# Compact Interpretable Tensor Graph Multi-Modal News Embeddings

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

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

Andrew Bauer
abauer@seekr.com
Seekr Technologies
San Diego, CA, USA

Arindam Khaled
akhaled@seekr.com
Seekr Technologies
San Diego, CA, USA

Stefanos Poulis
spoulis@seekr.com
Seekr Technologies
San Diego, CA, USA

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

## 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, while also being useful in a variety of downstream tasks? Recent advances in Large Language and Vision Models have made it possible to represent image and text data as embeddings, which can then be used to perform downstream tasks. Despite these developments, these embedding models tend to generate high-dimensional embeddings, making them problematic in terms of compactness and interpretability. In this paper, we propose CITEM (**C**ompact **I**nterpretable **T**ensor graph multi-modal news **EM**bedding), a tensor decomposition 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. Traditional tensor analysis has so far been restricted to transductive learning scenarios (e.g., in the form of semi-supervised learning), but CITEM includes two variants for inductive learning, which essentially allows us to represent unseen news articles. 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 baselines while producing 7× more compact embeddings.

## KEYWORDS

Tensor decomposition, Multi-modal tensor graph, Interpretable news embeddings

**ACM Reference Format:**
Dawon Ahn, William Shiao, Andrew Bauer, Arindam Khaled, Stefanos Poulis, and Evangelos Papalexakis. 2023. Compact Interpretable Tensor Graph Multi-Modal News Embeddings. In *The 29th ACM SIGKDD Conference on Knowledge Discovery & Data Mining Information and Knowledge Management (KDD2023) August 06–10, 2023.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456

## 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. Given news with those modalities, how can we effectively exploit them to learn concise and interpretable news representations? 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/click-bait news classification [1, 6, 31, 40], news recommendation [23, 39, 40], news topic detection [5], and stock trend prediction [13].

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 [23, 39]. Language pre-trained models such as BERT [7] trained with a massive corpus have greatly improved the performance of news embedding. Although many methods with those techniques accurately represent textual information from news, it is still insufficient to make an accurate news representation since news includes various information and different characteristics. Recently, leveraging multiple modalities to represent news has been actively studied with the success of multi-modal learning such as CLIP [28]. Many studies have developed multimodal news representations with various types of information such as category, images, knowledge graph, news relation graph, etc, in addition to text, which effectively enhances news representations [40].

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”.

*KDD'23, August 06–10, 2023, Long Beach, CA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/1122445.1122456

In this work, we propose CITEM (**C**ompact **I**nterpretable **T**ensor graph multi-modal news **EM**bedding), a tensor-based news representation learning framework that effectively combines the information contained in individual modalities employing pre-trained language and vision models, 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).

Traditionally, tensor-based methods which lend themselves to interpretability are restricted in the transductive learning setting [9, 10, 44], where representations for labeled and unlabeled instances are computed jointly (which is a typical scenario in Graph Neural Network evaluation as well [45]). In our work, however, we are the first to extend such interpretable tensor decomposition methods to the inductive setting where we can use our embeddings to classify unseen news articles. Our contributions are summarized as follows:

- **Compact & Interpretable News Embedding.** We propose a compact and interpretable news embedding model encoding information from multiple modalities and news relations.
- **Inductive Learning Algorithms.** We propose two variants of inductive learning methods to apply the proposed embedding method to consider unseen news.
- **Experiments.** Extensive experiments on five real-world datasets demonstrate that CITEM shows similar performances compared to baselines with 29.8x smaller embedding sizes.

The rest of this paper is organized as follows. In Sec. 2, we introduce the preliminaries of CITEM. In Sec. 3, we propose CITEM. We present experimental results in Sec. 4 and describe related works in Sec. 5. We summarize the key points and results of our paper in Sec. 6. The source code and datasets used in this paper are available at https://anonymous.4open.science/r/CITEM.

## 2 PRELIMINARIES

We introduce preliminaries including tensor decomposition and transductive and inductive learning.

### 2.1 Tensor Methods

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 methods are a popular tensor mining tool to discover underlying low-dimensional patterns in the tensor. CANDECOMP/PARAFAC (CP) decomposition model [2] which decomposes a tensor into a sum of rank-one components [15], is one of the most famous models because it is straightforward to analyze the underlying patterns.

