Compact Interpretable Tensor Graph Multi-Modal News Embeddings

Dawon Ahn
dahn017@ucr.edu
University of California, Riverside

William Shiao
wshia002@ucr.edu
University of California, Riverside

Arindam Khaled
akhaled@seekr.com
Seekr

Andrew Bauer
abauer@seekr.com
Seekr

Stefanos Poulis
spoulis@seekr.com
Seekr

Evangelos E. Papalexakis
epapalex@cs.ucr.edu
University of California, Riverside

ABSTRACT

Online news articles encompass a variety of modalities such as text and images. How can we learn a representation that incorporates information from all those modalities in a compact and interpretable manner? In this paper, we propose CITEM (**C**ompact **I**nterpretable **T**ensor graph multi-modal news **EM**bedding), a tensor based framework for compact and interpretable multi-modal news representations. CITEM generates a tensor graph consisting of a news similarity graph for each modality and employs a tensor decomposition to produce compact and interpretable embeddings, each dimension of which is a heterogeneous co-cluster of news articles and corresponding modalities. We extensively validate CITEM compared to baselines on two news classification tasks: misinformation news detection and news categorization. The experimental results show that CITEM performs within the same range of AUC as state-of-the-art baselines while producing 7× to 10.5× more compact embeddings. In addition, each embedding dimension of CITEM is interpretable, representing a latent co-cluster of articles.

CCS CONCEPTS

• **Information systems → Document representation**.

KEYWORDS

Tensor decomposition; Interpretable multi-modal embeddings

ACM Reference Format:
Dawon Ahn, William Shiao, Arindam Khaled, Andrew Bauer, Stefanos Poulis, and Evangelos E. Papalexakis. 2024. Compact Interpretable Tensor Graph Multi-Modal News Embeddings. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3651480

1 INTRODUCTION

Online news articles contain a variety of modalities such as text, image, and video, that are useful for representing the core content of the news. News representations aim to understand the main contents of news expressed as a real-valued vector and have played a fundamental role in solving a variety of news tasks such as fake/clickbait news classification [1, 3, 20] and news recommendation [11, 19, 20]. With advancements in natural language processing (NLP) techniques, most existing approaches have focused on accurately understanding textual information from news titles and bodies that contain concise and detailed information of key content [11, 19]. Recently, leveraging multiple modalities to represent news has been actively studied with the success of multi-modal learning such as CLIP [14]. Many studies have developed multimodal news representations with various types of information such as category, images, and knowledge graph, in addition to text, to enhance news representations [20].

Albeit very powerful and well-performing in a variety of downstream tasks, those state-of-the-art representations are usually high-dimensional with each dimension being disconnected from any semantically meaningful information that can help a practitioner understand, for instance, what features contribute to classifying a particular news article as "clickbait".

In this work, we propose CITEM (**C**ompact **I**nterpretable **T**ensor graph multi-modal news **EM**bedding), a tensor-based ensemble news representation that bridges the above gap by computing compact embeddings for news articles which effectively combine the information contained in individual modalities, while maintaining clustering-based interpretations for each of the new embedding dimensions, allowing for feature inspection and analysis in downstream tasks (as shown in Fig. 1). The source code and datasets are available at https://github.com/dawonahn/CITEM.

2 PRELIMINARIES & RELATED WORK

We introduce preliminaries including tensor decomposition and review literatures related to news embedding models.

Tensors are defined as multi-dimensional arrays that generalize one-dimensional arrays (or vectors) and two-dimensional arrays (or matrices) to higher dimensions. The dimension of a tensor is referred to as its order or mode; the length of each mode is called "dimensionality". We use boldface Euler script letters (e.g., $\boldsymbol{\mathcal{X}}$) to denote tensors, boldface capitals (e.g., $\mathbf{A}$) to denote matrices, and boldface lower cases (e.g., $\mathbf{a}$) to denote vectors. We denote the i-th row vector as $\mathbf{a}_{i,:}$ and i-th column vector as $\mathbf{a}_i$.

