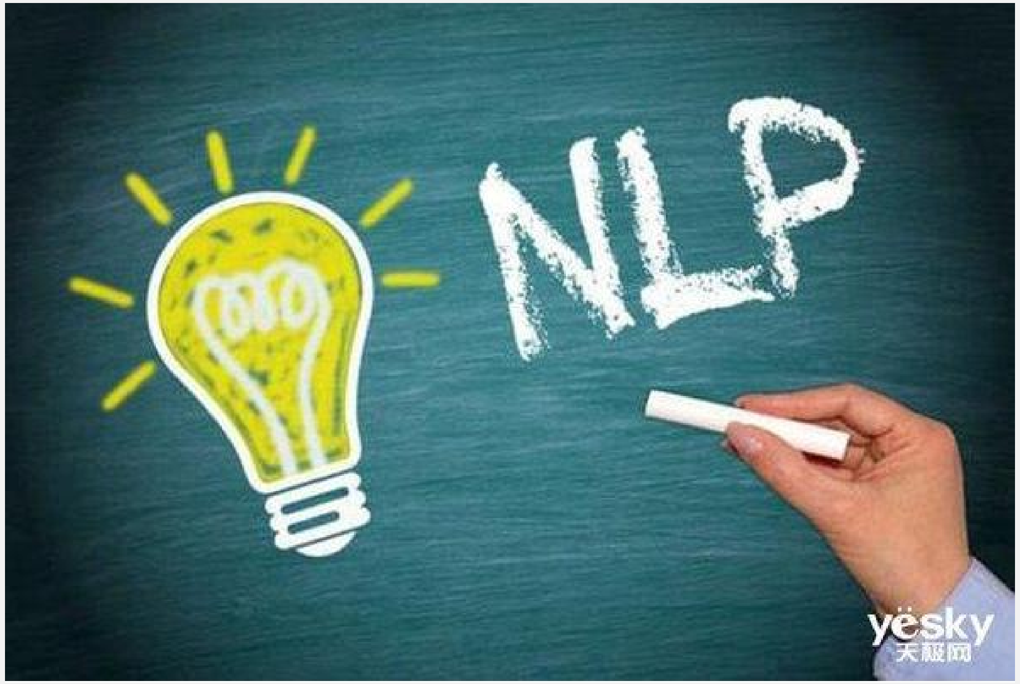
MINER: Multi-Interest Matching Network for News Recommendation

Jian Li[1], Jieming Zhu[1], Qiwei Bi[2], Guohao Cai[1], Lifeng Shang[1], Zhenhua Dong[1], Xin Jiang[1], Qun Liu[1]
[1]Huawei Noah's Ark Lab
[2]School of Statistics, Renmin University of China
{lijian703, jamie.zhu, caiguohao1, shang.lifeng, dongzhenhua}@huawei.com
{jiang.xin, qun.liu}@huawei.com

(ACL-2022)

Reported by Jia Wang



1. Introduction
2. Approach
3. Experiments

Introduction

News Click Sequence

Category	*Title*
Finance	5 excellent dividend stocks to buy for the holiday season.
Sports	Should NFL be able to fine players for criticizing officiating?
Sports	5 takeaways from the 49ers' dominant win over the Panthers.
Movies	Francis Ford Coppola says Marvel movies are 'despicable'.
Sports	Magic vs. Cavs Preview: Magic basketball is finally back.
Fitness	This guy altered his diet and training to drop 65 pounds and pack on muscle.

Figure 1: The news click history of one user, who has various interests including finance, sports and movies.

Introduction

The main contributions of this work can be summarized as follows:

- We propose a poly attention scheme in news recommendation to extract multiple interest vectors for each user. We further improve it with a disagreement regularization to make the extracted vectors more diverse.

- We propose a category-aware attention weighting strategy in the poly attention, which captures explicit category signals for user interest modeling.

- MINER achieves new state-of-the-art on the MIND benchmark and ranked the first on official leaderboard in September 2021.

