm0.3}$	$36.5_{\pm0.9}$	$31.1_{\pm1.0}$	$51.2_{\pm2.0}$	$35.1_{\pm2.1}$	$41.6_{\pm1.0}$	$41.8_{\pm0.1}$	$46.0_{\pm0.8}$	$41.7_{\pm0.7}$
MINS-PLM	$40.8_{\pm0.3}$	$25.0_{\pm0.3}$	$31.0_{\pm0.3}$	$34.7_{\pm0.2}$	$37.5_{\pm0.2}$	$32.1_{\pm0.3}$	$53.0_{\pm3.4}$	$33.9_{\pm0.7}$	$41.3_{\pm1.4}$	$41.8_{\pm0.1}$	$46.7_{\pm1.3}$	$41.5_{\pm1.0}$
$CAUM_{no\ entities}$-PLM	$41.0_{\pm1.9}$	$24.8_{\pm0.6}$	$30.9_{\pm1.0}$	$35.0_{\pm0.2}$	$37.8_{\pm0.9}$	$32.2_{\pm0.7}$	$52.0_{\pm1.2}$	$33.5_{\pm0.2}$	$39.6_{\pm1.1}$	$40.8_{\pm0.4}$	$46.3_{\pm0.5}$	$41.1_{\pm0.3}$
CAUM-PLM	$40.8_{\pm1.3}$	$25.1_{\pm0.3}$	$31.1_{\pm0.4}$	$35.0_{\pm0.1}$	$37.7_{\pm0.6}$	$32.3_{\pm0.3}$	–	–	–	–	–	–
TANR-PLM	$41.6_{\pm0.9}$	$25.2_{\pm0.5}$	$31.4_{\pm0.6}$	$35.0_{\pm0.2}$	$38.0_{\pm0.4}$	$32.5_{\pm0.5}$	$51.4_{\pm0.6}$	$34.0_{\pm0.5}$	$41.0_{\pm0.5}$	$41.8_{\pm0.1}$	$46.1_{\pm0.3}$	$41.2_{\pm0.4}$
SentiRec-PLM	$40.5_{\pm0.4}$	$24.2_{\pm0.3}$	$30.3_{\pm0.3}$	$34.6_{\pm0.0}$	$37.3_{\pm0.2}$	$31.6_{\pm0.2}$	$40.8_{\pm2.4}$	$32.4_{\pm0.3}$	$36.1_{\pm1.0}$	$39.3_{\pm0.1}$	$40.0_{\pm1.2}$	$37.1_{\pm0.7}$
SentiDebias-PLM	$32.2_{\pm2.0}$	$20.8_{\pm1.3}$	$25.2_{\pm1.5}$	$34.1_{\pm0.3}$	$33.1_{\pm1.2}$	$27.6_{\pm1.2}$	$44.2_{\pm2.9}$	$34.1_{\pm0.6}$	$38.5_{\pm1.3}$	$41.8_{\pm0.1}$	$42.9_{\pm1.4}$	$39.5_{\pm1.0}$
MANNeR (`CR-Module`)	**$43.2_{\pm0.6}$**	$24.7_{\pm0.1}$	$31.4_{\pm0.2}$	$35.1_{\pm0.1}$	$38.7_{\pm0.2}$	$32.6_{\pm0.2}$	**$54.3_{\pm2.5}$**	$34.5_{\pm0.1}$	$42.2_{\pm0.8}$	$42.0_{\pm0.1}$	$47.3_{\pm0.9}$	$42.1_{\pm0.5}$
MANNeR ($\lambda_{ctg} = 0.7/0.4$, $\lambda_{snt} = 0$)	$42.9_{\pm0.3}$	**$27.2_{\pm0.1}$**	**$33.3_{\pm0.1}$**	$35.2_{\pm0.0}$	$38.7_{\pm0.1}$	**$33.9_{\pm0.1}$**	$53.6_{\pm1.9}$	**$36.2_{\pm0.1}$**	**$43.2_{\pm0.7}$**	$42.1_{\pm0.1}$	$47.2_{\pm0.7}$	**$42.9_{\pm0.4}$**
MANNeR ($\lambda_{ctg} = 0$, $\lambda_{snt} = 0.2/0.1$)	$42.8_{\pm0.5}$	$24.7_{\pm0.1}$	$31.3_{\pm0.2}$	**$35.8_{\pm0.1}$**	**$39.0_{\pm0.2}$**	$32.7_{\pm0.1}$	$54.1_{\pm2.4}$	$34.7_{\pm0.1}$	$42.2_{\pm0.8}$	**$42.2_{\pm0.1}$**	**$47.4_{\pm0.9}$**	$42.2_{\pm0.5}$

Table 3: Single-aspect personalization. For MANNeR, we list the best results (cf. T_{A_p}) of single-aspect diversification as λ_{A_p} (MIND/Adressa). The best results per column are highlighted in bold, the second best underlined.

(i.e., news with similar content tend to belong to the same category). Fig. 2d explores the trade-off between content and category personalization, for positive values of λ_{ctg} on MIND. The best topical category personalization (PS_{ctg}), obtained for $\lambda_{ctg} > 0.7$, comes at the small expense of content personalization: too much weight on the category similarity of news dilutes the impact of content relevance. Increased sentiment personalization (cf. Fig. 2c), however, is much more detrimental to content personalization. Intuitively, users do not choose articles based on sentiment. Tailoring recommendations according to the sentiment of previously clicked news thus leads to more content-irrelevant suggestions.