Tensor Decomposition is a popular tensor mining tool to discover underlying low-dimensional patterns in the tensor. We focus on CANDECOMP/PARAFAC (CP) decomposition model [2], one of the most famous models, which decomposes a tensor into a sum of rank-one components [7]. We choose the CP decomposition because of its interpretability and simplicity, which makes it easy to

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05.
https://doi.org/10.1145/3589335.3651480

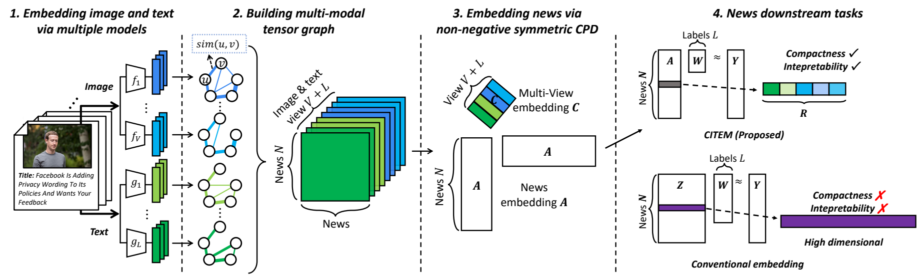


Figure 1: Illustration of main ideas of CITEM. We extract multiple textual and visual features from news via various pre-trained models. We convert each feature space into a graph and stack them into a tensor. With non-negative symmetric CPD, we decompose the given tensor to obtain compact, interpretable embeddings where each dimension is interpretable with regard to each modality and news relations.

Table 1: AUC comparison of utilizing multiple pre-trained models. To represent each modality, using multiple pre-trained models are more effective than using a single model.

Dataset	LLM	LVM	AUC
Seekr	CLIP	CLIP	0.730
	CLIP, SBERT	CLIP, ResNet	0.732
	CLIP, SBERT, BART	CLIP, ResNet, ViT	**0.772**
News Category	CLIP	CLIP	0.871
	CLIP, SBERT	CLIP, ResNet	0.955
	CLIP, SBERT, BART	CLIP, ResNet, ViT	**0.968**

analyze the low-dimensional patterns. Given a third-order tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$ and a rank R, CP decomposition approximates $\mathcal{X}$ to find factor matrices $\{\mathrm{A} \in \mathbb{R}^{I \times R}, \mathrm{B} \in \mathbb{R}^{J \times R}, \mathrm{C} \in \mathbb{R}^{K \times R}\}$ that minimize: $\min_{\mathrm{A,B,C}} ||\mathcal{X} - \tilde{\mathcal{X}}||$ where $\tilde{\mathcal{X}} = [\![\mathrm{A}, \mathrm{B}, \mathrm{C}]\!] = \sum_{r=1}^{R} \mathrm{a}_r \circ \mathrm{b}_r \circ \mathrm{c}_r$. Note that $\circ$ represents an outer product and $\mathrm{a}_r \in \mathbb{R}^I$ is a rth column factor of A (similarly for b_r and c_r). Each row of factor matrices indicates a representation of an entity of each mode. For example, $\mathrm{a}_{i,:}$ is a representation of i-th element of the first mode. The rth rank-one component (e.g., $\mathrm{a}_r \circ \mathrm{b}_r \circ \mathrm{c}_r$) corresponds to the rth latent co-cluster which groups modes that share similar relationships with each other. With this interpretable property of CP decomposition, we are able to describe each dimension of the representation by identifying latent clusters while other news embedding models are not interpretable due to their high dimensionality and semantically irrelevant dimensions.

3 PROPOSED METHOD

We propose CITEM (**C**ompact **I**nterpretable **T**ensor graph multi-modal news **EM**bedding) which 1) expresses multi-modal information of news, 2) explains each dimension of the news embedding with latent clusters of news, as illustrated in Fig. 1.

We describe how we fuse embeddings—each corresponding to different news modalities from multiple pre-trained models—into a single compact and interpretable embedding. We first extract the title and the top image from each of the N news articles. We then compute multiple text (title) and image (top image) embeddings from each article using different types of language and vision models, which allow for more accurate and discriminative news article representations, as shown in Table 1. More specifically, each news article can be represented as text feature vectors $\mathrm{U}^{(\ell)} = \{\mathrm{u}_n^{(\ell)}\}_{n=1}^{N}$ produced by the ℓth language model for $1 \leq \ell \leq L$ and image feature vectors $\mathrm{V}^{(v)} = \{\mathrm{v}_n^{(v)}\}_{n=1}^{N}$ produced by the vth vision model for $1 \leq v \leq V$. However, these text and image feature vectors are embedded in different feature spaces since different pre-trained models have been trained with different data and learning methods and may have different embedding sizes.

