Multi-Aspect Matching between Disentangled Representations of User Interests and Content for News Recommendation

Yingzhi Miao[1], Martin Pavlovski[2], Zhiqiang Chen[1], and Fang Zhou[1],(✉)

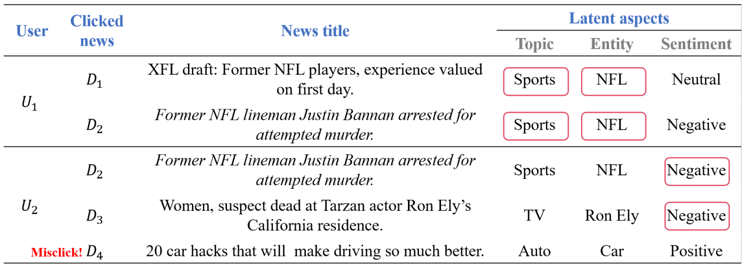
[1] School of Data Science and Engineering,
East China Normal University, Shanghai, China
{yzmiao21,zqchen}@stu.ecnu.edu.cn, fzhou@dase.ecnu.edu.cn
[2] Temple University, Philadelphia, USA
martin.pavlovski@temple.edu

Abstract. Personalized news recommendation is a crucial technique to help users find the content of interest from massive news. While most news recommendation approaches learn a single representation for both users and news, they overlook the nuanced diversity of user interests. Some recent works focused on learning multi-aspect representations of user interests. However, they ignore that news can encompass various aspects of a user's interests, failing to capture the intricate interactions between news content and user preferences. Meanwhile, a user could occasionally click on some news by mistake. In this paper, we propose a novel news recommendation model which learns disentangled representations for both user interests and news content. This allows for capturing the characteristics of different aspects of news content and user interests. An aspect-wise matching is then applied to capture the fine-grained interactions between news and users. A disentanglement loss is proposed to encourage independence of different aspects. Furthermore, we leverage contrastive learning on a news-level to emphasize the aspect-related information as well as on a user-level to mitigate the impact of misclicked news and thus further improve the model's robustness. Extensive experiments on two real-world datasets demonstrate the effectiveness of our model.

Keywords: News recommendation · Disentangled representation learning · Multi-aspect matching.

1 Introduction

With the development of online products and networks, a plethora of people are habituated to reading news on online platforms, such as Microsoft News and Google News [21]. However, the amount of news articles generated on online platforms every day is massive, making it rather difficult for people to find the news articles they are really interested in, which greatly affects the user experience. Thus, personalised news recommendation gets more and more attention in the recent years [19].



User	Clicked news	News title	Latent aspects		
			Topic	Entity	Sentiment
U_1	D_1	XFL draft: Former NFL players, experience valued on first day.	Sports	NFL	Neutral
	D_2	*Former NFL lineman Justin Bannan arrested for attempted murder.*	Sports	NFL	Negative
U_2	D_2	*Former NFL lineman Justin Bannan arrested for attempted murder.*	Sports	NFL	Negative
	D_3	Women, suspect dead at Tarzan actor Ron Ely's California residence.	TV	Ron Ely	Negative
	Misclick! D_4	20 car hacks that will make driving so much better.	Auto	Car	Positive

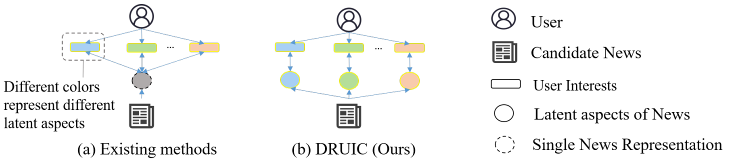
Fig. 1. Examples of two users' historically clicked news. The topic, entity and sentiment are latent aspects of users' interests. A red rectangle indicates that a user clicked a news title because of a certain aspect.