Approach

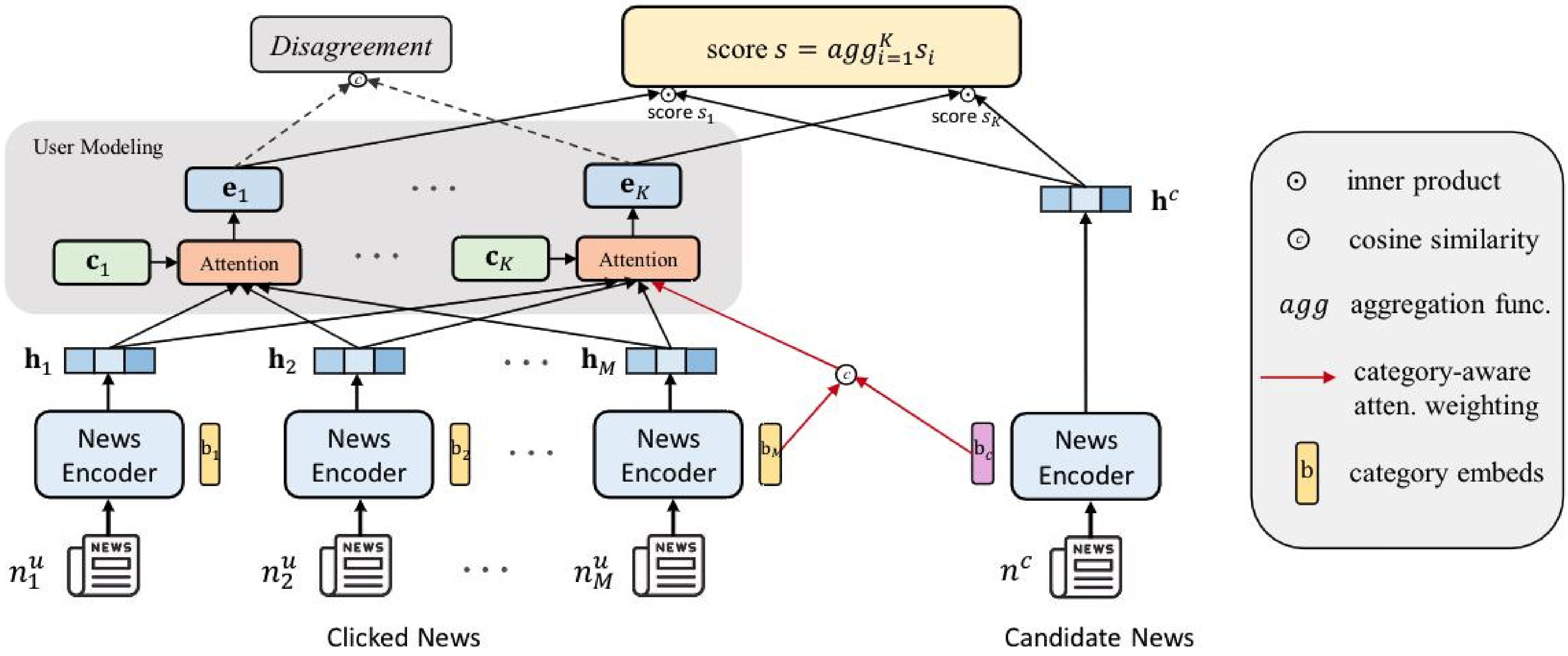
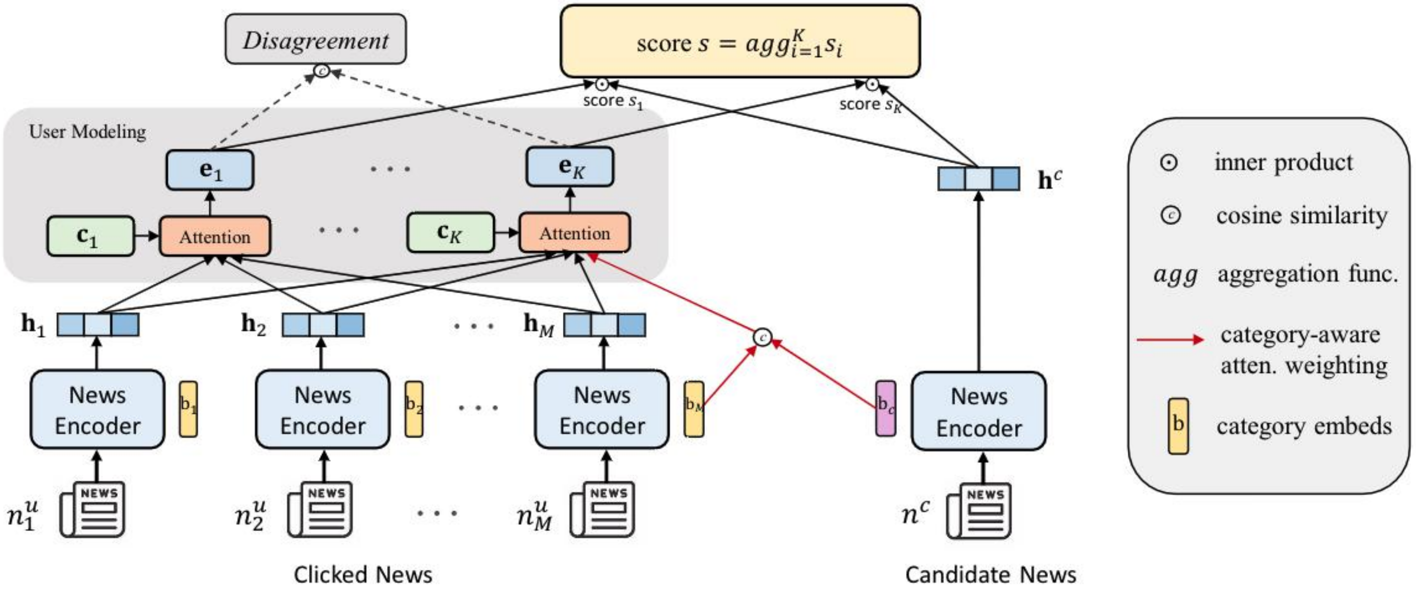


Figure 2: The overall framework of MINER, which consists of a news encoder, a multi-interest user modeling module, and a click score predictor. Disagreement regularization is introduced to make the multiple interest representations more diverse. Category-aware attention weighting is used to re-weight historical news according to the category similarity to candidate news.

Approach

Problem Formulation



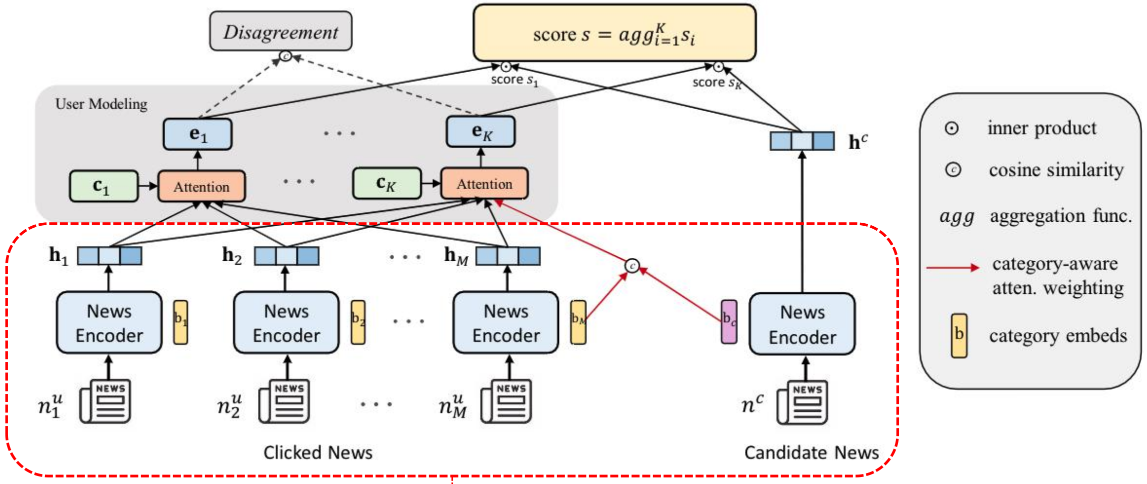
$$N^u = [n_1^u, n_2^u, \ldots, n_M^u],$$