This raises a question: *how can we represent each representation in the same space without losing the rich information obtained from pre-trained models?* The common intermediate representation we choose is a similarity *graph*, where nodes represent news articles and edge weights are similarities between the news representations produced by each embedding model. We then calculate the pairwise cosine similarity between each normalized embedding vector to produce similarity graphs $\mathrm{X}_u^{(\ell)} = \mathrm{U}^{(\ell)}\mathrm{U}^{(\ell)\intercal}$ and $\mathrm{X}_v^{(v)} = \mathrm{V}^{(v)}\mathrm{V}^{(v)\intercal}$. Next, we stack all graphs and build a multi-modal tensor graph $\mathcal{X} \in \mathbb{R}^{N \times N \times K}$ where $K = L+V$. Extensive prior work [12] has demonstrated that tensor analysis in such multi-graphs, where there is an expectation of overlapping (but not entirely identical) structure across different graphs, can yield expressive representations where each latent factor corresponds to a co-cluster of news articles and different modalities [13].

The multi-modal tensor graph we form is symmetric with respect to the first and the second modes and is non-negative. As such, we impose two constraints on the CP decomposition: 1) symmetry—the first and the second factor matrix are identical, and 2) non-negativity—all factor matrices are non-negative. The above constraints result in a non-negative symmetric CP decomposition (NS-CPD), which allows us to generate compact and interpretable multi-modal news embeddings given a multi-modal tensor graph. Given a third-order multi-modal tensor graph $\mathcal{X} \in \mathbb{R}^{N \times N \times K}$ and a rank R, the NS-CPD approximates $\mathcal{X}$ to find factor matrices $\{\mathrm{A} \in \mathbb{R}_+^{N \times R}, \mathrm{C} \in \mathbb{R}_+^{K \times R}\}$ that solves:

$$\min_{\mathrm{A,C}} ||\mathcal{X} - \tilde{\mathcal{X}}|| \text{ where } \tilde{\mathcal{X}} = [\![\mathrm{A}, \mathrm{A}, \mathrm{C}]\!] = \sum_{r=1}^{R} \mathrm{a}_r \circ \mathrm{a}_r \circ \mathrm{c}_r. \quad (1)$$

The nth row vector of factor matrix A corresponds to the nth news embedding whose dimensions are co-clustered with other news entities. The kth row vector of factor matrix C corresponds to a representation of the kth pre-trained model, which indicates its influence on each dimension of the news embedding. This allows us to identify which modality or model significantly influences a dimension of news embedding. We use the Adam [6] optimizer to train factor matrices.

4 EXPERIMENTS

We perform experiments to answer the following questions: **Q1**. How accurately does CITEM perform in news downstream tasks? (Sec. 4.2), **Q2**. How compact are CITEM embeddings? (Sec. 4.3), and **Q3**. Can CITEM produce interpretable results? (Sec. 4.4)

4.1 Experimental Settings

We evaluate CITEM on two news downstream tasks: misinformation news detection and news categorization.

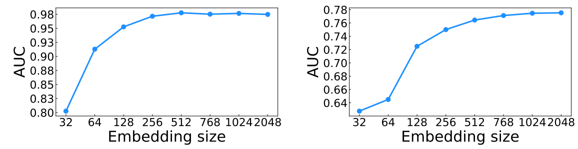
Dataset. We evaluate the performance of CITEM and baselines on four real-world datasets. All datasets except for News Category are related to misinformation detection tasks. We construct a multi-modal tensor graph with six pre-trained models to extract text and image feature vectors. The first and the second modes of the tensor represent news entities while the last mode represents modalities. GossipCop[1] [16–18] is a fake news benchmark dataset about celebrity gossip including 16,817 real and 5,323 fake news articles. We sample 30 percent of the real articles (5,045 articles) to create a balanced dataset. We construct the tensor of size $10,368 \times 10,368 \times 6$. PolitiFact[1] [16–18] is a fake news benchmark dataset about politicians' statements including 625 real and 432 fake news articles. We construct the tensor of size $1,056 \times 1,056 \times 6$. Seekr[2] is a click-bait news dataset with $1,148$ click-baits and $4,563$ non-click baits collected from Seekr—a news aggregator that rates the credibility of articles. We construct the tensor graph of size $5,711 \times 5,711 \times 6$. News Category[3] [9] is a HuffPost news dataset from 2010 to 2022, containing 38 different topics. We sample datasets with 15 categories for labels and sample 10 percent of the original datasets. We construct the tensor graph of size $9,976 \times 9,976 \times 3$.

Pre-trained models. We utilize various pre-trained language and vision models to extract text and image features from news articles. We use Hugging Face's[5] implementations. We employ two language models, SBERT [15] and BART [8], two vision models, ResNet [5] and ViT [4], and a multi-modal model, CLIP [14].