The key to personalised news recommendation is to model the news features and user interests. Explicit user feedback such as reviews and ratings is typically absent on the news platforms, thus most works are based on implicit feedback, such as clicks, to learn user interests [19, 20]. Nonetheless, these methods adopt single representations for both news and users, neglecting the diversity of users' interests and news content. Recently, an increasing number of works focus on exploring multi-aspect representations learning [16, 14, 17, 8, 18]. However, these efforts concentrate solely on modeling multi-aspects for either news content or user interests. They overlook the consideration that a news article may encompass multiple latent aspects related to user interests, and a user might click on a news article due to one or more aspects. As illustrated in Figure 1, the user U_1 clicked on the news D_1 and D_2 because of his/her interest in their corresponding topic and entity, while the user U_2 clicked on the news D_2 and D_3 due to his/her interest in their sentiment. Despite both U_1 and U_2 clicked on D_2, their interests are attributed to different aspects. *Typically, obtaining direct labels for various aspects of users' interests is challenging, and there may be undiscovered latent aspects within the vast amount of news data.*

Furthermore, as illustrated in Figure 2, existing methods addressing the diversity of user interests [17, 8, 18], often employ a single representation for a candidate news article. Despite disentangling user interests, these methods struggle to capture the fine-grained matching between user interests and news content for each aspect. In addition, a user may occasionally misclick on a news title that he or she is not genuinely interested in. For example, consider the user U_2 who prefers to read news with a negative sentiment (see Figure 1); then the news D_4 with a positive sentiment is clicked by mistake. In real-world application scenarios, *the absence of explicit labels to indicate genuine clicks and misclicks poses an additional challenge.* As a result, misclicked news articles can potentially impact the learning of a user's real interests.

To tackle these challenges, in this paper, we propose a model for learning **D**isentangled **R**epresentations of **U**ser **I**nterests and **C**ontent (DRUIC)[3], aimed at capturing diverse user interests w.r.t. multiple hidden aspects of news. Instead of using a single vector to represent news and a user, DRUIC separately learns disentangled representations for news and a user based on the news titles

[3] Source code: https://github.com/myz000/DRUIC.


Different colors represent different latent aspects
...
User
Candidate News
User Interests
Latent aspects of News
Single News Representation

(a) Existing methods

(b) DRUIC (Ours)

Fig. 2. (a) Existing methods match multiple user interests to a single representation of a candidate news [17, 8, 18]. (b) Our method performs fine-grained matching of user interests with news content across every aspect.

and the user's historically clicked news. Subsequently, an aspect-wise matching is applied to examine the fine-grained interactions between news content and user interests. To enhance the independence of various aspects, a disentanglement loss is proposed to enforce the learning of disentangled representations of news that are close to their corresponding aspect centers and away from the other aspects' centers. Furthermore, to ensure the model's capability of capturing aspect-related features of the news and mitigate the impact of noisy words, we propose a news-level contrastive loss that maximizes the agreement between two masked news that are generated by randomly masking subsets of words from the same news' title. Analogously, to improve the robustness of our model, we propose a user-level contrastive loss which maximizes the agreement between two masked variants of a user-clicked news sequence. By introducing the user-level contrastive loss, even if there are some misclicked news in the user's clicked news list, the model would still be capable of capturing the user's true interests.

To summarize, the main contributions of this paper are:

- We propose a novel model to learn disentangled representations for both user interests and news content, which allows for (1) capturing diverse user interests w.r.t. multiple aspects hidden in the news content and (2) capturing fine-grained matching between user interests and news content.
- We propose a disentanglement loss to promote the learning of independent news representations, a news-level contrastive loss to enforce the proposed model to focus on aspect-related information, and a user-level contrastive loss to mitigate the impact of misclicked news.
- We conduct extensive experiments on two real-world datasets, the results of which demonstrate the effectiveness of our proposed model.

2 Related work

News recommendation is an important task that has been widely explored in the recent years. Its aim is to select news articles that users may have interest in from a massive set of candidate news [21]. Numerous studies have invested significant efforts in exploring various deep neural networks, including GRU [1], CNN [15], and attention-based networks [20], to learn representations of news and users. Most of these methods learn a single representation for both users and news. Recently, increasing works proposed to model fine-grained representations of either news or users. Some works focused on learning multiple representations of words or entire news articles [16], but they did not consider learning