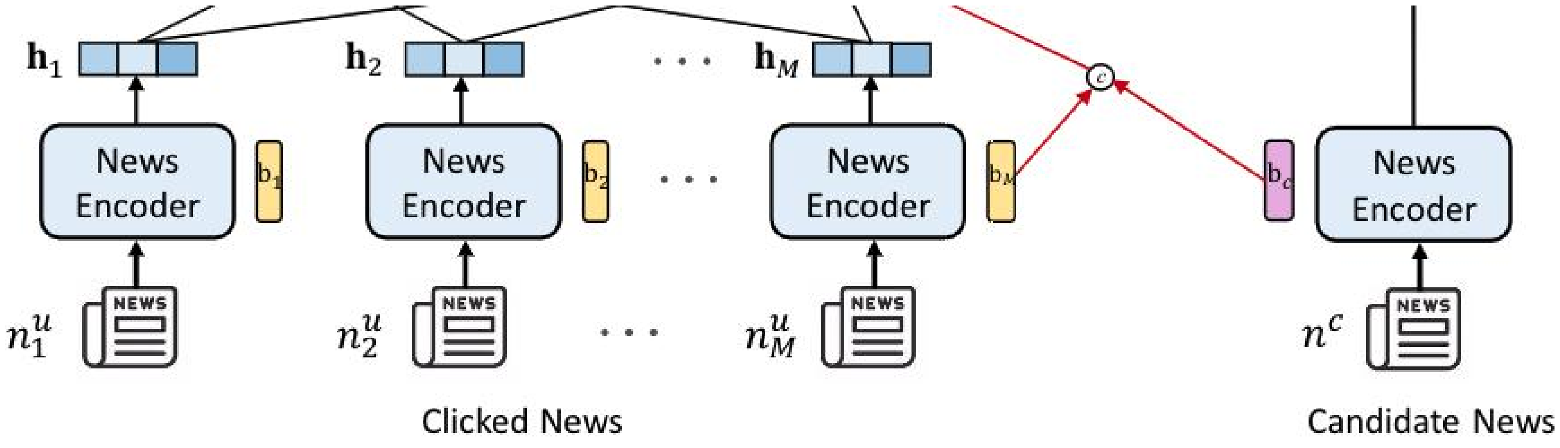
texts T

category ct.

a candidate news n^c

News Encoder





the user u

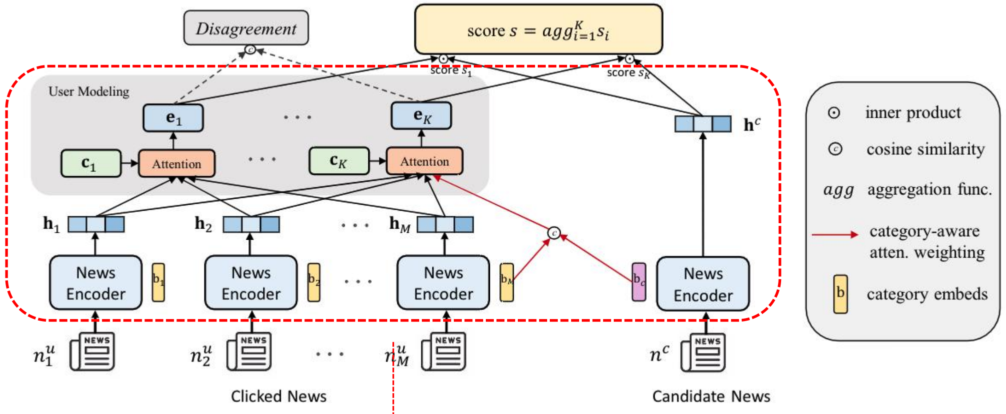
$$\mathbf{H}^u = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_M]$$

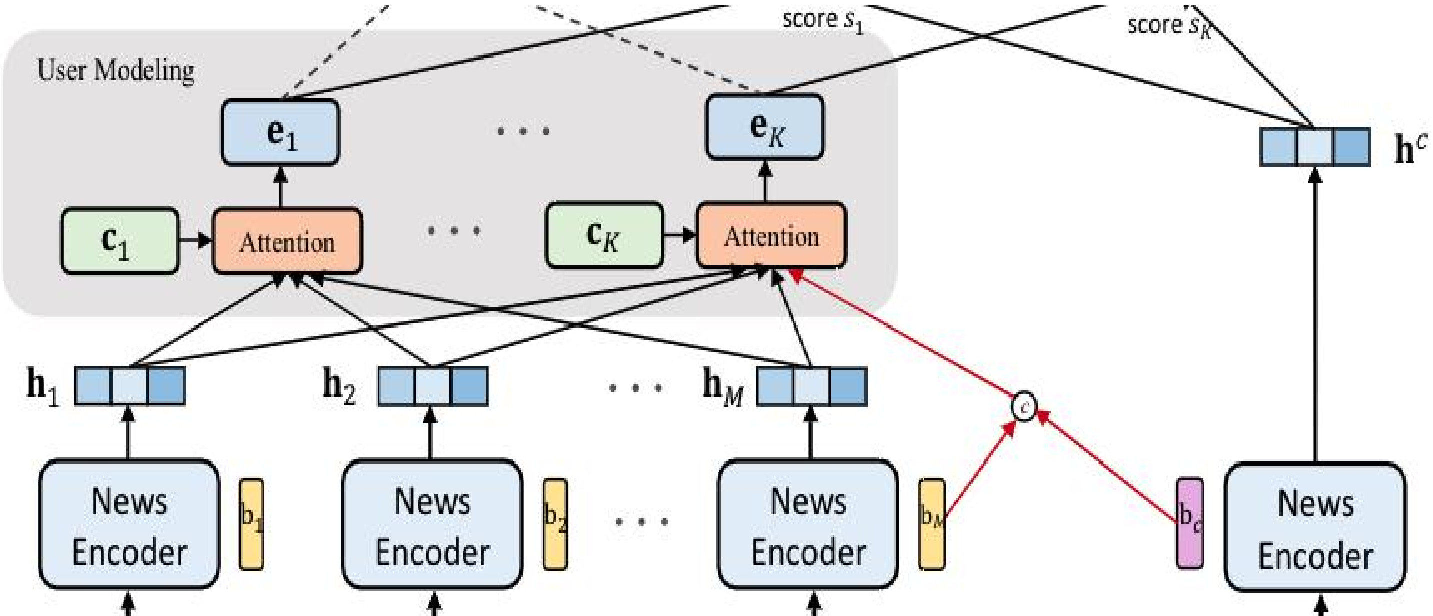
candidate news n^c

$$\mathbf{h}^c$$

Approach

Multi-Interest User Modeling





$$\mathbf{e}_i = \sum_{j=1}^{M} w_j^{c_i} \mathbf{h}_j, \quad w_j^{c_i} = \mathrm{softmax}(\phi_h^{c_i}(\mathbf{h}_j)), \quad (1)$$