5.3 Multi-Aspect Customization

We further explore the trade-off between content personalization and multi-aspect diversification, i.e. diversifying over both topical categories and sentiments. We achieve the highest T_{all} for $\lambda_{ctg} = -0.2$ and $\lambda_{snt} = -0.25$ on MIND (cf. Fig. 4a). In line with single-aspect diversification results, we observe that improving diversity in terms of topical categories rather than sentiments has a more negative effect on content personalization quality, i.e. steeper decline in T_{all}. These results confirm that MANNeR can generalize to diversify for multiple aspects at once by weighting individual aspect relevance scores less than in the single-aspect task. Weighting several aspects higher simultaneously acts as a double discounting for content personalization, diluting content relevance disproportionately. Similarly, for multi-aspect personalization, we achieve the best multi-aspect trade-off on MIND (cf. Fig. 4b) for $\lambda_{ctg} = 0.45$ and $\lambda_{snt} = 0.25$. Stronger enforcing of alignment of candidate news with the user's history is needed for topical categories than for sentiment (i.e., $\lambda_{ctg} > \lambda_{snt}$). This is because sentiment exhibits low variance within categories (e.g., *politics* news are mostly negative) and enforcing categorical personalization partly also achieves sentiment personalization.[7]

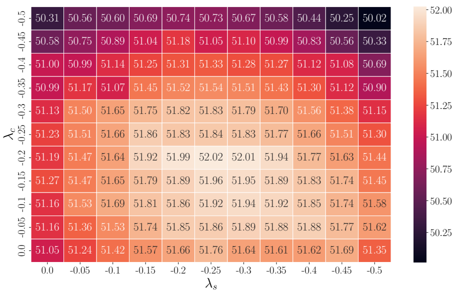
5.4 Cross-Lingual Transfer

We next analyze the transferability of MANNeR across datasets and languages in single-aspect customization experiments.[8] Concretely, we train the `CR-Module` and `A-Modules` on both MIND (i.e., in English) and Adressa (i.e., in Norwegian), respectively. At inference, we evaluate all combinations of pretrained `CR-Module` and `A-Modules` on the test set of MIND. We now use a multilingual DistilBERT Base (Sanh et al., 2019) as MANNeR's NE to enable cross-lingual transfer (`XLT`). Fig. 5 summarizes the `XLT` results for single-aspect diversification.[9] As expected, MANNeR trained fully on Adressa suffers a large drop in content personalization performance, compared to the counterpart trained on MIND. In contrast, transferring only the `A-Module`, i.e., training the `CR-Module` on MIND and the `A-Module` on Adressa, yields performance comparable to that of complete in-language training (i.e., both `CR-Module` and `A-Module` trained on MIND). This is particularly the case for the sentiment `A-Module`, since the sentiment labels between the datasets are more aligned than those for topical categories. These results indicate that the plug-and-play of `A-Modules` enables zero-shot `XLT`, as modules trained on the much smaller Norwegian Adressa transfer well to the large English MIND. This suggests that, coupled with multilingual PLMs, MANNeR can be used for effec-

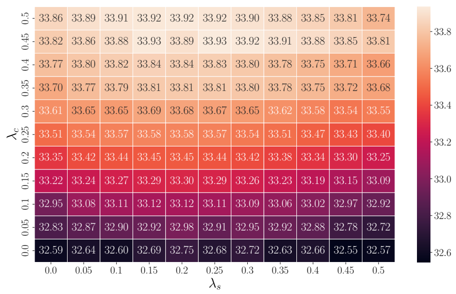
[7]We refer to Fig. 9 for analogous results on Adressa.

[8]We evaluate only the title-based version of MANNeR, as the full version cannot be trained on Adressa.

[9]Figs. 10 and 11 provide similar results for single-aspect personalization and single-aspect customization on MIND, and respectively, Adressa, as target-language datasets.

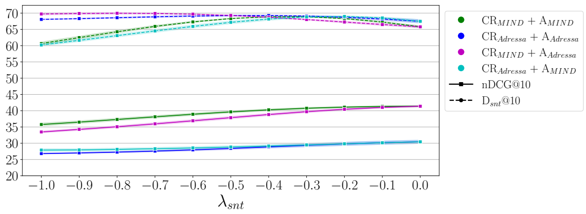


(a) Multi-aspect diversification on MIND.

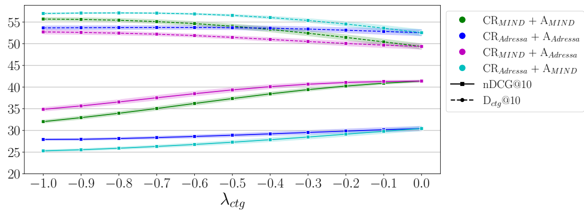


(b) Multi-aspect personalization on MIND.

Figure 4: Results of multi-aspect customization for MANNeR on MIND.



(a) Sentiment diversification.



(b) Category diversification.

Figure 5: `XLT` in single-aspect diversification, with modules trained on different (combinations of) source-language datasets and evaluated on the target-language dataset MIND. The line style indicates the metric, the color the source-language datasets used in training.

tive news recommendation in lower-resource languages, where training data and aspectual labels are scarce. Furthermore, the results demonstrate that the `A-Modules` could be trained on general-purpose classification datasets (e.g. topic or sentiment classification datasets), alleviating the need for aspect-specific annotation of news stories.