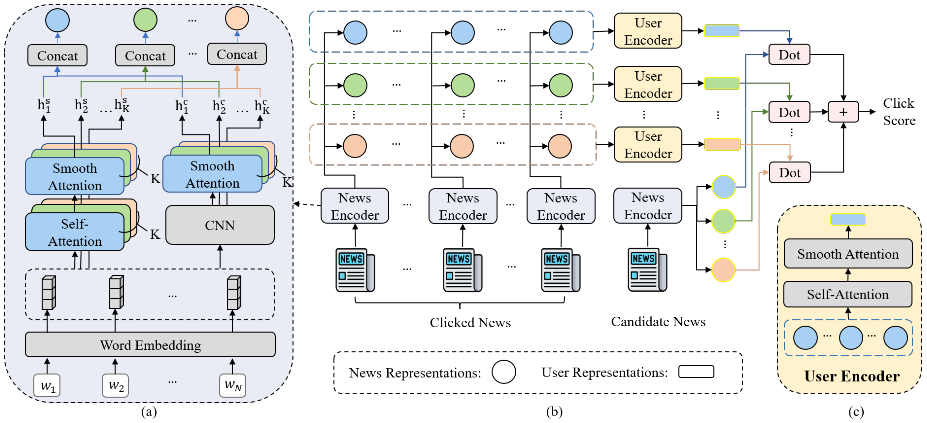


Fig. 3. Overview of DRUIC. (a) Illustration of the news encoder. (b) Framework of DRUIC. (c) Illustration of the user encoder.

diverse representations at the user interest level. Other works aimed to learn multiple representations of users interests [8]. However, most of these methods assume that each news article is tied to a single aspect of user interests, such as a specific topic. They typically leverage a single representation of news, and then extract multiple user interests from a user's historical clicked news. Unlike these methods, we propose to (1) learn disentangled news representations from the news' content, (2) model multiple user interests from the corresponding disentangled representations of a user's history of clicked news, and (3) apply an aspect-wise matching to capture the fine-grained interactions between (1) and (2). Other recent works consider incorporating additional information, like entities mentioned in news [10]. However, such additional information is not easily accessible. Thus, the setting considered in this work relies only on news' titles.

3 Problem definition

A news title d is composed of a word sequence $[w_1, w_2, ..., w_N]$, where N is the number of words in the news title. For a user u, his/her historically clicked news, ordered by their respective timestamps, can be denoted as $s_u = [d_1, d_2, ..., d_T]$, where T is the number of news clicked by the user u. Given a sequence of user's historically clicked news s_u and a candidate news title d, the goal is to calculate the click score y which indicates the user's interest in the news.

4 Method

In this section, we will introduce the proposed model DRUIC in detail. Figure 3 shows the overall architecture of the model. It contains three modules. The first module is a *disentangled news encoder* designed to capture the news features w.r.t. multiple aspects, and a *disentanglement loss* is proposed to force the disentanglement of different aspects. The second module incorporates multiple *disentangled user encoders* to learn a disentangled user interest representation for each aspect. The third module performs *click prediction* by calculating a click score y. In addition, a *news-level contrastive loss* is proposed to enforce the news encoder to focus on aspect-related words and a *user-level contrastive loss* is proposed to mitigate the impact of misclicked news. Next, we introduce each module in detail.

Disentangled news encoder. Since news contain a variety of information that can influence user interests, the goal is to learn K representations to represent multiple aspects of a news title, where K is the number of latent aspects. As illustrated in Figure 3(a), a news title d is first transformed into a sequence of vectors $E = [e_1, e_2, ..., e_N]$ by a word embedding layer. Then, the vectors E are fed into a *self-attention-based news feature extractor (SA-FE)* which consists of a multi-head self-attention (MSA) network and multiple smooth attention (SAT) networks aimed to capture the global information of news. Each head of the MSA network is combined with a SAT network to capture the features w.r.t. an aspect. SAT can prevent the model from using a single word to represent a sequence of words, thus increasing the robustness of the model. For the k^{th} aspect, the representations of words from an article's title, i.e., $[r^s_{1,k}, r^s_{2,k}, ..., r^s_{N,k}]$, can be obtained through the k^{th} head of MSA; after which the k^{th} SAT is applied to select informative words, and the weight of i^{th} word in the title is computed as $\alpha^s_{i,k} = \frac{exp(a^s_{i,k})}{\sum_{j=1}^{N} exp(a^s_{j,k})}$, $a^s_{i,k} = \mu \cdot tanh({q^s_k}^T(Q^s_k \times r^s_{i,k} + b^s_k))$, where Q^s_k and b^s_k are projection parameters, q^s_k is the query vector, and μ is a smooth factor. The representation of a news title for the k^{th} aspect is calculated as the weighted sum of words representations, i.e., $h^s_k = \sum_{i=1}^{N} \alpha^s_{i,k} r^s_{i,k}$. Then, one group of K representations $(h^s_1, h^s_2, ..., h^s_K)$ of a single news title can be obtained.