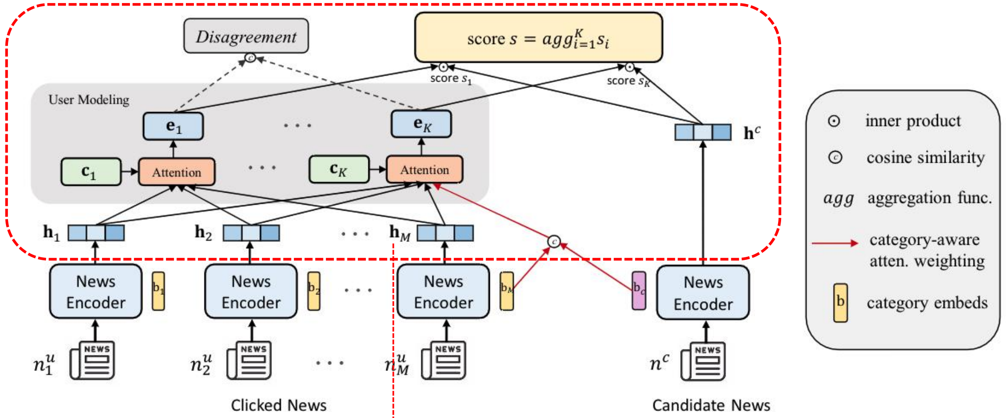
where $w_j^{c_i}$ denotes the attention weight of the j-th historical news. $\phi_h^{c_i}(\cdot)$ is a dense network over the context code $\mathbf{c}_i$ and news representation $\mathbf{h}$:

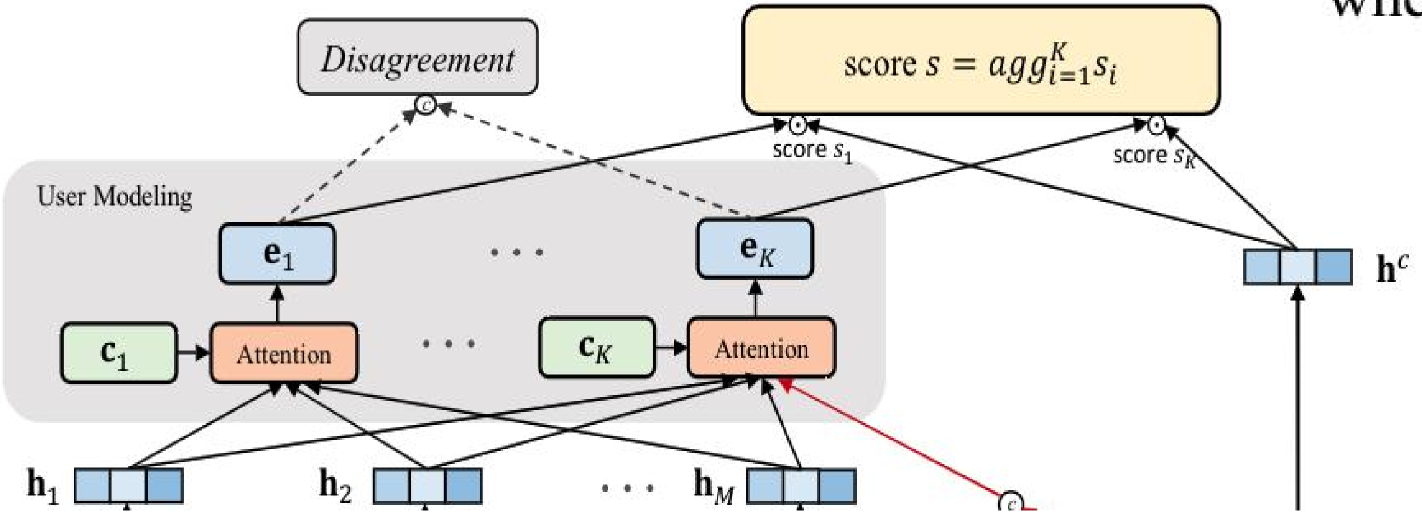
$$\phi_h^{c_i}(\mathbf{h}_j) = \mathbf{c}_i^{\top} \tanh(\mathbf{W}^h \mathbf{h}_j), \qquad (2)$$

where $\mathbf{c}_i$ and $\mathbf{W}^h$ are both trainable parameters.