Definition 1 (CANDECOMP/PARAFAC (CP) Decomposition). *Given a third-order tensor $\boldsymbol{\mathcal{X}} \in \mathbb{R}^{I\times J\times K}$ and a rank R, CP decomposition approximates $\boldsymbol{\mathcal{X}}$ to find factor matrices* $\{\mathbf{A} \in \mathbb{R}^{I\times R}, \mathbf{B} \in \mathbb{R}^{J\times R}, \mathbf{C} \in \mathbb{R}^{K\times R}\}$ *that minimize:*

$$\min_{\tilde{\boldsymbol{\mathcal{X}}}} ||\boldsymbol{\mathcal{X}} - \tilde{\boldsymbol{\mathcal{X}}}|| \text{ where } \tilde{\boldsymbol{\mathcal{X}}} = [\![\mathbf{A}, \mathbf{B}, \mathbf{C}]\!] = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r. \tag{1}$$

Note that $\circ$ represents an outer product and $\mathbf{a}_r \in \mathbb{R}^I$ is a $r$th column factor of $\mathbf{A}$ (similarly for $\mathbf{b}_r \in \mathbb{R}^J$ and $\mathbf{c}_r \in \mathbb{R}^K$). There are two ways to explain factor matrices in terms of their rows and columns. Each factor matrix is an embedding matrix corresponding to each mode, and each row of the factor matrix represents an embedding of an entity in that mode. For example, the row factor vector $\mathbf{a}_{i,:}$ of the factor matrix $\mathbf{A}$ is the $i$th entity embedding of the first mode. Each column of each factor matrix represents each feature of the embedding. What each feature means can be described as a hidden concept representing an entity's underlying relationship across different modes in a given tensor. For example, the $r$th rank-1 component $\mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \in \mathbb{R}^{I\times J\times K}$ corresponds to the $r$th hidden concept. We order values in each column of each factor matrix and identify which entities are strongly associated with the given concept. 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 semantically irrelevant high dimensions.

### 2.2 Inductive & Transductive Learning

In inductive learning, we train a model based on a training dataset that we already have and predict labels of a dataset that has never been seen before with the trained model. On the other hand, the goal of transductive learning is to train a model by leveraging information from unlabeled test data, resulting in better predictive performance. However, the model cannot classify new datasets after the model has been trained. Inductive learning methods are preferred over transductive learning methods [43] for news application tasks since embedding newly-uploaded news with a trained model is desirable rather than re-learning the whole data whenever new data are generated.

Tensor decomposition methods have been inherently restricted to transductive learning scenarios [9, 10, 44]. When new datasets become available, we have to decompose tensors made with existing and new datasets, which is not practical for news embedding when the existing datasets are large. Hence, we propose two variants of inductive learning algorithms which do not decompose an entire of existing datasets.

## 3 PROPOSED METHOD

We propose CITEM (**C**ompact **I**nterpretable **T**ensor graph multi-modal news **EM**bedding) which 1) effectively expresses multi-modal information of news, 2) explains each dimension of the news embedding with latent clusters of news, and 3) allows inductive learning for new unseen data Fig. 1 illustrates the main ideas of CITEM and we describe each of the steps in detail below.

### 3.1 Compact & Interpretable News Embeddings

In this section, 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

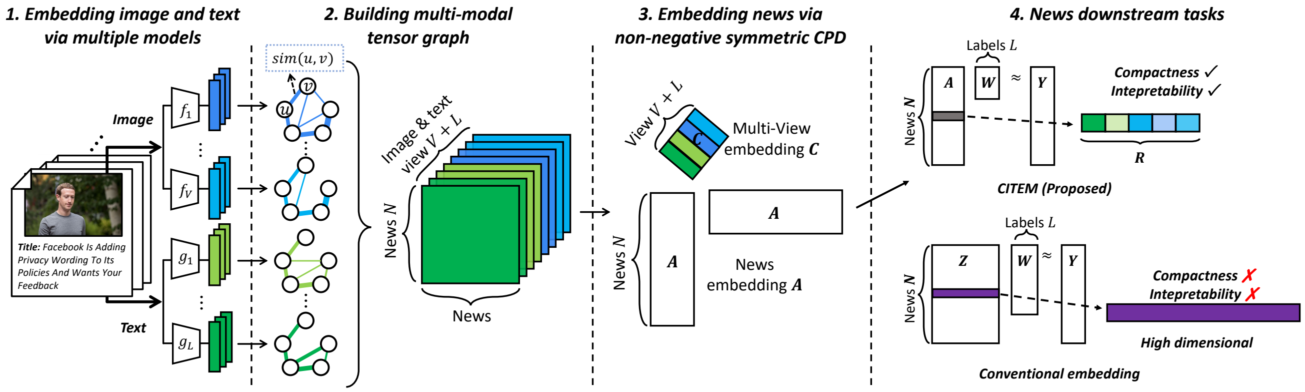


**Figure 1: Illustration of main ideas of CITEM. We extract multiple textual and visual features from news via various pre-trained models. We convert each feature space into a graph and stack them into a tensor. With non-negative symmetric CPD, we decompose the given tensor to obtain compact, interpretable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