To enforce the news representations to be independent of each other, we propose a disentanglement loss that uses the representations' *centers* to guide the representations learning. It can not only focus on distinguishing vectors from different aspects but also promote similarity between vectors of the same aspect. We denote the center representations of the K aspects as $(o^s_1, o^s_2, ..., o^s_K)$. For the k^{th} representation of a news title $h^s_{i,k}$, we calculate the similarity between $h^s_{i,k}$ and its corresponding center o^s_k, i.e., $P(h^s_{i,k}, o^s_k) = \frac{exp(sim(h^s_{i,k}, o^s_k))}{\sum_{j=1}^{K} exp(sim(h^s_{i,k}, o^s_j))}$, and $sim(h^s_k, o^s_k) = \frac{h^s_k o^s_k}{\left\|h^s_k\right\|_2 \left\|o^s_k\right\|_2}$. The disentanglement loss for the *SA-FE* is formulated as $\mathcal{L}^s_{dis} = -\frac{1}{K} \sum_{k=1}^{K} \sum_{i \in D} lnP(h^s_{i,k}, o^s_k)$, where D is a set that contains all training news titles. The disentanglement loss enforces $h^s_{i,k}$ to be similar to its corresponding center and further apart from the other centers. Similarly, to capture the local information of news, we fedd the vectors E into a *CNN-based news feature extractor (CNN-FE)* composed of a CNN network and multiple SAT networks, yielding another group of K representations $(h^c_1, h^c_2, ..., h^c_K)$. The disentanglement loss for the *CNN-FE*, denoted as $\mathcal{L}^c_{dis}$, is formulated in the same manner as $\mathcal{L}^s_{dis}$, by replacing the superscript s with c. The final disentanglement loss is formulated as $\mathcal{L}_{dis} = \mathcal{L}^s_{dis} + \mathcal{L}^c_{dis}$. The final representation of a news title is $h = (h_1, h_2, ..., h_K)$, where h_k is the concatenation of the two groups of representations, i.e., $h_k = [h_k^s, h_k^c]$.

Disentangled user encoder. Given a user historically clicked news $[d_1, d_2, ..., d_T]$, through the disentangled news encoder, we obtain the disentangled representations $[h_{i,1}, h_{i,2}, ..., h_{i,K}]$ for d_i. To model the multiple user interests w.r.t. different aspects, as Figure 3(b) shows, we applied K parallel disentangled user encoders to learn the interests representations of a single user, where the input of the k^{th} user encoder is the sequence of the k^{th} representations of all clicked news, i.e., $[h_{1,k}, h_{2,k}, ..., h_{T,k}]$. For the k^{th} disentangled user encoder, as Figure 3(c) shows, it first adopts a self-attention network to capture the interactions of clicked news. Then, it uses a SAT to learn the k^{th} user interest representation u_k by aggregating news representations. Through the multiple disentangled user encoders, the representations of a user, denoted as $(u_1, u_2, ..., u_K)$, are obtained.

News-level contrastive learning. To guide the model to focus on the aspect-related information, we propose a news-level contrastive loss. The objective is to ensure that the news encoder retains the ability to capture the information w.r.t. various aspects of the news, even when certain words in the title are masked. For a clicked news d, we generate two positive samples, d^1 and d^2, by randomly masking a subset of words in d. The limitation of many general contrastive learning methods [2] is that they consider other news as negative samples. However, we cannot guarantee that other news are different from d in certain aspects. Inspired by a recent work [3] on contrastive learning frameworks that exclusively rely on positive samples, we propose to solely utilize positive samples of news within the Simsiam architecture [3].