Baselines. We describe baselines to evaluate our proposed method. Concat-Text is 2,048-dim. text embeddings from SBERT, BART, and CLIP models concatenated together. Concat-Img is 3,328-dim. image embeddings from ViT, ResNet, and CLIP models concatenated together. Concat-Both is 5,376-dim. text and image embeddings from all six models concatenated together. Concat-PCA is 768-dim. text and image embeddings obtained from applying Principal Component Analysis (PCA) on Concat-Both. CPD is CP decomposition with Alternating Least Square (ALS) optimization. RESCAL [10] is a symmetric Tucker with L2 regularization optimized via Adam.

Hyper-parameters. We vary the embedding size (rank), ranging from 32 to 2024. We fix a learning rate of 0.001 and a weight decay of 0.001. We employ a linear classifier as the downstream classifier for downstream tasks. This choice is motivated by two reasons: 1) we focus on news embedding models rather than complex nonlinear classifiers, and 2) we can easily identify the influential dimensions of embeddings through the weight of linear classifiers.

[1] https://github.com/KaiDMML/FakeNewsNet
[2] https://www.seekr.com/
[3] https://www.kaggle.com/datasets/rmisra/news-category-dataset
[5] https://huggingface.co



(a) Seekr (b) News Category

Figure 2: The AUC of **CITEM** according to the different embedding sizes (rank). **CITEM** performs well at embedding sizes 256 and 768.

4.2 Classification Performance

We show performance comparisons in Table 2. We can see that CITEM performs on par with the best-performing methods. We must note here that our goal is not necessarily to beat the best-performing method but to perform comparably to it, since this indicates that CITEM is able to successfully distill the multi-modal information in a *compact* and *effective* manner while allowing for intuitive interpretations of the newly computed embedding dimensions. Concat-Both demonstrates that incorporating multi-modal information is crucial for accurately representing news, rather than relying solely on uni-modal information. The embeddings generated by CPD and RESCAL are smaller in size while showing good performance but are not as interpretable as the proposed method because the embeddings are non-negative.

4.3 Compactness

Fig. 2 demonstrates the compactness of CITEM for news embedding. We investigate the compactness of the proposed method while adjusting embedding sizes (rank) ranging from 32 to 2024 as shown in Fig. 2. The AUC of CITEM increases as the rank size increases. For the PolitiFact dataset, CITEM achieves the highest AUC of baselines with 256 embedding size, which is 21× smaller than Concat-Both.

4.4 Interpretability

We examine each dimension of news embedding to analyze the interpretability of CITEM. We discover important dimensions of news embedding based on intercepts and coefficients of a logistic regression model. After training the logistic regression model, we compute an influence score $|\mathrm{a}_{n,:}\mathrm{w}_n^l + d^l|$ with the nth news embedding $\mathrm{a}_{n,:}$ and corresponding coefficients w^l of a label l and an intercept d^l. With the highest influence scores, we find the top-k influential dimensions of nth embeddings from model decisions. We then identify which dimensions correspond to which modalities based on multi-view embedding C. Fig. 3 illustrates interpretation of dimensions from CITEM on News Category dataset. Given a news article, we select the top-5 influential dimensions and display each of their representative articles.

Table 2: Performance comparison on news downstream tasks. Note that the best method is in bold, and the second-best method is underlined. **CITEM** integrates different embeddings accurately and is interpretable due to its non-negativity.

Model	GossipCop				PolitiFact				Seekr				News Category			
	Size	Acc.	F1	AUC	Size	Acc.	F1	AUC	Size	Acc.	F1	AUC	Size	Acc.	F1	AUC
Concat-Text	2,048	0.811	0.779	0.886	2,048	0.915	0.894	0.965	2,048	0.818	0.378	0.791	2,048	0.741	0.741	0.969
Concat-Img	3,328	0.740	0.667	0.829	3,328	0.736	0.548	0.786	3,328	0.799	0.202	0.660	3,328	0.615	0.615	0.930
Concat-Both	5,376	<u>0.854</u>	0.828	**0.922**	5,376	0.934	0.916	0.973	5,376	0.822	0.384	0.776	5,376	**0.769**	**0.769**	<u>0.972</u>
Concat-PCA	768	0.853	0.826	<u>0.918</u>	768	0.934	0.916	0.973	768	0.815	0.385	0.766	768	<u>0.767</u>	<u>0.767</u>	<u>0.972</u>
CPD	512	0.845	0.820	0.905	64	<u>0.943</u>	<u>0.930</u>	0.971	256	<u>0.827</u>	**0.453**	0.767	256	0.741	0.741	0.956
RESCAL	512	0.830	0.797	0.898	256	0.934	0.916	<u>0.976</u>	512	0.822	0.342	**0.792**	512	0.756	0.756	**0.973**
CITEM (Proposed)	768	0.844	0.814	0.903	256	**0.953**	**0.941**	**0.977**	768	**0.831**	<u>0.425</u>	<u>0.772</u>	768	0.752	0.752	0.968