articles. We then compute multiple text (title) and image (top image) embeddings from each article using different types of language and vision models, which allows for more accurate and discriminative news article representations even without fine-tuning the models. More specifically, text features extracted via the $\ell(1 \leq \ell \leq L)$th language model for all $N$ articles are denoted as $\mathbf{U}^{(\ell)} \in \mathbb{R}^{N \times d_l}$ where $d_l$ is an embedding dimension size fixed by the given model. image features extracted via the $v(1 \leq v \leq V)$th vision model for all $N$ articles are denoted as $\mathbf{V}^{(v)} \in \mathbb{R}^{N \times d_v}$. 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 $\mathbf{X}_u^{(\ell)} = \mathbf{U}^{(\ell)}\mathbf{U}^{(\ell)\top}$ and $\mathbf{X}_v^{(v)} = \mathbf{V}^{(v)}\mathbf{V}^{(v)\top}$. Next, we stack all graphs and build a multi-modal tensor graph or multi-layer graph $\mathcal{X} \in \mathbb{R}^{N \times N \times K}$ where $K = L + V$. Extensive prior work [9, 24] has demonstrated that tensor analysis in such multi-layer graphs, where there is an expectation of overlapping yet 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 [25].

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 $\{\mathbf{A} \in \mathbb{R}_+^{N \times R}, \mathbf{C} \in \mathbb{R}_+^{K \times R}\}$ that solves:

$$\min_{\tilde{\mathcal{X}}} ||\mathcal{X} - \tilde{\mathcal{X}}|| \text{ where } \tilde{\mathcal{X}} = [\![\mathbf{A}, \mathbf{A}, \mathbf{C}]\!] = \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{a}_r \circ \mathbf{c}_r. \quad (2)$$

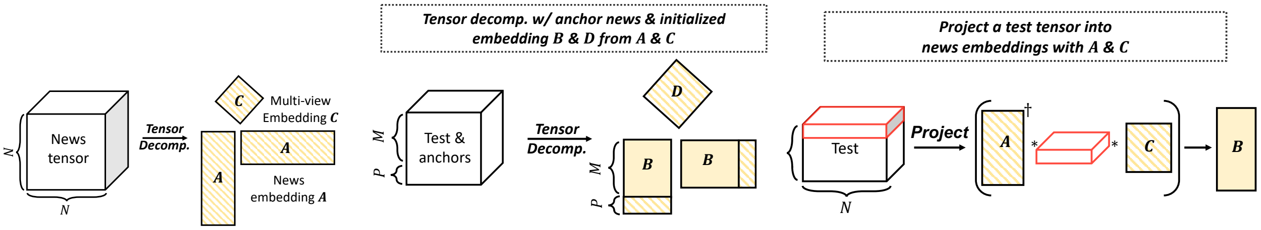
The $n$th row vector of factor matrix $\mathbf{A}$ corresponds to the $n$th news embedding whose dimensions are co-clustered with other news entities. The $k$th row vector of factor matrix $\mathbf{C}$ corresponds to a representation of the $k$th pre-trained model, which indicates its influence on each 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 [14] optimizer to compute the factor matrices.

## 3.2 Transductive & Inductive Learning

In this section, we discuss how we extend CITEM from transductive learning to inductive learning as shown in Fig. 2.

*3.2.1 Transductive Learning.* As we mention in Sec. 1, CPD-based tensor decomposition methods, even though it lends themselves to well-performing and interpretable representations, they have traditionally been restricted to transductive scenarios [10], where we decompose labeled and unlabeled data points into the same set of factors. However, when an unseen data point arrives, transductive-only models are not able to properly handle it. The reason behind this is that the computed factors *do not define* a projection to the embedding space as one would expect from matrix methods such as Singular Value Decomposition (SVD) [12] or even Tucker-based tensor methods [36] where the computed factors are proper projector matrices but do not define a set of interpretable latent factors but merely model the subspaces of the tensor [15].

We introduce CITEM where we assume that we have both training and test datasets already observable except labels of a test dataset. We build a tensor graph with both training and test dataset and decompose it together as shown in Fig. 2(a). We obtain embeddings from NS-CPD and we split them into training and test set when we obtain few labels. This embedding model takes advantage of information from data points in test dataset *but not from labels* such that it would show an inflated performance. However, in this


News
tensor
N
N
Tensor
Decomp.
C
Multi-view
Embedding C
A
A
News
embedding A
Tensor decomp. w/ anchor news & initialized
embedding B & D from A & C
M
P
Test &
anchors
Tensor
Decomp.
D
M
P
B
B
Project a test tensor into
news embeddings with A & C
Test
N
Project
A
†
*
*
C
B

(a) Transductive learning (CITEM) (b) Inductive variant 1 (CITEM-I1) (c) Inductive variant 2 (CITEM-I2)

**Figure 2: Comparison of transductive and two variants of inductive learning methods.**

transductive setting, the embedding model can not be used for additional unseen data points. In this section, we introduce two inductive variants that allow us to use our interpretable latent factor-based embeddings for unseen data.