5.5 Computational Complexity

While A-Modules add extra parameters, their average training time is two orders of magnitude faster than that of the CR-Module.[10] This is a *one-time* increase in training time: the resulting modules can then be arbitrarily combined for any recommendation goal without additional training. In contrast, all other NNRs require re-training or fine-tuning if the recommendation objective changes as the model weights have to be adjusted each time. This translates into much higher computational costs in practice. We emphasize that MANNeR also has a much lower inference latency due to the (i) CR-Module's lean architecture without a parameterized UE, and (ii) ability to parallelize loading and deploying different modules, for which only the final score has to be combined.[11] Overall, considering both training and inference, MANNeR is more efficient and flexible in a realistic setup with differing recommendation goals that may vary by user or for the same user over time.

6 Conclusion

We proposed MANNeR, a modular framework for multi-aspect neural news recommendation. It learns aspect-specialized NEs with supervised contrastive learning, and linearly combines the corresponding aspect-specific similarity scores for final ranking. Its modular design allows defining ad-hoc multi-aspect ranking functions at inference. Our experiments show that MANNeR consistently outperforms state-of-the-art NNRs on both (i) standard content-based recommendation, and on single- and multi-aspect (ii) diversification and (iii) personalization of recommendations. Moreover, we can identify on-the-fly optimal trade-offs between content-based recommendation performance and aspect-based customization. Equipped with a multilingual PLM, MANNeR can successfully cross-lingually transfer aspect-specific NEs. This supports use cases where aspect-specific labels (e.g., sentiment) are not available for news in the target languages of interest. We hope that our work stimulates more research towards modular, easily extendable, and reusable news recommenders.

[10] On MIND (Adressa), the A-Module for topical category trains 277 (51) times faster and that for sentiment 204 (53) times faster on average per epoch than the CR-Module.

[11] We provide the average inference times in Appendix C.5.

Limitations

MANNeR targets exclusively content-based neural news recommendation and leverages solely textual features. In practice, recommender systems may incorporate content features from various other modalities (e.g., image, video), as well as similarities between users in a collaborative filtering manner. We leave the extension of MANNeR with multi-modal content for future work.

MANNeR independently handles each aspect and aggregates them by weighting the aspect-specific similarities. While it could be argued that direct interactions between different aspects might improve the recommendation performance, training a separate A-Module for each aspect is exactly what drives MANNeR's flexibility. The A-modules allow the user to arbitrarily define the preferences for any concrete recommendation, by defining how much diversification or personalization is desired over each aspect. As illustrated by the results of our experiments, MANNeR outperforms all the state-of-the-art systems, including the ones where additional aspects are directly integrated in the training objective or in the news encoder. MANNeR is thus, besides being drastically more flexible (as it supports arbitrary recommendation objectives at inference time), also more performant, despite the fact that no interactions exist between the aspect modules at training time.

Our framework fully fine-tunes a PLM per aspect-specific module (either for content-relevance in the `CR-Module` or for aspect similarity in the `A-Module`). As all modules share the same PLM as backbone, parameter efficient fine-tuning (PEFT), e.g. LoRA (Hu et al., 2021), would bypass the need to repeatedly load the entire PLM per module into memory. PEFT has been shown to closely meet the performance of full fine-tuning. This represents a key advantage for deploying MANNeR in real-world applications. We however fully fine-tuned models to avoid PEFT as a confounding factor in our experiments. We further chose base-sized PLMs as the backbone of the NE in all models due to computational constraints. While in fine-tuning they remain competitive to large language models (LLMs), the latter may capture richer semantics, which can prove particularly useful for `XLT` applications. With a PEFT approach, MANNeR could easily leverage LLMs without a corresponding increase in computational resources.

Lastly, there exist varied approaches for measuring the descriptive (Castells et al., 2021) and normative (Vrijenhoek et al., 2023) diversity of recommendations. While some of these metrics can be tailored to support arbitrary aspects (i.e., to measure the diversity of recommendations w.r.t. to a particular categorical feature), we opted to quantify aspect-based diversity as generally as possible, leveraging only the distribution of an aspect's values in the recommendation list. We leave exploration of further diversity metrics to future work.

Ethical Considerations

We consider several ethical considerations that arise when working with recommender systems and open benchmark datasets. On the one hand, any biases or misinformation that might exist in the news and user data provided in the public datasets could be propagated through the recommendation pipeline. Similarly, the PLMs used as the recommenders' backbone could contain social biases captured from the training data. On the other hand, the `A-Modules` in MANNeR could be abused to reduce the diversity of recommendations by over-weighting the aspectual-similarity with the user's history, particularly for sensitive aspects such as news stance. This, in turn, could lead to reinforcing the users' existing worldviews and stances (Li and Wang, 2019). Therefore, safeguards should be incorporated in the recommendation models to ensure not only that the outputs are accurate and truthful, but also that the systems are not misused to constrain access to diverse viewpoints.

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. Andreea Iana was supported by the ReNewRS project grant, which is funded by the Baden-Württemberg Stiftung in the Responsible Artificial Intelligence program.

References

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*, pages 336–345.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *ICLR*.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266.

Vincenzo Bonnici. 2020. Kullback-leibler divergence between quantum distributions, and its upper-bound. *arXiv preprint arXiv:2008.05932*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, 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*, pages 2787–2795.

Pablo Castells, Neil Hurley, and Saul Vargas. 2021. Novelty and diversity in recommender systems. In *Recommender systems handbook*, pages 603–646. Springer.