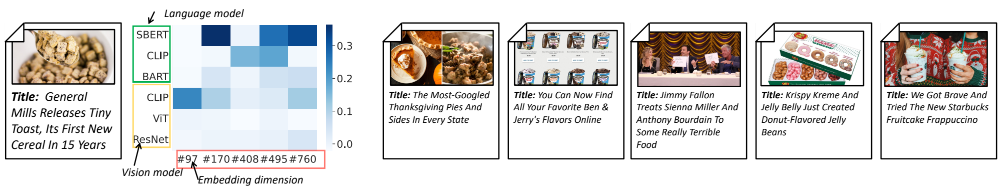


Figure 3: Thanks to interpretable embeddings and multi-view embedding, we can identify the top-5 dimensions and the most influential multi-view for news category classification. This information allows us to determine which dimensions have a significant impact on the model's decisions.

5 CONCLUSION

We propose CITEM, a tensor-based framework for interpretable multi-modal news representation. With non-negative symmetric CPD, CITEM successfully integrates multi-modal information extracted from pre-trained models into compact embeddings that are interpretable with regard to related articles based on their modalities. We further develop the framework with end-to-end training, which can be subsequently applied to specific downstream tasks.

6 ACKNOWLEDGEMENT

This research was supported by the National Science Foundation under CAREER grant no. IIS 2046086 and CREST Center for Multidisciplinary Research Excellence in Cyber-Physical Infrastructure Systems (MECIS) grant no. 2112650, a CISCO Research Faculty Award, and by the Combat Capabilities Development Command Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-13-2-0045 (ARL Cyber Security CRA). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Combat Capabilities Development Command Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes not withstanding any copyright notation here on.

REFERENCES

[1] Sara Abdali, Gisel G Bastidas, Neil Shah, and Evangelos E Papalexakis. 2020. Tensor Embeddings for Content-Based Misinformation Detection with Limited Supervision. In *Disinformation, Misinformation, and Fake News in Social Media*. Springer, 117–140.

[2] J Douglas Carroll and Jih-Jie Chang. 1970. Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika* 35, 3 (1970), 283–319.

[3] Limeng Cui, Suhang Wang, and Dongwon Lee. 2019. Same: sentiment-aware multi-modal embedding for detecting fake news. In *Proceedings of the 2019 IEEE/ACM international conference on advances in social networks analysis and mining*. 41–48.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[6] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[7] Tamara G Kolda and Brett W Bader. 2009. Tensor decompositions and applications. *SIAM review* 51, 3 (2009), 455–500.

[8] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[9] Rishabh Misra. 2018. *News Category Dataset.* https://doi.org/10.13140/RG.2.2.20331.18729

[10] Maximilian Nickel, Volker Tresp, Hans-Peter Kriegel, et al. 2011. A three-way model for collective learning on multi-relational data.. In *Icml*, Vol. 11. 3104482–3104584.

[11] Shumpei Okura, Yukihiro Tagami, Shingo Ono, and Akira Tajima. 2017. Embedding-based news recommendation for millions of users. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1933–1942.

[12] Evangelos E Papalexakis, Leman Akoglu, and Dino Ience. 2013. Do more views of a graph help? community detection and clustering in multi-graphs. In *Proceedings of the 16th International Conference on Information Fusion*. IEEE, 899–905.

[13] Evangelos E Papalexakis, Nicholas D Sidiropoulos, and Rasmus Bro. 2012. From k-means to higher-way co-clustering: Multilinear decomposition with sparse latent factors. *IEEE transactions on signal processing* 61, 2 (2012), 493–506.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[15] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[16] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. FakeNewsNet: A Data Repository with News Content, Social Context and Dynamic Information for Studying Fake News on Social Media. *arXiv preprint arXiv:1809.01286* (2018).

[17] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations Newsletter* 19, 1 (2017), 22–36.

[18] Kai Shu, Suhang Wang, and Huan Liu. 2017. Exploiting Tri-Relationship for Fake News Detection. *arXiv preprint arXiv:1712.07709* (2017).

[19] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with attentive multi-view learning. *arXiv preprint arXiv:1907.05576* (2019).

[20] Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2022. Personalized News Recommendation: Methods and Challenges. *ACM Transactions on Information Systems (TOIS)* (2022).