$$\mathbf{E}^u = [\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_K]$$

Disagreement Regularization

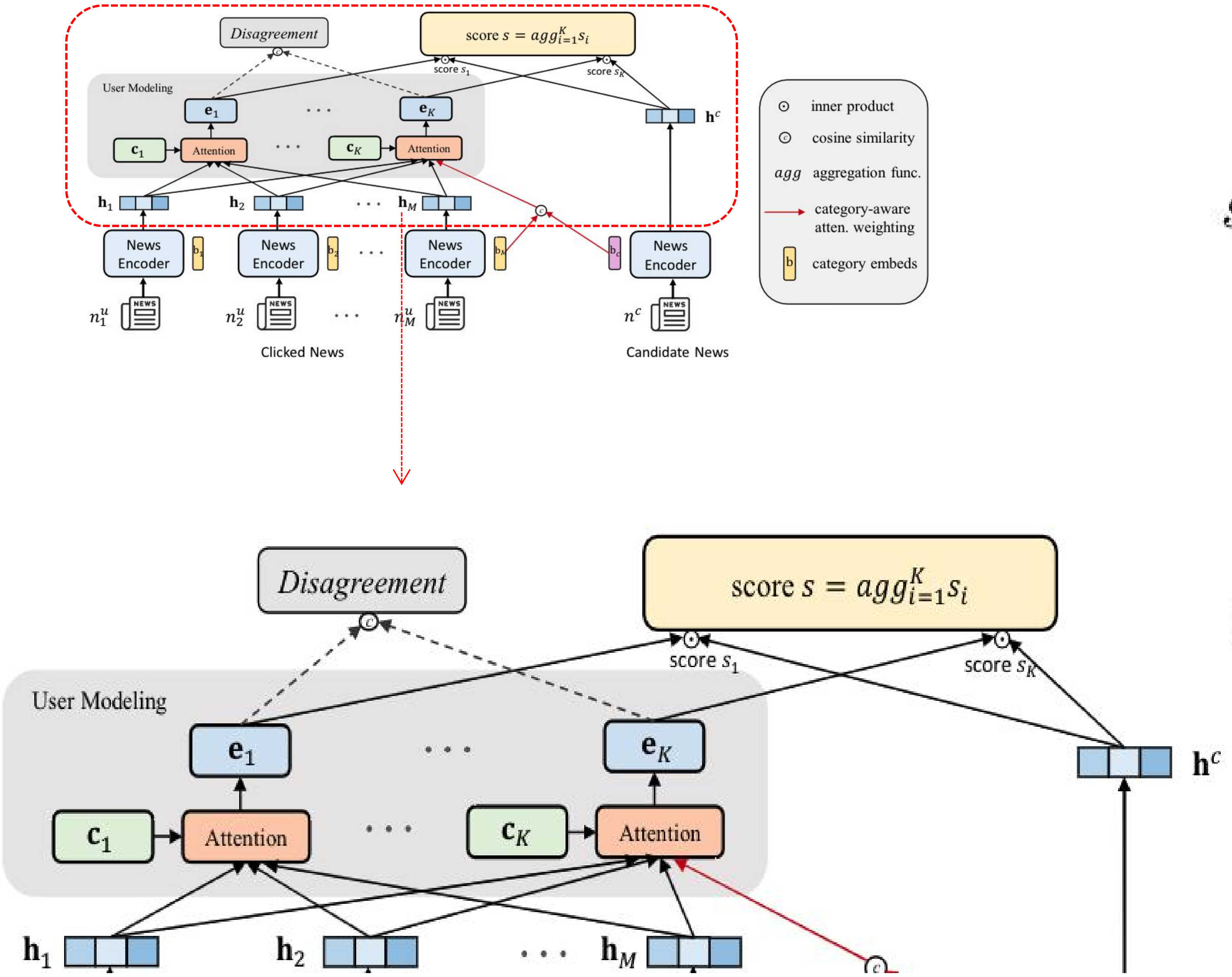




$$\mathcal{L}_D = \frac{1}{K^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{\mathbf{e}_i^\top \mathbf{e}_j}{\|\mathbf{e}_i\| \|\mathbf{e}_j\|}, \qquad (3)$$

where K is the number of interest vectors.

Click Predictor



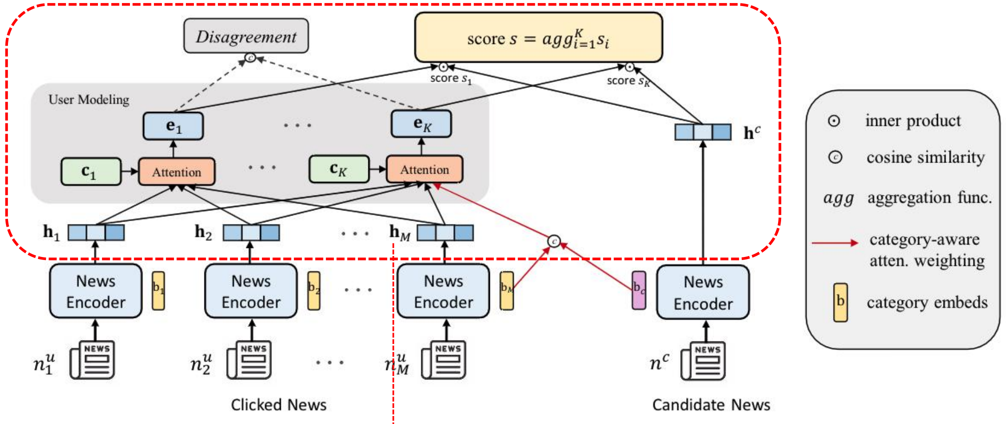
$$s_i = \mathbf{e}_i^\top \mathbf{h}^c. \qquad (4)$$

$$s = \sum_{i=1}^{K} w_i s_i,$$

$$w_i = \mathrm{softmax}(\mathbf{e}_i^\top \mathrm{gelu}(\mathbf{W}^e \mathbf{h}^c)),$$