Cheng Chen, Xiangwu Meng, Zhenghua Xu, and Thomas Lukasiewicz. 2017. Location-aware personalized news recommendation with deep semantic analysis. *IEEE Access*, 5:1624–1638.

Seonghwan Choi, Hyeondey Kim, and Manjun Gim. 2022. Do not read the same news! enhancing diversity and personalization of news recommendation. In *Companion Proceedings of the Web Conference 2022*, pages 1211–1215.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jonathan L Freedman and David O Sears. 1965. Selective exposure. In *Advances in experimental social psychology*, volume 2, pages 57–97. Elsevier.

P Moreira Gabriel De Souza, Dietmar Jannach, and Adilson Marques Da Cunha. 2019. Contextual hybrid session-based news recommendation with recurrent neural networks. *IEEE Access*, 7:169185–169203.

Jie Gao, Xin Xin, Junshuai Liu, Rui Wang, Jing Lu, Biao Li, Xin Fan, and Ping Guo. 2018. Fine-grained deep knowledge-aware network for news recommendation with self-attention. In *2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 81–88. IEEE.

Alireza Gharahighehi and Celine Vens. 2023. Diversification in session-based news recommender systems. *Personal and Ubiquitous Computing*, 27(1):5–15.

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*, pages 1042–1048.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*.

Lucien Heitz, Juliane A Lischka, Alena Birrer, Bibek Paudel, Suzanne Tolmeijer, Laura Laugwitz, and Abraham Bernstein. 2022. Benefits of diverse news recommendations for democracy: A user study. *Digital Journalism*, 10(10):1710–1730.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Linmei Hu, Siyong Xu, Chen Li, Cheng Yang, Chuan Shi, Nan Duan, Xing Xie, and Ming Zhou. 2020. Graph neural news recommendation with unsupervised preference disentanglement. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4255–4264.

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*, pages 2333–2338.

Andreea Iana, Mehwish Alam, and Heiko Paulheim. 2024. A survey on knowledge-aware news recommender systems. *Semantic Web*, 15(1):21–82.

Andreea Iana, Goran Glavaš, and Heiko Paulheim. 2023a. Newsreclib: A pytorch-lightning library for neural news recommendation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 296–310.

Andreea Iana, Goran Glavas, and Heiko Paulheim. 2023b. 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*, pages 2384–2388.

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*, pages 18661–18673.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *ICLR*.

Per E Kummervold, Javier De la Rosa, Freddy Wetjen, and Svein Arne Brygfjeld. 2021. Operationalizing a national digital library: The case for a norwegian transformer model. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 20–29.

Jian Li, Jieming Zhu, Qiwei Bi, Guohao Cai, Lifeng Shang, Zhenhua Dong, Xin Jiang, and Qun Liu. 2022. Miner: Multi-interest matching network for news recommendation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 343–352.

Miaomiao Li and Licheng Wang. 2019. A survey on personalized news recommendation technology. *IEEE Access*, 7:145861–145879.

Danyang Liu, Jianxun Lian, Shiyin Wang, Ying Qiao, Jiun-Hung Chen, Guangzhong Sun, and Xing Xie. 2020. KRED: Knowledge-aware document representation for news recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 200–209.

Ping Liu, Karthik Shivaram, Aron Culotta, Matthew A Shapiro, and Mustafa Bilgic. 2021. The interaction between political typology and filter bubbles in news recommendation algorithms. In *Proceedings of the Web Conference 2021*, pages 3791–3801.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Feng Lu, Anca Dumitrache, and David Graus. 2020. Beyond optimizing for clicks: Incorporating editorial values in news recommendation. In *Proceedings of the 28th ACM conference on user modeling, adaptation and personalization*, pages 145–153.

Cornelius A Ludmann. 2017. Recommending news articles in the clef news recommendation evaluation lab with the data stream management system odysseus. In *CLEF (Working Notes)*.

Dan Saattrup Nielsen. 2023. ScandEval: A benchmark for scandinavian natural language processing. In *The 24rd Nordic Conference on Computational Linguistics*.

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*, pages 1933–1942.

Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. penguin UK.

Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021a. Personalized news recommendation with knowledge-aware interactive matching. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 61–70.

Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021b. 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)*, pages 5457–5467.

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*, pages 1917–1921.

Tao Qi, Fangzhao Wu, Chuhan Wu, Peiru Yang, Yang Yu, Xing Xie, and Yongfeng Huang. 2021c. 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)*, pages 5446–5456.

Junyang Rao, Aixia Jia, Yansong Feng, and Dongyan Zhao. 2013. Taxonomy based personalized news recommendation: novelty and diversity. In *Web Information Systems Engineering–WISE 2013: 14th International Conference, Nanjing, China, October 13-15, 2013, Proceedings, Part I 14*, pages 209–218. Springer.

Shaina Raza. 2023. Bias reduction news recommendation system. *Digital*, 4(1):92–103.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Mete Sertkan and Julia Neidhardt. 2022. Exploring expressed emotions for neural news recommendation. In *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, pages 22–28.

Mete Sertkan and Julia Neidhardt. 2023. On the effect of incorporating expressed emotions in news articles on diversity within recommendation models. *decision-making*, 3:11.

Heng-Shiou Sheu and Sheng Li. 2020. Context-aware graph embedding for session-based news recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 657–662.