*3.2.2 Inductive Learning Variant 1: Matching Anchor Points.* We propose CITEM-I1 which exploits anchor points as shown in Fig. 2(b). Assume that we construct a training tensor, decompose it, obtain training embeddings, and train a classifier for a given task. When we receive test datasets, we obtain test embeddings by following the same process as above. However, we can not directly use test news embeddings since their dimensions are not aligned since NS-CPD is unique up to component permutations. To address this issue, we carefully initialize test factor matrices from training factor matrices with anchor points. In the training phase, we obtain a news factor matrix $\mathbf{A} \in \mathbb{R}^{N \times R}$ and a modality factor matrix $\mathbf{C} \in \mathbb{R}^{K \times R}$ from a training tensor $\mathcal{X}$ via NS-CPD. In the inference phase, we use anchor points set $P_a = \{p | 1 \leq p \leq N\}$, $(|P_a| = P)$ that overlapped data points in training and test datasets to match the dimensions of training and test factors. We randomly sample anchor points of training datasets and make a new test dataset by adding anchor points. Hence, we build a new test tensor graph $\mathcal{Y} \in \mathbb{R}^{M \times M \times K}$ where $M = M' + P$ and $M'$ is the original number of test news and obtain a news factor matrix $\mathbf{B} \in \mathbb{R}^{M \times R}$ and a modality factor matrix $\mathbf{D} \in \mathbb{R}^{K \times R}$. Before decomposing $\mathcal{Y}$, we initialize $\mathbf{D}$ with $\mathbf{C}$. Also, we initialize a test news embedding corresponding to anchor points $\tilde{\mathbf{B}} \in \mathbb{R}^{P \times R}$ with a training news embedding corresponding to anchor points $\tilde{\mathbf{A}} \in \mathbb{R}^{P \times R}$. With this method, whenever we receive test datasets, we do not need to decompose again about whole training news datasets but care about test datasets and anchor datasets.

*3.2.3 Inductive Learning Variant 2: Projection.* We propose CITEM-I2 which transforms the tensor into compact embedding with results of training datasets without tensor decomposition as shown in Figure 2(c). We create a training tensor graph $\mathcal{X} \in \mathbb{R}^{N \times N \times K}$ and decompose it to obtain training factor matrices, $\mathbf{A}$ and $\mathbf{C}$. We create a test tensor graph $\mathcal{Y} \in \mathbb{R}^{M \times N \times K}$ by calculating a similarity between the training and test dataset. Note that the first mode indicates test data and the second one indicates existing training data. Each $m$th piece of news $\mathbf{Y}_m \in \mathbb{R}^{N \times K}$ represents similarity between $m$th test news and all training news with regard to $K$ views.

Then we exploit training factor matrices $\mathbf{A}$ and $\mathbf{C}$ to create test news embeddings without decomposing the test tensor. Our goal is to acquire a test factor matrices $\mathbf{B} \in \mathbb{R}^{M \times R}$. The $m$th news embedding $\mathbf{b}_{m,:}$ is generated as follows:

$$\mathbf{b}_{m,:} = \text{colsum}(\mathbf{E}_m) \text{ where } \mathbf{E}_m = \mathbf{A}^{\dagger}\mathbf{Y}_m\mathbf{C} \tag{3}$$

where colsum$(\cdot)$ denotes a column-wise summation.

# 4 EXPERIMENTS

We perform experiments to answer the following questions.

- **Q1 Overall Performance (Sec. 4.2).** How accurately does CITEM perform in news downstream tasks?
- **Q2 Compactness (Sec. 4.3).** How compact are CITEM embeddings?
- **Q3 Interpretability (Sec. 4.4).** Can CITEM produce interpretable results?
- **Q4 Inductive Learning (Sec. 4.5).** How do the two inductive variants behave differently?

## 4.1 Experimental Settings

We evaluate CITEM on two news-related downstream tasks: misinformation news detection and news categorization. We describe experimental settings for datasets, baselines, metrics, and hyperparameters. Note that all experiments are conducted on a machine equipped with an AMD Ryzen CPU and an NVIDIA RTX A6000.