Note that the *SA-FE* and *CNN-FE* learn different features, independently. Therefore, we leverage news-level contrastive learning for them separately. For d^1 and d^2, through the *SA-FE*, we obtain their concatenated representations $h^{s,1} = [h_1^{s,1}, h_2^{s,1}, ..., h_K^{s,1}]$ and $h^{s,2} = [h_1^{s,2}, h_2^{s,2}, ..., h_K^{s,2}]$. Then, the news-level contrastive loss for *SA-FE* is constructed as: $\mathcal{L}^s_{news\_ctr} = -\frac{1}{T}\sum_{u \in U}\sum_{i=1}^{T}(\frac{1}{2}sim(h^{s,1}_{u,i}, stop\_g(predict(h^{s,2}_{u,i})) + \frac{1}{2}sim(h^{s,2}_{u,i}, stop\_g(predict(h^{s,1}_{u,i}))))$, where *sim* is the cosine similarity, U are all training users in a batch, *stop_g* is a stop-gradient operation, and *predict* is an MLP network. *stop_g* and *predict* are the key components proposed in Simsiam. Similarly, the news-level contrastive loss for the *CNN-FE*, denoted as $\mathcal{L}^c_{news\_ctr}$, can be constructed in the same manner as $\mathcal{L}^s_{news\_ctr}$ by replacing the superscript s with c. The final news-level contrastive loss is the summation of $\mathcal{L}^s_{news\_ctr}$ and $\mathcal{L}^c_{news\_ctr}$, that is $\mathcal{L}_{news\_ctr} = \mathcal{L}^s_{news\_ctr} + \mathcal{L}^c_{news\_ctr}$.

User-level contrastive learning. In the real-world application scenarios, users sometimes mistakenly click on news that they are not really interested in, and then misclicked news may affect the learning of the users' genuine interests. To mitigate the impact of misclicked news and improve the model's robustness, we propose a user-level contrastive loss which minimizes the similarities of pairs of randomly masked title sequences. The intuition is that, despite of randomly masking certain news in a user's sequence of clicked news, the model can still capture the user's interests w.r.t. each aspect. More precisely, given a training batch with U users, for the i^{th} user's clicked sequence $s_{u_i} = [d_1, d_2, ..., d_T]$, we generate two variants of it, $s_{u_i}^1$ and $s_{u_i}^2$, by randomly masking some clicked news, respectively for each variant. Then, through the user encoders, the representations of these two variants can be obtained as $\hat{u}_i^1 = [u_{i,1}^1, u_{i,2}^1, ..., u_{i,K}^1]$ and $\hat{u}_i^2 = [u_{i,1}^2, u_{i,2}^2, ..., u_{i,K}^2]$. Assuming that the interests of users are different from each other, the representations of other users can be used as negative samples. The variants generated from the same user's clicked news sequence are taken as positive samples, while the variants generated from other users' sequences in the same batch are considered as negative samples. Then, for each variant, there is one positive sample and $2 * (U - 1)$ negative samples. For the i^{th} user, the user-level contrastive loss is formulated as: $\mathcal{L}_{u\_ctr}(\hat{u}_i^1, \hat{u}_i^2) + \mathcal{L}_{u\_ctr}(\hat{u}_i^2, \hat{u}_i^1)$, where $\mathcal{L}_{u\_ctr}(\hat{u}_i^1, \hat{u}_i^2) = -log \frac{exp(sim(\hat{u}_i^1, \hat{u}_i^2))}{\sum_{j \in U, j \neq i} \sum_{l=[1,2]} exp(sim(\hat{u}_i^1, \hat{u}_j^l))}$ and sim is the cosine similarity. The final user-level contrastive loss is the summation of the contrastive losses for all users, that is: $\mathcal{L}_{user\_ctr} = \sum_{i \in U} \mathcal{L}_{u\_ctr}(\hat{u}_i^1, \hat{u}_i^2) + \mathcal{L}_{u\_ctr}(\hat{u}_i^2, \hat{u}_i^1)$.

Click prediction. Given a user's clicked history news, his multiple latent interests $[u_1, u_2, ..., u_K]$ can be obtained through the disentangled user encoders. For a candidate news, through the disentangled news encoders , its multiple disentangled representations $[h_1, h_2, ..., h_K]$ can be obtained. Then the click probability score is computed as the summation of the inner products between corresponding user representations and news representations, which can be denoted as $y = \sum_{k=1}^{K} (u_k)^{\top} h_k$.