Hao Shi, Zi-Jiao Wang, and Lan-Ru Zhai. 2022. DCAN: Diversified news recommendation with coverage-attentive networks. *arXiv preprint arXiv:2206.02627*.

Jeong-Woo Son, A-Yeong Kim, and Seong-Bae Park. 2013. A location-based news article recommendation with explicit localized semantic analysis. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 293–302.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Sanne Vrijenhoek, Gabriel Bénédict, Mateo Gutierrez Granada, and Daan Odijk. 2023. Radio*–an introduction to measuring normative diversity in news recommendations. *ACM Transactions on Recommender Systems*.

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*, pages 1835–1844.

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)*, pages 7942–7946. IEEE.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019a. Neural news recommendation with attentive multi-view learning. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3863–3869.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019b. NPA: neural news recommendation with personalized attention. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2576–2584.

Chuhan Wu, Fangzhao Wu, Mingxiao An, Yongfeng Huang, and Xing Xie. 2019c. Neural news recommendation with topic-aware news representation. In *Proceedings of the 57th Annual meeting of the association for computational linguistics*, pages 1154–1159.

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

Chuhan Wu, Fangzhao Wu, and Yongfeng Huang. 2022a. Rethinking InfoNCE: How many negative samples do you need? In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2509–2515. International Joint Conferences on Artificial Intelligence Organization.

Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized news recommendation: Methods and challenges. *ACM Transactions on Information Systems*, 41(1):1–50.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2020a. 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*, pages 44–53.

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*, pages 1652–1656.

Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2022b. End-to-end learnable diversity-aware news recommendation. *arXiv preprint arXiv:2204.00539*.

Chuhan Wu, Fangzhao Wu, Tao Qi, Chenliang Li, and Yongfeng Huang. 2022c. Is news recommendation a sequential recommendation task? In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2382–2386.

Chuhan Wu, Fangzhao Wu, Tao Qi, Wei-Qiang Zhang, Xing Xie, and Yongfeng Huang. 2022d. Removing ai's sentiment manipulation of personalized news delivery. *Humanities and Social Sciences Communications*, 9(1):1–9.

Fangzhao Wu, Ying Qiao, Jiun-Hung Chen, Chuhan Wu, Tao Qi, Jianxun Lian, Danyang Liu, Xing Xie, Jianfeng Gao, Winnie Wu, and Mind Zhou. 2020b. MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3597–3606.

Hongyan Xu, Qiyao Peng, Hongtao Liu, Yueheng Sun, and Wenjun Wang. 2023. Group-based personalized news recommendation with long-and short-term fine-grained matching. *ACM Transactions on Information Systems*.

Jiahao Xun, Shengyu Zhang, Zhou Zhao, Jieming Zhu, Qi Zhang, Jingjie Li, Xiuqiang He, Xiaofei He, Tat-Seng Chua, and Fei Wu. 2021. Why do we click: visual impression-aware news recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3881–3890.

Jingwei Yi, Fangzhao Wu, Chuhan Wu, Ruixuan Liu, Guangzhong Sun, and Xing Xie. 2021. Efficient-fedrec: Efficient federated learning framework for privacy-preserving news recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2814–2824.

Mi Zhang and Neil Hurley. 2008. Avoiding monotony: improving the diversity of recommendation lists. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 123–130.

A Dataset Statistics

	MIND (large)		Adressa (one week)	
Statistic	Train	Test	Train	Test
# News	101,527	72,023	11,207	11,207
# Users	698,365	196,444	96,801	68,814
# Impressions	2,186,683	365,201	218,848	146,284
# Categories	18	17	18	18
Avg. history length	33.7	33.6	13.9	15.6
Avg. # candidates / user	37.4	37.4	21.0	21.0

Table 4: MIND and Adressa dataset statistics.

B Reproducibility Details

B.1 Model Parameters.

		MIND		Adressa	
Model	Non-trainable	Trainable	Total	Trainable	Total
NRMS-PLM	56.7	73	129	126	182
MINER	56.7	68.2	124	121	178
NAML-PLM	56.7	70.8	127	124	180
LSTUR-PLM	56.7	633	690	200	257
MINS-PLM	56.7	73.3	130	126	183
$\text{CAUM}_{\text{no entities}}$-PLM	56.7	73.2	129	126	183
CAUM-PLM	56.7	74.9	131	–	–
TANR-PLM	56.7	70.6	127	123	180
SentiRec-PLM	56.7	73	129	126	182
SentiDebias-PLM	56.7	73.3	130	126	183
MANNeR ($\text{CR-Module}_{\text{title}}$ / $\text{A-Module}_{\text{title}}$) – monolingual	56.7	67.9	124	121	177
MANNeR (CR-Module / A-Module) – monolingual	56.7	70.3	126	–	–
MANNeR ($\text{CR-Module}_{\text{title}}$ / $\text{A-Module}_{\text{title}}$) – multilingual	0	134	134	134	134

Table 5: Number of model parameters (in millions). $\text{CR-Module}_{\text{title}}$ / $\text{A-Module}_{\text{title}}$ denote the MANNeR modules trained with only the news title as input to the NE.