*4.1.1 Dataset.* We evaluate the performance of CITEM and baselines on five real-world datasets. All datasets except for News Category are related to misinformation detection tasks. We construct a multi-modal tensor graph with five pre-trained models to extract six different feature vectors as described in Sec. 4.1.2. The first and the second modes of the tensor represent news entities while the last mode represents modalities.

- **Fakeddit**[1] [21]: multi-modal fake news benchmark dataset, providing text and image data and fine-grained fake news labels. We sample 10 percent of the dataset and make 13,633 samples which have 8,249 real news and 5,384 fake news. We construct the tensor of size $13,633 \times 13,633 \times 6$.
- **GossipCop**[2] [32–34]: 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 then construct a tensor of size $10,368 \times 10,368 \times 6$.

[1] https://github.com/entitize/Fakeddit
[2] https://github.com/KaiDMML/FakeNewsNet

- **PolitiFact**[2] [32–34]: 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**[3]: 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 a tensor graph of size $5,711 \times 5,711 \times 6$.
- **News Category**[4] [20]: 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 a tensor graph of size $9,976 \times 9,976 \times 3$.

Each dataset is randomly split into training and test sets with an 8:2 split ratio for classification tasks.

*4.1.2 Pre-trained models.* We utilize five pre-trained language and vision models to extract text and image features from news articles. Note that we extract both text and image features from CLIP, a multi-modal model. We use Hugging Face's[5] implementations for our experiments.

- **SBERT**[6] [29] is a language model and a variant of BERT [7] that specializes in generating sentence embedding. It generates 768-dim. text embeddings.
- **BART**[7] [17] is a language model specialized in handling complicated language tasks (e.g., text summarization). It generates 768-dim. text embeddings.
- **ResNet**[8] [11] is a vision model commonly used for various image-related tasks. It generates 2,048-dim. image embeddings.
- **ViT**[9] [8] is a vision transformer model that leverages a transformer architecture for image classification. It generates 768-dim. image embeddings.
- **CLIP**[10] [28] is a multi-modal model trained on a variety of image and text pairs. It generates 512-dim. text and image embeddings.

*4.1.3 Baselines.* We describe various baselines to evaluate our proposed method.

- Concat-Text: 2,048-dim. text embeddings from SBERT, BART, and CLIP models concatenated together.
- Concat-Img: 3,328-dim. image embeddings from ViT, ResNet, and CLIP models concatenated together.
- Concat-Both: 5,376-dim. text and image embeddings from all six models concatenated together.
- Concat-PCA: 768-dim. text and image embeddings obtained from applying Principal Component Analysis (PCA) on Concat-Both.
- CPD: standard CP decomposition with Alternating Least Square (ALS) optimization.
- RESCAL [22]: symmetric Tucker with L2 regularization optimized via Adam.
- CAFE [3]: a state-of-the-art method in misinformation detection where it learns cross-modal and unimodal features by estimating the divergence of different types of modality. For a fair comparison, we replace the feature extractor models in the original paper with the pretrained models in Sec. 4.1.2.

*4.1.4 Metrics.* We use the following five metrics: accuracy, recall, precision, F1-score, and Area under the curve (AUC) for fake/clickbait classification tasks and Accuracy, Micro-Recall, Micro-Precision, Micro-F1 score, and Micro AUC for news category classification.

*4.1.5 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 Logistic Regression (LR) model as the downstream classifier to evaluate the performance of 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.

## 4.2 Classification Performance

We show performance comparisons in Table 1. 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. Note that CAFE is trained with a supervised learning method, unlike all other baselines including the proposed method are unsupervised learning methods.

## 4.3 Compactness

Fig. 3 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. 3. 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.

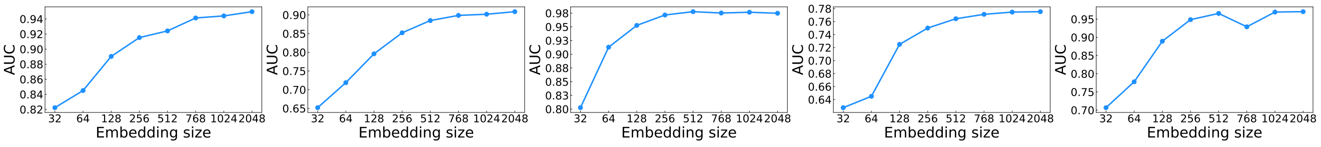
## 4.4 Interpretability

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 $|\mathbf{a}_{n,:}\mathbf{w}_n^l + d^l|$ with the $n$th news embedding $\mathbf{a}_{n,:}$ and corresponding coefficients $\mathbf{w}^l$ of a label $l$ and an intercept $d^l$. With the highest influence scores, we find the top-$k$ influential dimensions of $n$th embeddings from model decisions.