Model training. Upon obtaining the click probability scores, following [20], we use negative sampling techniques for training. For each clicked news d^+ in the training dataset S, we randomly select g non-clicked news $[d_1^-, d_2^-, ..., d_g^-]$ from the same set of news viewed by a user. For a given clicked news and its g corresponding non-clicked news, their click probability scores are denoted as $[y^+]$ and $[y_1^-, y_2^-, ..., y_g^-]$. Then the recommendation loss is formulated as: $\mathcal{L}_{rec} = -\sum_{i \in S} log \frac{exp(y_i^+)}{exp(y_i^+) + \sum_{j=1}^{g} exp(y_j^-)}$. Finally, the overall training loss can be written as: $\mathcal{L} = \mathcal{L}_{rec} + \lambda_1 \mathcal{L}_{dis} + \lambda_2 \mathcal{L}_{news\_ctr} + \lambda_3 \mathcal{L}_{user\_ctr} + \lambda_4 \|\Theta\|_2$, where Θ are all trainable model parameters; λ_1, λ_2, λ_3 and λ_4 are hyper-parameters that control the importance of different loss.

5 Experiments

5.1 Dataset and experimental settings

Dataset and baselines. We evaluate the performance of various methods on two widely used and publicly available datasets, MIND-Small [21] and Adressa-1week [5]. We compared our model with some state-of-the-art recommendation methods: LibFM [13], Wide&Deep [4], DeepFM [6], DSSM [7], DKN [15], DAN [22], NPA [19], NRMS [20], LSTUR [1], FIM [14], FUM [11], CAUM [12], MCCM [16], MINS [17], MIECL [18], and MINER [8]. For fair comparison, we only used news' titles in the experiments for all methods.

Settings. In the experiments, all news were preprocessed by removing digits and stop words and we filtered out users having clicked news sequences shorter

Table 1. The news recommendation performance of different methods. '*' indicates that our model have a significant improvement at $p < 0.05$ over this method. The type 'X' indicates absence of disentangled representation learning; 'N' and 'U' symbolize the disentangling of news content and user interests, respectively.

Model	Type	MIND				Adressa			
		AUC	MRR	nDCG@5	nDCG@10	AUC	MRR	nDCG@5	nDCG@10
LibFM	X	59.88*	24.75*	26.42*	33.02*	62.85*	20.21*	28.10*	30.59*
Wide&Deep	X	61.01*	25.71*	27.61*	34.28*	72.58*	32.67*	51.21*	51.71*
DeepFM	X	62.06*	26.91*	29.50*	35.75*	73.72*	32.82*	51.38*	52.04*
DSSM	X	61.47*	26.95*	29.15*	35.82*	62.18*	27.29*	42.70*	41.92*
DKN	X	64.34*	28.92*	31.93*	38.29*	76.05*	34.89*	54.82*	55.38*
DAN	X	66.63*	29.57*	32.63*	39.19*	77.75*	35.39*	55.64*	56.42*
NPA	X	66.54*	30.14*	33.35*	39.79*	72.04*	31.97*	50.33*	50.12*
NRMS	X	66.85*	30.81*	34.10*	40.51*	75.93*	36.27*	56.80*	57.38*
LSTUR	X	68.05*	31.31*	34.75*	41.17*	77.50*	35.94*	56.21*	57.14*
FUM	X	68.07*	31.05*	34.65*	41.14*	74.29*	37.01*	58.19*	58.34*
CAUM	X	67.76*	30.82*	34.22*	40.82*	75.42*	37.10*	58.37*	58.58*
FIM	N	66.42*	29.79*	32.82*	39.44*	75.98*	36.18*	57.03*	57.38*
MCCM	N	68.13*	31.42*	35.03*	41.40*	76.51*	35.02*	54.65*	55.33*
MINS	U	67.20*	31.29*	34.49*	40.68*	78.45*	36.47*	56.75*	57.45*
MIECL	U	67.27*	30.08*	33.32*	39.98*	78.48*	36.17*	56.56*	57.45*
MINER	U	67.96*	31.21*	34.78*	41.20*	76.76*	34.52*	54.02*	54.84*
DRUIC	N+U	**69.59**	**32.73**	**36.52**	**42.77**	**82.60**	**37.31**	**58.55**	**59.63**

Table 2. DRUIC's performance measured when ablating different losses.