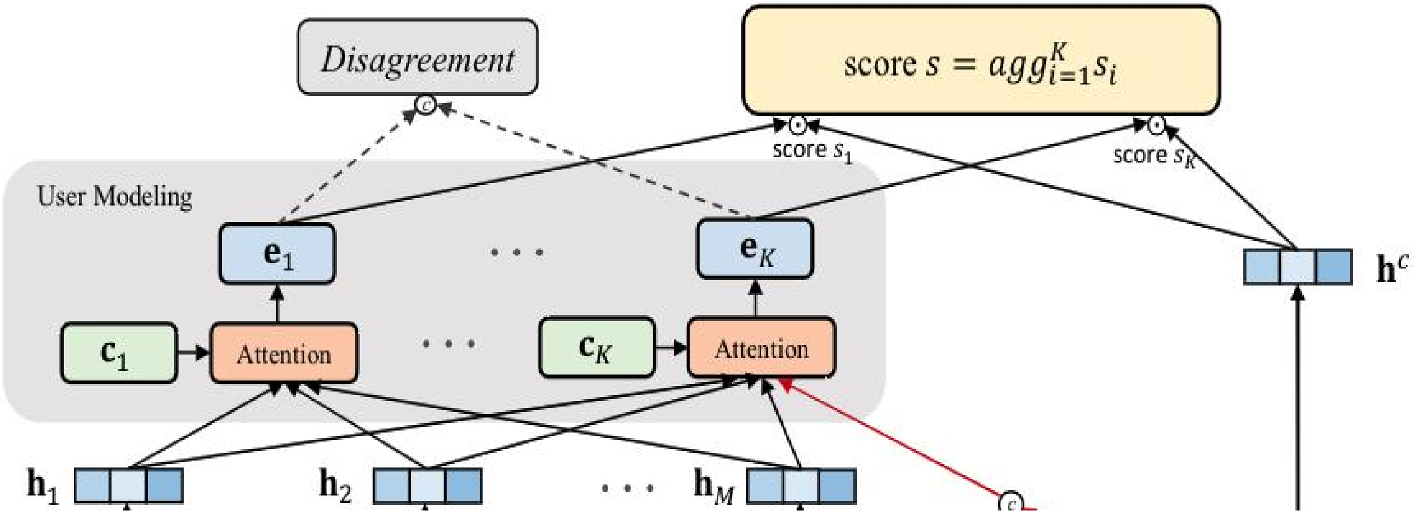
where $\mathrm{gelu}(\cdot)$ is the activation function and $\mathbf{W}^e$ is trainable parameter.

Category-aware Attention Weighting



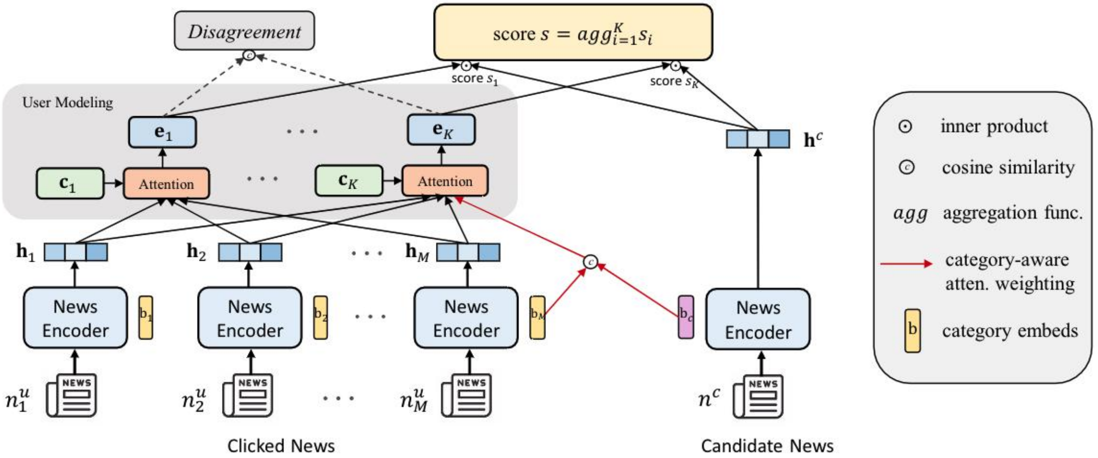
$$\mathbf{e}_i = \sum_{j=1}^{M} w_j^{c_i} \mathbf{h}_j, \quad w_j^{c_i} = \mathrm{softmax}(\phi_h^{c_i}(\mathbf{h}_j)), \quad (1)$$

$$w_j^{c_i} = \mathrm{softmax}(\phi_h^{c_i}(\mathbf{h}_j) + \underline{\lambda \cos(\mathbf{b}_j, \mathbf{b}_c)}), \quad (5)$$



where $\mathbf{b}_j$ and $\mathbf{b}_c$ denote the category embedding of the j-th historical news and the candidate news. $\cos(\cdot)$ denotes the cosine similarity between the two category embeddings and λ is a learnable scalar. Note that, due to the exponential operation in *softmax* function, adding the original logit similarity $\phi_h^{c_i}(\mathbf{h}_j)$ with a bias term $\lambda \cos(\cdot)$ equals to multiplying the attention distribution by a scaling factor. In this way, we learn to re-weight the historical news according to category information.

Model Training



$$\mathcal{L}_{NCE} = -\sum_{i=1}^{|\mathcal{D}|} \log \frac{\exp(s_i^+)}{\exp(s_i^+) + \sum_{j=1}^{L} \exp(s_i^j)}. \tag{6}$$

$$\mathcal{L}_D = \frac{1}{K^2} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{\mathbf{e}_i^\top \mathbf{e}_j}{\|\mathbf{e}_i\| \|\mathbf{e}_j\|}, \tag{3}$$

$$\mathcal{L} = \mathcal{L}_{NCE} + \beta * \mathcal{L}_D, \tag{7}$$

where β is a hyper-parameter and is set to 0.8 based on validation set performance.

Experiments

	MIND-small	MIND-large
# News	65,238	161,013
# Categories	18	20
# Impressions	230,117	15,777,377
# Clicks	347,727	24,155,470

Table 1: Statistics of the two datasets.

Experiments

		MIND-small				MIND-large			
#	Methods	AUC	MRR	nDCG@5	nDCG@10	AUC	MRR	nDCG@5	nDCG@10
1	LibFM	59.74	26.33	27.95	34.29	61.85	29.45	31.45	37.13
2	DeepFM	59.89	26.21	27.74	34.06	61.87	29.30	31.35	37.05
3	DKN	62.90	28.37	30.99	37.41	64.07	30.42	32.92	38.66
4	NPA	64.65	30.01	33.14	39.47	65.92	32.07	34.72	40.37
5	NAML	66.12	31.53	34.88	41.09	66.46	32.75	35.66	41.40
6	LSTUR	65.87	30.78	33.95	40.15	67.08	32.36	35.15	40.93
7	NRMS	65.63	30.96	34.13	40.52	67.66	33.25	36.28	41.98
8	HieRec†	67.95	32.87	36.36	42.53	69.03	33.89	37.08	43.01
9	LSTUR+BERT‡	68.28	32.58	35.99	42.32	69.49	34.72	37.97	43.70
10	NRMS+BERT‡	68.60	32.97	36.55	42.78	69.50	34.75	37.99	43.72
11	UNBERT§	67.62	31.72	34.75	41.02	70.68	35.68	39.13	44.78
12	MINER-*max*	67.39	32.37	35.93	42.11	69.97	35.03	38.37	44.05
13	MINER-*mean*	69.49	33.44	37.37	43.53	71.37	36.06	39.56	45.21
14	MINER-*weighted*	**69.61**	**33.97**	**37.62**	**43.90**	**71.51**	**36.18**	**39.72**	**45.34**

Table 2: Performance of different methods. Previously best results are underlined (the higher, the better) and MINER significantly outperforms all baselines ($p < 0.01$). †: results are from Qi et al. (2021). ‡: results on MIND-large are from Wu et al. (2021). §: results are from Zhang et al. (2021). Our ensemble model on MIND-large ranked the first on official leaderboard: https://msnews.github.io/#leaderboard in September 2021.