[2] https://github.com/KaiDMML/FakeNewsNet
[3] https://www.seekr.com/
[4] https://www.kaggle.com/datasets/rmisra/news-category-dataset
[5] https://huggingface.co
[6] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[7] https://huggingface.co/facebook/bart-large
[8] https://huggingface.co/microsoft/resnet-50
[9] https://huggingface.co/google/vit-base-patch16-224
[10] https://huggingface.co/sentence-transformers/clip-ViT-B-32

**Table 1: Performance comparison on news downstream tasks. Note that the best method is in bold, and the second-best method is underlined. CITEM produces a compact embedding by integrating different modalities and interpretable embedding due to its non-negativity.**

| Dataset | Model | Size | Acc. | Rec. | Prec. | F1 | AUC |
|---|---|---|---|---|---|---|---|
| Fakeddit | Concat-Text | 2,048 | 0.841 | 0.774 | 0.824 | 0.798 | 0.926 |
| | Concat-Img | 3,328 | 0.817 | 0.789 | 0.767 | 0.778 | 0.901 |
| | Concat-Both | 5,376 | 0.881 | 0.857 | 0.859 | 0.853 | **0.955** |
| | Concat-PCA | 768 | 0.877 | 0.844 | 0.853 | 0.848 | 0.952 |
| | CPD | 768 | 0.881 | 0.844 | 0.862 | 0.853 | 0.948 |
| | RESCAL | 256 | 0.875 | 0.841 | 0.850 | 0.846 | 0.940 |
| | CAFE | 96 | **0.897** | **0.882** | **0.861** | **0.872** | 0.894 |
| | CITEM (Proposed) | 768 | 0.869 | 0.827 | 0.848 | 0.837 | 0.945 |
| GossipCop | Concat-Text | 2,048 | 0.811 | 0.765 | 0.792 | 0.779 | 0.886 |
| | Concat-Img | 3,328 | 0.740 | 0.601 | 0.749 | 0.667 | 0.829 |
| | Concat-Both | 5,376 | 0.854 | 0.812 | **0.845** | 0.828 | **0.922** |
| | Concat-PCA | 768 | 0.853 | 0.809 | 0.844 | 0.826 | 0.918 |
| | CPD | 512 | 0.845 | 0.814 | 0.825 | 0.820 | 0.905 |
| | RESCAL | 512 | 0.830 | 0.771 | 0.826 | 0.797 | 0.898 |
| | CAFE | 96 | **0.865** | **0.903** | 0.771 | **0.832** | 0.873 |
| | CITEM (Proposed) | 768 | 0.844 | 0.789 | 0.841 | 0.814 | 0.903 |
| PolitiFact | Concat-Text | 2,048 | 0.915 | 0.884 | 0.905 | 0.894 | 0.965 |
| | Concat-Img | 3,328 | 0.736 | 0.395 | 0.895 | 0.548 | 0.786 |
| | Concat-Both | 5,376 | 0.934 | 0.884 | 0.950 | 0.916 | 0.973 |
| | Concat-PCA | 768 | 0.934 | 0.884 | 0.950 | 0.916 | 0.973 |
| | CPD | 64 | 0.943 | **0.930** | 0.930 | 0.930 | 0.971 |
| | RESCAL | 256 | 0.934 | 0.884 | 0.950 | 0.916 | 0.976 |
| | CAFE | 96 | 0.906 | 0.880 | 0.880 | 0.880 | 0.902 |
| | CITEM(Proposed) | 256 | **0.953** | **0.930** | **0.952** | **0.941** | **0.977** |
| Seekr | Concat-Text | 2,048 | 0.818 | 0.255 | 0.727 | 0.378 | 0.791 |
| | Concat-Img | 3,328 | 0.799 | 0.117 | 0.733 | 0.202 | 0.660 |
| | Concat-Both | 5,376 | 0.822 | 0.255 | 0.774 | 0.384 | 0.776 |
| | Concat-PCA | 768 | 0.815 | 0.266 | 0.694 | 0.385 | 0.766 |
| | CPD | 256 | 0.827 | 0.330 | 0.721 | **0.453** | 0.767 |
| | RESCAL | 512 | 0.822 | 0.213 | **0.870** | 0.342 | **0.792** |
| | CAFE | 96 | 0.802 | **0.696** | 0.188 | 0.296 | 0.752 |
| | CITEM (Proposed) | 768 | **0.831** | 0.287 | 0.818 | 0.425 | 0.772 |
| News Category | Concat-Text | 2,048 | 0.741 | 0.741 | 0.741 | 0.741 | 0.969 |
| | Concat-Img | 3,328 | 0.615 | 0.615 | 0.615 | 0.615 | 0.930 |
| | Concat-Both | 5,376 | **0.769** | **0.769** | **0.769** | **0.769** | 0.972 |
| | Concat-PCA | 768 | 0.767 | 0.767 | 0.767 | 0.767 | 0.972 |
| | CPD | 256 | 0.741 | 0.741 | 0.741 | 0.741 | 0.956 |
| | RESCAL | 512 | 0.756 | 0.756 | 0.756 | 0.756 | **0.973** |
| | CITEM (Proposed) | 768 | 0.752 | 0.752 | 0.752 | 0.752 | 0.968 |