Model	AUC	MRR	nDCG@5	nDCG@10
DRUIC	**69.59**	**32.73**	**36.52**	**42.77**
-$\mathcal{L}_{dis}$-$\mathcal{L}_{news\_ctr}$-$\mathcal{L}_{user\_ctr}$	68.05 ↓2.21%	31.20 ↓4.67%	34.67 ↓5.07%	41.05 ↓4.02%
-$\mathcal{L}_{news\_ctr}$-$\mathcal{L}_{user\_ctr}$	68.19 ↓2.01%	31.22 ↓4.61%	34.74 ↓4.87%	41.11 ↓3.88%
-$\mathcal{L}_{dis}$-$\mathcal{L}_{user\_ctr}$	68.15 ↓2.07%	31.30 ↓4.37%	34.93 ↓4.33%	41.20 ↓3.65%
-$\mathcal{L}_{dis}$-$\mathcal{L}_{news\_ctr}$	68.90 ↓0.99%	32.21 ↓1.59%	35.87 ↓1.78%	42.14 ↓1.47%
-$\mathcal{L}_{user\_ctr}$	68.26 ↓1.91%	31.44 ↓3.94%	35.05 ↓4.03%	41.33 ↓3.37%
-$\mathcal{L}_{dis}$	68.99 ↓0.86%	32.37 ↓1.10%	36.04 ↓1.31%	42.29 ↓1.12%
-$\mathcal{L}_{news\_ctr}$	69.33 ↓0.37%	32.49 ↓0.73%	36.22 ↓0.82%	42.48 ↓0.68%

than 10 and 5, from the MIND and Adressa datasets, respectively. The word embeddings for the MIND dataset were initialized using pre-trained Glove embedding vectors [9], whereas the word embeddings for the Adressa dataset were randomly initialized. The output dimension of all self-attention networks was set to 300. We set the number of aspects K to 5 for the MIND dataset and 8 for the Adressa dataset. In both news-level contrastive learning and user-level contrastive learning, the mask ratios were uniformly set to 0.5. The model was optimized using the Adam optimizer with a learning rate of 0.0001, λ_1=0.5, λ_2=0.3, λ_3=0.1, λ_4=0.0001, the negative sampling ratio g=4 and a batch size of 32. All hyper-parameters were tuned on the validation dataset. We repeated each experiment 10 times and reported the average results. Following [21], we used AUC, MRR, and nDCG for evaluation.

5.2 Results

Performance comparison. Table 1 shows the news recommendation performances of DRUIC and the baselines. We have several observations: (1) The news recommendation methods such as DAN, NPA and NRMS show superior performance compared to general recommendation methods including LibFM, Wide&Deep and DSSM. (2) DRUIC statistically significantly outperforms other

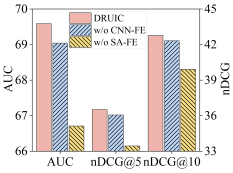


Fig. 4. DRUIC's performance measured when ablating different news feature extractors.

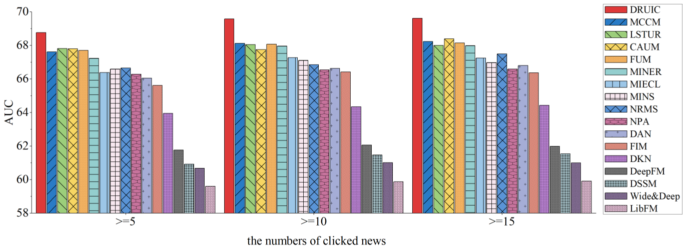


Fig. 5. Performances obtained under different minimum length limits of clicked news sequences.

news recommendation methods which use only a single representation to represent news and user interest, such as NRMS, LSTUR and CAUM. (3) DRUIC surpasses news recommendation methods that learn multiple representations of one of news and users, such as FIM, MCCM, and MINS. This superiority stems from the fact that these methods solely focus on disentangled learning of either news or users, while DRUIC considers both aspects comprehensively.

Ablation study. Next, we conduct two ablation studies on the MIND dataset. (1) We investigate how each loss contributes to the overall performance of DRUIC. The results are summarized in Table 2. We observed that removing any loss either individually, or in combination with other losses, results in a decline in performance. In addition, with more losses removed, the performance tends to decline more. (2) We investigate how the CNN-based news feature extractor (CNN-FE) and the Self-Attention-based news feature extractor (SA-FE) contribute to the overall performance of DRUIC. We remove each of them to assess how DRUIC's performance is affected. We can observe from Figure 4 that removing either extractor leads to a performance drop suggesting that all news feature extractors are necessary.

Robustness to different minimum sequence lengths. Next, we change the minimum sequence length of clicked news histories to investigate the robustness of DRUIC on the MIND dataset. We separately set this limit to 5, 10 and 15. For each limit, we filter out the users whose clicked news sequence length is less than the limit value. The results are presented in Figure 5. We can observe that DRUIC consistently performs better than the other baselines. Besides, DRUIC's performance is higher when the limit value is larger, meaning that DRUIC is better suited for longer user sequences.