Experiments

Model	AUC	MRR	nDCG@10
HieRec (Qi et al., 2021)	67.95	32.87	42.53
MINER w/o BERT	68.07	32.93	42.62
w/o disagreement	67.42	32.38	42.12
w/o category	67.13	32.06	41.73
MINER with BERT	69.61	33.97	43.90
w/o disagreement	69.49	33.46	43.56
w/o category	69.38	33.60	43.60

Table 3: Effects of different MINER components.

Experiments

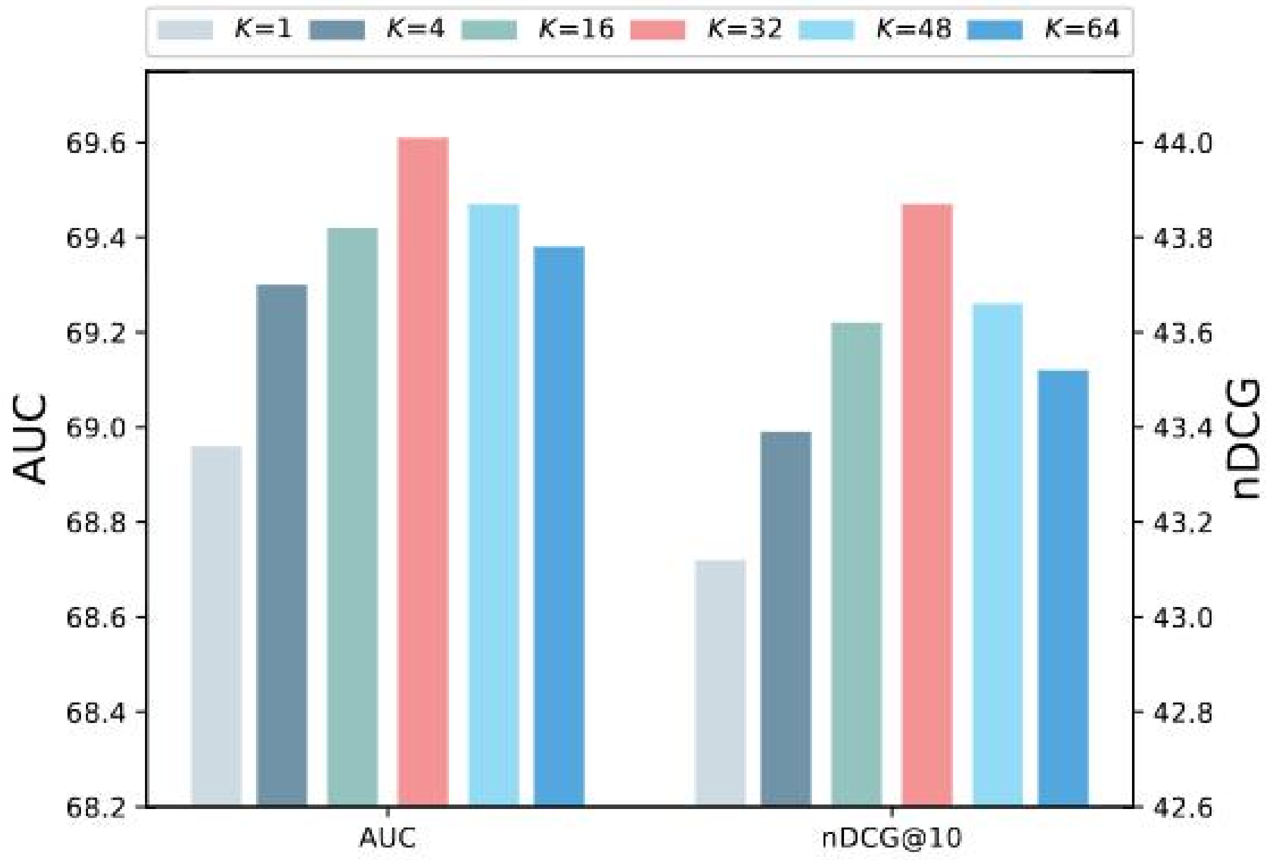


Figure 3: Influence of the number of interest vectors.