B.2 Hyperparameters and Implementation

Hyperparameter Optimization. We use RoBERTa Base (Liu et al., 2019) and NB-BERT Base (Kummervold et al., 2021; Nielsen, 2023) as the backbone PLMs of all models, in experiments on MIND and Adressa, respectively. In both cases, we fine-tune only the PLM's last four layers.[12] In the cross-lingual transfer experiments from §5.4, we fine-tune all of the 6 layers of DistilBERT. We use 100-dimensional TransE embeddings (Bordes et al., 2013) pretrained on Wikidata as input to the entity encoder in the NE of the knowledge-aware NNRs. We perform hyperparameter tuning on the main hyperparameters of MANNeR and the baselines using grid search. Table 6 lists the search spaces for all the optimized hyperparameters, as well as the best values. We repeat each experiment five times with the seeds ($\{42, 43, 44, 45, 46\}$) set with PyTorch Lightning's `seed_everything`.

Code. We train MANNeR, as well as all the baselines, using the implementations provided in the NewsRecLib library (Iana et al., 2023a).[13]

Infrastructure and Compute. We conduct all experiments on a cluster with virtual machines. We train MANNeR on both datasets, as well as the baselines on MIND, on a single NVIDIA A100 40GB GPU. We train the baselines on Adressa on a single NVIDIA Tesla V100 32GB GPU.

C Additional Results

C.1 Content Personalization

Fig. 6a shows MANNeR's performance on MIND for different inputs to the NE. Even the `CR-Module` exposed to titles only (i.e., no abstract or entity information) outperforms all of the baselines on content recommendation. Fig. 6b illustrates MANNeR's performance for alternative architecture designs and training objectives (cf. §5.1).[14]

C.2 Single-Aspect Customization

Figure 7 explores the trade-off between content and aspect diversification, and respectively, personalization tasks for different values of λ_{ctg} and λ_{snt} on the Adressa dataset. Fig. 8 shows the 2-dimensional t-SNE visualizations (Van der Maaten and Hinton, 2008) of the news embeddings produced with aspect-specialized NEs trained on Adressa.

C.3 Multi-Aspect Customization

Fig. 9 explores the trade-off between content personalization and multi-aspect diversification on Adressa.

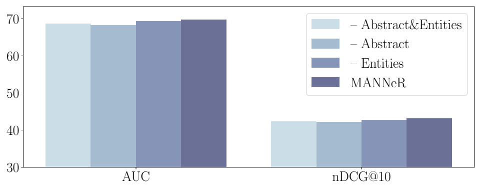
[12] In preliminary results, we did not see significant differences between full fine-tuning of all layers and fine-tuning only the last four layers. In the interest of computational efficiency, we thus froze the first eight layers of the transformer.

[13] https://github.com/andreeaiana/newsreclib

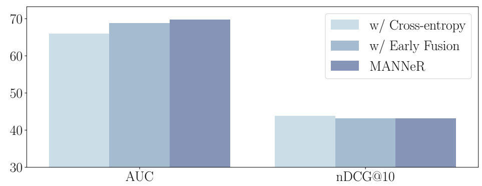
[14] For brevity, we report results on MIND; findings on Adressa exhibit identical trends.

	`lr`	`num`$_{\texttt{heads}}$	`query`$_{\texttt{dim}}$	`UE agg`	K	`score agg`	λ	μ	α	$\tau_{CR-Module}$	$\tau_{A-Module}$
Search Space	$[1e^{-4}, 1e^{-6}]$	{8, 12, 16, 24, 32}	[50, 200]	{ini, con}	{8, 16, 32, 48}	{mean, max, weighted}	[0.1, 0.3]	[5, 15]	[0.05, 0.2]	[0.1, 0.5]	[0.1, 0.9]
Step	$1e^{-1}$	–	50	–	–	–	0.05	5	0.05	0.02	0.05
NRMS-PLM	$1e^{-5}$ / $1e^{-6}$	32 / 8	150 / 200	–	–	–	–	–	–	–	–
MINER	$1e^{-5}$ / $1e^{-6}$	–	–	–	32 / 48	mean / mean	–	–	–	–	–
NAML-PLM	$1e^{-5}$ / $1e^{-6}$	16 / 8	200 / 200	–	–	–	–	–	–	–	–
LSTUR-PLM	$1e^{-5}$ / $1e^{-6}$	24 / 8	150 / 50	ini / ini	–	–	–	–	–	–	–
MINS-PLM	$1e^{-5}$ / $1e^{-6}$	32 / 12	100 / 200	–	–	–	–	–	–	–	–
CAUM-PLM	$1e^{-5}$ / $1e^{-6}$	16 / 16	50 / 150	–	–	–	–	–	–	–	–
TANR-PLM	$1e^{-5}$ / $1e^{-6}$	32 / 8	150 / 50	–	–	–	0.3 / 0.15	–	–	–	–
SentiRec-PLM	$1e^{-5}$ / $1e^{-6}$	32 / 8	200 / 200	–	–	–	–	5 / 5	–	–	–
SentiDebias-PLM	$1e^{-5}$ / $1e^{-6}$	8 / 12	100 / 150	–	–	–	–	–	0.15 / 0.15	–	–
MANNeR	$1e^{-5}$ / $1e^{-6}$	–	200 / 200	–	–	–	–	–	–	0.36 / 0.14	0.9 / 0.9

Table 6: Search spaces used for hyperparameter optimization and best values found for all models. We report the optimal values in the format *value*$_{MIND}$ / *value*$_{Adressa}$. We use the following abbreviations: `lr` = learning rate, `num`$_{\texttt{heads}}$ = number of attention heads, `query`$_{\texttt{dim}}$ = dimensionality of the query vector in additive attention, `UE agg` = aggregation method used to combine the long-term and the short-term user representations into a final user embedding in LSTUR (An et al., 2019), K = number of context codes in MINER (Li et al., 2022), `score agg` = aggregation function for the final user click score calculation in MINER (Li et al., 2022), λ = weight of the topic classification task in TANR (Wu et al., 2019c), μ = weight of the sentiment diversity regularization loss in SentiRec (Wu et al., 2020a), α = adversarial loss coefficient in SentiDebias (Wu et al., 2022d), τ = temperature parameter in SCL in MANNeR, *ini* = initialize, *con* = concatenate, *categ* = category.