(a) Fakeddit (b) GossipCop (c) PolitiFact (d) Seekr (e) News Category

**Figure 3: The AUC of CITEM according to the different embedding sizes (rank). As the rank size increases, the AUC increases. However, it performs well at embedding sizes 512 and 768 (10.5x and 7x smaller than the size of Concat-Both).**

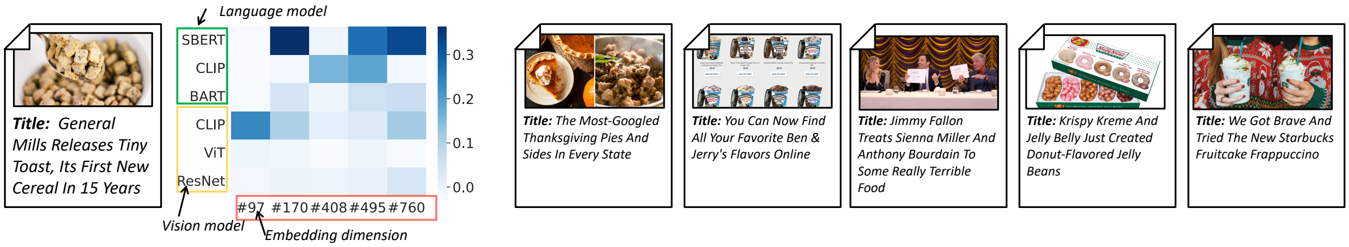


**Figure 4: Thanks to interpretable embeddings and multi-view embedding, we can identify the top-5 dimensions and the most influential multi-view for news category classification. This information allows us to determine which dimensions have a significant impact on the model's decisions.**

We then identify which dimensions correspond to which modalities based on multi-view embedding C. Fig. 4 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.

## 4.5 Inductive Learning

**Table 2: An AUC of the transductive and inductive variants of CITEM. We find that CITEM-I2 is highly effective and is able to match the performance of the transductive CITEM in most cases, and even exceed it (in the case of the Seekr and News Category datasets).**

| Model | CITEM | CITEM-I1 | CITEM-I2 |
|---|---|---|---|
| **Fakeddit** | 0.945 | 0.468 | 0.939 |
| **GossipCop** | 0.907 | 0.868 | 0.895 |
| **PolitiFact** | 0.979 | 0.960 | 0.977 |
| **Seekr** | 0.788 | 0.738 | 0.802 |
| **News Category** | 0.970 | 0.959 | 0.971 |

We compare the AUC score of a transductive method to two inductive variants as shown in Table 2. CITEM-I2 performs better than the CITEM-I1 since when we decompose the training and testing dataset separately, even in the presence of anchor points, there may be cases where the two decompositions have extracted non-fully-intersecting set of components, which can ultimately result in noisy features.

## 5 RELATED WORK

We review previous work on news embedding models with regard to types of information and their applications. Online news has become so popular that people are exposed to it every day and is so fast and massive that it is difficult to recommend news to individuals. Due to this problem, news recommendations have had attention, and news modeling methods have been greatly improved recently [40] since news embedding where its goal is to learn the main content of the news is an essential step for accurate news recommendations. Natural language processing (NLP) techniques have been incredibly successful in many fields, numerous methods employed pre-trained language models to encode textual information such as headlines and bodies consisting of the key contents of news articles [16, 18, 23, 39].

For news recommendations, Okura et al. [23] exploited a body of news to represent news; Ma et al. [19] proposed a news network embedding containing semantic features and the relationship between news and its event elements; Liu et al. [18] proposed a news embedding based on a document level exploiting title and body. To enhance news embedding, many methods actively incorporated various information from the news. Wang et al. [38] proposed a knowledge-aware recommendation method exploiting text entities to learn common sense and knowledge information. Santosh et al. [30] utilized news-news relatedness in addition to titles, bodies, and categories. Also, several methods exploited categories and topics categories [26, 30, 37, 39] and leverage visual information [41, 42] in addition to text information. Recently, various methods attempted to exploit different types of information such as tags [41], sentiment [37], popularity [4], and temporal and spatial information [27, 35, 42] to understand the characteristics of news better. However, those approaches are not easily interpretable.