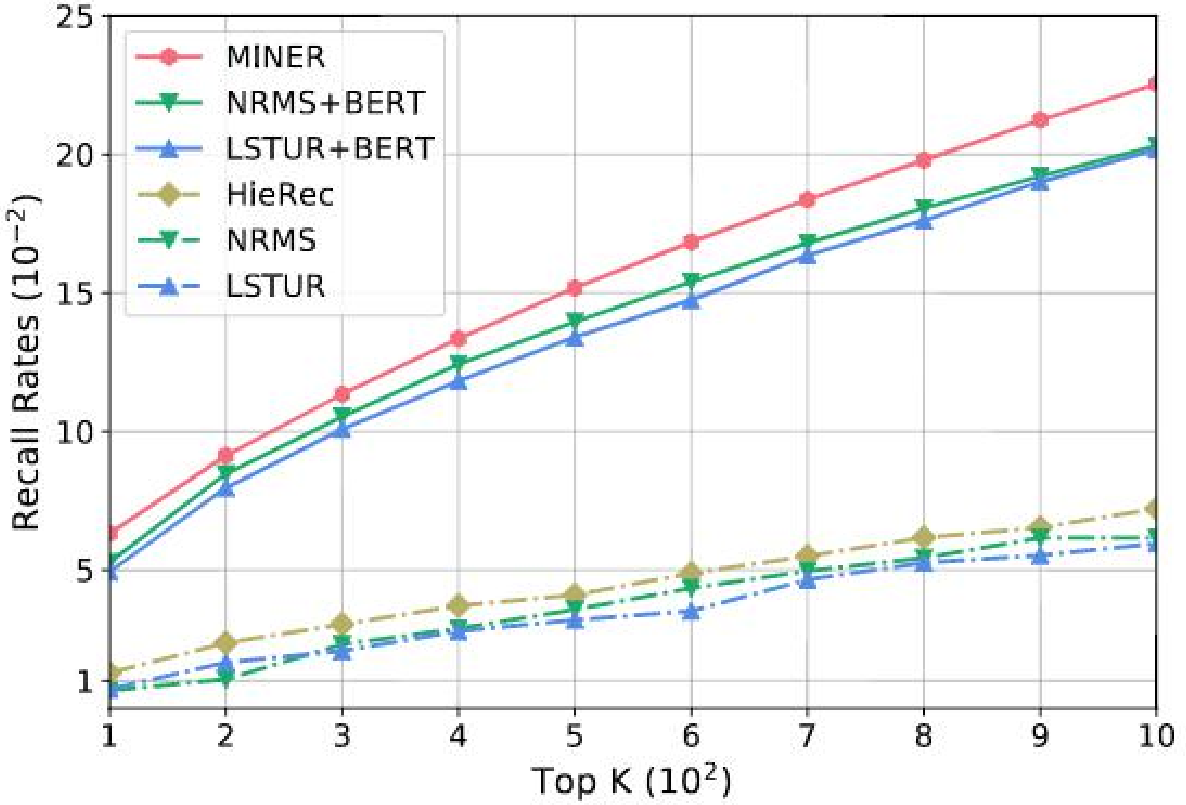


Figure 4: The performance on news recall.

Experiments

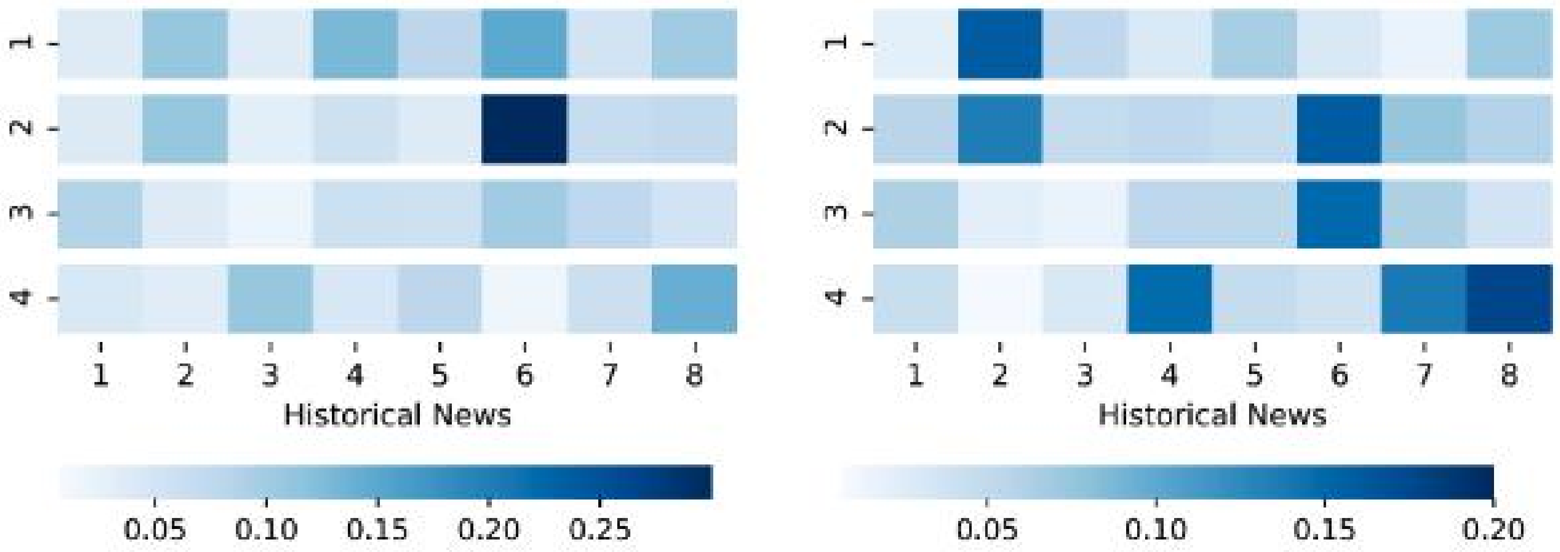
Historical Clicked News		
1	Finance	Man who inherited 6 figures shares advice he'd give his younger self.
2	Sports	Foles will start for Jaguars over Minshew after bye week.
3	Sports	Pete Carroll takes swipe at Patriots over their strict culture.
4	Food	The best Trader Joe's desserts of all time.
5	Politics	Senate to try to override Trump emergency declaration veto Thursday.
6	Sports	NFL had no choice but to send a clear message with Garrett punishment.
7	Sports	Umpire Jeff Nelson leaves game with concussion after being hit by foul balls.
8	Food	Wendy's is turning 50 years old, and is gifting us free food through 2020.

Recommended by NRMS+BERT	
Sports	NFL week 8 power rankings: old-school football rules the day.
Sports	Patriots wanted a test. Now, they need some answers.
Politics	40 conservative groups sign ethics complaint against Pelosi.

Recommended by MINER	
Sports	Patriots wanted a test. Now, they need some answers.
Food	**National Dessert Day: Where to get free dessert at Wendy's.**
Health	Simple diet changes helped this guy lose 75 pounds in 9 months.

Figure 5: Case study on top 3 news recommended by *NRMS+BERT* and MINER in a sampled impression. The news actually clicked by the user is highlighted in blue.

Experiments



(a) w/o disagreement (b) with disagreement

Figure 6: Visualize the attention weights on the historical news in Figure 5 (a) before and (b) after applying disagreement regularization.

Thanks！