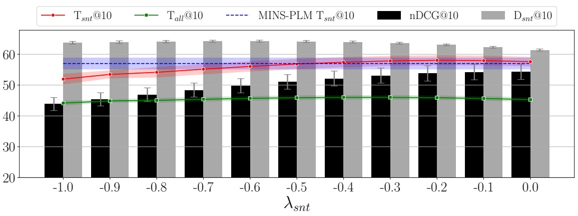


(a) Input features for the News Encoder (NE).

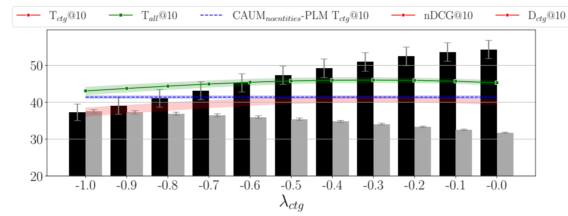


(b) `CR-Module` design/training alternatives.

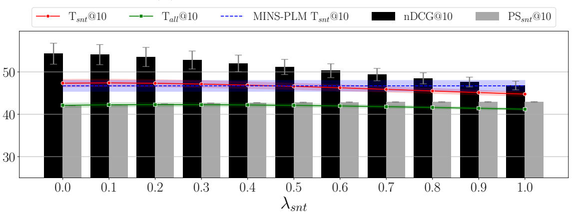
Figure 6: Effect of different (a) NE inputs and (b) model design/training choices on MANNeR's content-based personalization performance.



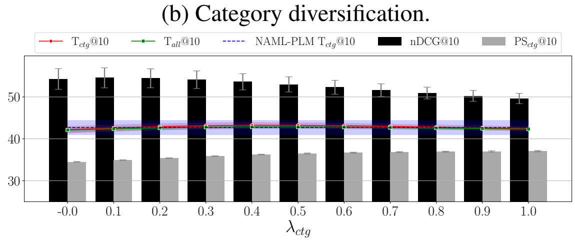
(a) Sentiment diversification.



(b) Category diversification.



(c) Sentiment personalization.



(d) Category personalization.

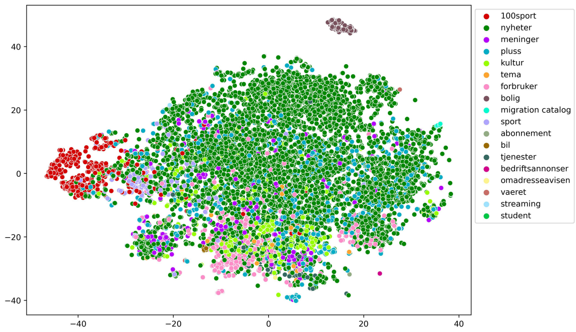
Figure 7: Results of single-aspect customization for MANNeR and the best baseline on Adressa.

C.4 Cross-Lingual Transfer

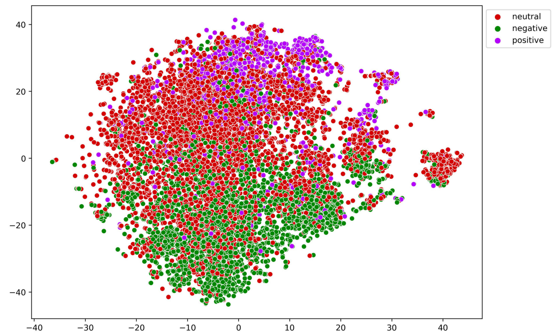
Fig. 10 summarizes the `XLT` results for single-aspect personalization on the target-language dataset MIND, whereas Fig. 11 shows the analogous `XLT` results for single-aspect diversification and personalization, respectively, on the target-language dataset Adressa.

C.5 Time Complexity Analysis

Table 7 shows the average inference time for the entire MIND (365,201 impressions), and respectively, Adressa (146,284 impressions) test sets. Note that runtimes heavily depend on the computing infrastructure used, as well as on the parallel usage of the infrastructure for other tasks, as experiments are conducted on a HPC cluster. We highlight that

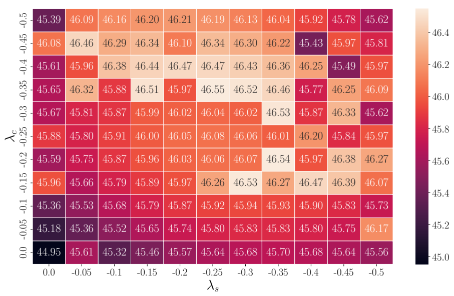


(a) Category-shaped embedding space.

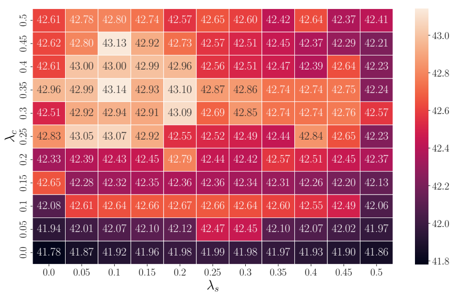


(b) Sentiment-shaped embedding space.