## 6 CONCLUSION

We propose CITEM, a tensor-based framework for 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. CITEM is effective in the transductive setting, where we can compute a decomposition across all of the articles at training time. However, this case is not always realistic, so we propose two variants of CITEM for the more realistic inductive setting: CITEM-I1 and CITEM-I2. This allows us to embed articles unseen at training time and, in the case of CITEM-I2, maintain similar performance to the transductive variant. Our experimental results show that CITEM is up to 7× more compact than baselines while achieving similar performance. We further develop the framework with end-to-end training, which can be subsequently applied to specific downstream tasks.

## 7 ACKNOWLEDGEMENTS

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, 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] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022*. 2897–2905.

[4] Sungmin Cho, Hongjun Lim, Keunchan Park, Sungjoo Yoo, and Eunhyeok Park. 2021. On the Overlooked Significance of Underutilized Contextual Features in Recent News Recommendation Models. *arXiv preprint arXiv:2112.14370* (2021).

[5] Lingyang Chu, Yanyan Zhang, Guorong Li, Shuhui Wang, Weigang Zhang, and Qingming Huang. 2014. Effective multimodality fusion framework for cross-media topic detection. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 3 (2014), 556–569.

[6] 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.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] 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).

[9] Ekta Gujral. 2021. *Modeling and Mining Multi-Aspect Graphs With Scalable Streaming Tensor Decomposition*. University of California, Riverside.

[10] Ekta Gujral and Evangelos E Papalexakis. 2018. Smacd: Semi-supervised multi-aspect community detection. In *Proceedings of the 2018 SIAM International Conference on Data Mining*. SIAM, 702–710.

[11] 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.

[12] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. 50–57.

[13] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 261–269.

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

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

[16] Vaibhav Kumar, Dhruv Khattar, Shashank Gupta, and Vasudeva Varma. 2017. Word semantics based 3-d convolutional neural networks for news recommendation. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 761–764.

[17] 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).

[18] Jialu Liu, Tianqi Liu, and Cong Yu. 2021. Newsembed: Modeling news through pre-trained document representations. *arXiv preprint arXiv:2106.00590* (2021).

[19] Ye Ma, Lu Zong, Yikang Yang, and Jionglong Su. 2019. News2vec: News network embedding with subnode information. 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)*. 4843–4852.

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

[21] Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. *arXiv preprint arXiv:1911.03854* (2019).

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Keunchan Park, Jisoo Lee, and Jaeho Choi. 2017. Deep neural networks for news recommendations. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2255–2258.

[27] Tao Qi, Fangzhao Wu, Chuhan Wu, and Yongfeng Huang. 2021. Pp-rec: News recommendation with personalized user interest and time-aware news popularity. *arXiv preprint arXiv:2106.01300* (2021).

[28] 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.

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

[30] TYSS Santosh, Avirup Saha, and Niloy Ganguly. 2020. MVL: Multi-view learning for news recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1873–1876.

[31] Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 395–405.

[32] 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).

[33] 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.

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

[35] 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*. 293–302.

[36] M Alex O Vasilescu and Demetri Terzopoulos. 2007. Multilinear projection for appearance-based recognition in the tensor framework. In *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 1–8.

[37] Heyuan Wang, Fangzhao Wu, Zheng Liu, and Xing Xie. 2020. Fine-grained interest matching for neural news recommendation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 836–845.

[38] Hongwei Wang, Fuzheng Zhang, Xing Xie, and Minyi Guo. 2018. DKN: Deep knowledge-aware network for news recommendation. In *Proceedings of the 2018 world wide web conference*. 1835–1844.

[39] 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).

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

[41] Chuhan Wu, Fangzhao Wu, Tao Qi, and Yongfeng Huang. 2021. Mm-rec: multimodal news recommendation. *arXiv preprint arXiv:2104.07407* (2021).

[42] 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*. 3881–3890.

[43] Zhilin Yang, William Cohen, and Ruslan Salakhudinov. 2016. Revisiting semi-supervised learning with graph embeddings. In *International conference on machine learning*. PMLR, 40–48.

[44] Anil R Yelundur, Vineet Chaoji, and Bamdev Mishra. 2019. Detection of review abuse via semi-supervised binary multi-target tensor decomposition. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2134–2144.

[45] Daniel Zügner. 2022. *Adversarial Robustness of Graph Neural Networks*. Ph. D. Dissertation. Technische Universität München.