6 Conclusion

In this paper, we introduce a novel news recommendation method which learns disentangled representations for both user interests and news content w.r.t. multiple aspects. We propose a disentanglement loss, to enforce the disentanglement of representations of news. Furthermore, we additionally leverage a news-level contrastive loss to mitigate the impact of noisy words in news' titles, and a user-level contrastive loss to improve the model's robustness.

References

1. An, M., Wu, F., Wu, C., Zhang, K., Liu, Z., Xie, X.: Neural news recommendation with long-and short-term user representations. In: ACL. pp. 336–345 (2019)
2. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML (2020)
3. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758 (2021)
4. Cheng, H.T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., et al.: Wide & deep learning for recommender systems. In: DLRS. pp. 7–10 (2016)
5. Gulla, J.A., Zhang, L., Liu, P., Özgöbek, O., Su, X.: The adressa dataset for news recommendation. In: WI. p. 1042–1048 (2017)
6. Guo, H., Tang, R., Ye, Y., Li, Z., He, X.: Deepfm: A factorization-machine based neural network for ctr prediction. In: IJCAI. p. 1725–1731 (2017)
7. Huang, P.S., He, X., Gao, J., Deng, L., Acero, A., Heck, L.: Learning deep structured semantic models for web search using clickthrough data. In: CIKM. p. 2333–2338 (2013)
8. Li, J., Zhu, J., Bi, Q., Cai, G., Shang, L., Dong, Z., Jiang, X., Liu, Q.: MINER: Multi-interest matching network for news recommendation. In: Findings of ACL. pp. 343–352 (2022)
9. Pennington, J., Socher, R., Manning, C.: GloVe: Global vectors for word representation. In: EMNLP. pp. 1532–1543 (2014)
10. Qi, T., Wu, F., Wu, C., Huang, Y.: Personalized news recommendation with knowledge-aware interactive matching. In: SIGIR. pp. 61–70 (2021)
11. Qi, T., Wu, F., Wu, C., Huang, Y.: Fum: Fine-grained and fast user modeling for news recommendation. In: SIGIR. p. 1974–1978 (2022)
12. Qi, T., Wu, F., Wu, C., Huang, Y.: News recommendation with candidate-aware user modeling. In: SIGIR. p. 1917–1921. New York, NY, USA (2022)
13. Rendle, S.: Factorization machines with libfm. TIST pp. 1–22 (2012)
14. Wang, H., Wu, F., Liu, Z., Xie, X.: Fine-grained interest matching for neural news recommendation. In: ACL. pp. 836–845 (2020)
15. Wang, H., Zhang, F., Xie, X., Guo, M.: Dkn: Deep knowledge-aware network for news recommendation. In: WWW. pp. 1835–1844 (2018)
16. Wang, J., Jiang, Y., Li, H., Zhao, W.: Improving news recommendation with channel-wise dynamic representations and contrastive user modeling. In: WSDM. p. 562–570 (2023)
17. Wang, R., Wang, S., Lu, W., Peng, X.: News recommendation via multi-interest news sequence modelling. In: ICASSP. pp. 7942–7946 (2022)
18. Wang, S., Guo, S., Wang, L., Liu, T., Xu, H.: Multi-interest extraction joint with contrastive learning for news recommendation. In: ECML PKDD. p. 606–621 (2023)
19. Wu, C., Wu, F., An, M., Huang, J., Huang, Y., Xie, X.: Npa: neural news recommendation with personalized attention. In: KDD. pp. 2576–2584 (2019)
20. Wu, C., Wu, F., Ge, S., Qi, T., Huang, Y., Xie, X.: Neural news recommendation with multi-head self-attention. In: EMNLP-IJCNLP. pp. 6389–6394 (2019)
21. Wu, F., Qiao, Y., Chen, J.H., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Wu, W., Zhou, M.: MIND: A large-scale dataset for news recommendation. In: ACL. pp. 3597–3606 (2020)
22. Zhu, Q., Zhou, X., Song, Z., Tan, J., Guo, L.: Dan: Deep attention neural network for news recommendation. In: AAAI. pp. 5973–5980 (2019)