Figure 8: t-SNE plots of the news embeddings in the test set of Adressa.

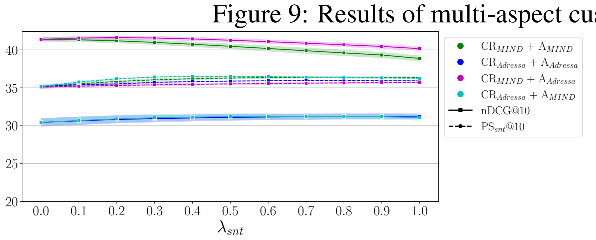


(a) Multi-aspect diversification on Adressa.

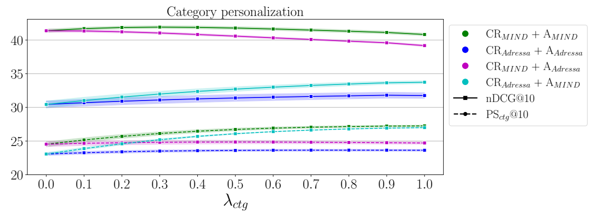


(b) Multi-aspect personalization on Adressa.

Figure 9: Results of multi-aspect customization for MANNeR on Adressa.

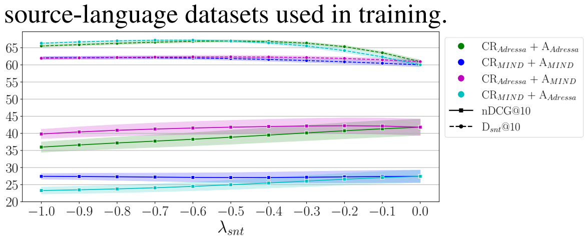


(a) Sentiment personalization.

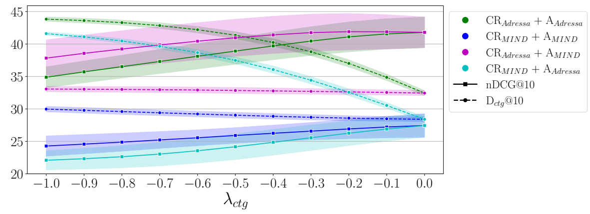


(b) Category personalization.

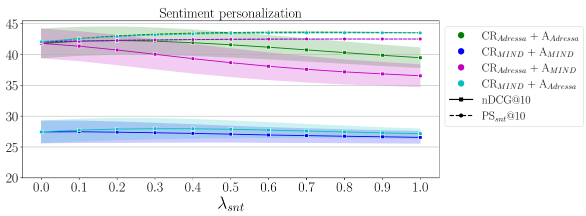
Figure 10: XLT in single-aspect personalization, with modules trained on different (combinations of) source-language datasets and evaluated on the target-language dataset MIND. The line style indicates the metric, the color the source-language datasets used in training.



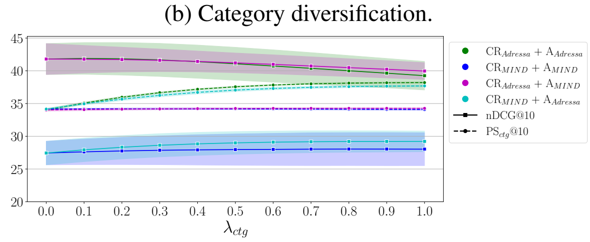
(a) Sentiment diversification.



(b) Category diversification.



(c) Sentiment personalization.



(d) Category personalization.

Figure 11: XLT in single-aspect diversification and personalization, with modules trained on different (combinations of) source-language datasets and evaluated on the target-language dataset Adressa. The line style indicates the metric, the color the source-language datasets used in training.

MANNeR achieves a much lower inference time than the other NNRs.

Model	MIND	Adressa
NRMS-PLM	$17.53_{\pm 0.48}$	$7.13_{\pm 0.27}$
MINER	$16.03_{\pm 1.66}$	$9.96_{\pm 0.73}$
NAML-PLM	$33.99_{\pm 0.51}$	$7.09_{\pm 0.14}$
MINS-PLM	$27.50_{\pm 10.87}$	$7.81_{\pm 0.21}$
$\text{CAUM}_{\text{no entities}}$-PLM	$22.67_{\pm 2.46}$	$8.12_{\pm 0.13}$
CAUM-PLM	$25.22_{\pm 0.45}$	–
TANR-PLM	$17.02_{\pm 1.07}$	$6.98_{\pm 0.08}$
SentiRec-PLM	$17.93_{\pm 0.34}$	$7.02_{\pm 0.08}$
SentiDebias-PLM	$21.01_{\pm 3.03}$	$13.28_{\pm 0.83}$
MANNeR (CR-Module)	$1.34_{\pm 0.03}$	$2.09_{\pm 0.06}$
MANNeR (CR-Module + *ctg* A-Module)	$1.68_{\pm 0.08}$	$2.78_{\pm 0.10}$
MANNeR (CR-Module + *snt* A-Module)	$1.65_{\pm 0.01}$	$2.73_{\pm 0.06}$
MANNeR (CR-Module + 2 A-Modules)	$2.13_{\pm 0.05}$	$3.17_{\pm 0.01}$

Table 7: Inference time (in thousands of seconds) for the different NNRs on the test portions of the MIND and Adressa datasets, respectively.