Cross-modal Ambiguity Learning for Multimodal Fake News Detection

Yixuan Chen
School of Computer Science
Shanghai Key Laboratory of Data Science
Fudan University
Shanghai, China
yixuanchen20@fudan.edu.cn

Dongsheng Li
Microsoft Research Asia
Shanghai, China
dongsli@microsoft.com

Peng Zhang
School of Computer Science
Shanghai Key Laboratory of Data Science
Fudan University
Shanghai, China
zhangpeng_@fudan.edu.cn

Jie Sui
University of Chinese Academy of Sciences
Beijing, China
suijie@ucas.ac.cn

Qin Lv
University of Colorado Boulder
Boulder, United States
qin.lv@colorado.edu

Tun Lu, Li Shang*
School of Computer Science
Shanghai Key Laboratory of Data Science
Fudan University
Shanghai, China
{lutun,lishang}@fudan.edu.cn

ABSTRACT

Cross-modal learning is essential to enable accurate fake news detection due to the fast-growing multimodal contents in online social communities. A fundamental challenge of multimodal fake news detection lies in the inherent ambiguity across different content modalities, i.e., decisions made from unimodalities may disagree with each other, which may lead to inferior multimodal fake news detection. To address this issue, we formulate the cross-modal ambiguity learning problem from an information-theoretic perspective and propose CAFE — an ambiguity-aware multimodal fake news detection method. CAFE consists of 1) a cross-modal alignment module to transform the heterogeneous unimodality features into a shared semantic space, 2) a cross-modal ambiguity learning module to estimate the ambiguity between different modalities, and 3) a cross-modal fusion module to capture the cross-modal correlations. CAFE improves fake news detection accuracy by judiciously and adaptively aggregating unimodal features and cross-modal correlations, i.e., relying on unimodal features when cross-modal ambiguity is weak and referring to cross-modal correlations when cross-modal ambiguity is strong. Experimental studies on two widely used datasets (Twitter and Weibo) demonstrate that CAFE outperforms state-of-the-art fake news detection methods by 2.2-18.9% and 1.7-11.4% on accuracy, respectively.

CCS CONCEPTS

• **Information systems** → **Data mining**; **Social networks**; • **Computing methodologies** → **Artificial intelligence**.

*Corresponding author.

WWW '22, April 25–29, 2022, Virtual Event, Lyon, France.

ACM ISBN 978-1-4503-9096-5/22/04…$15.00
https://doi.org/10.1145/3485447.3511968

KEYWORDS

Multimodal Learning, Fake News Detection, Cross-modal Ambiguity Learning

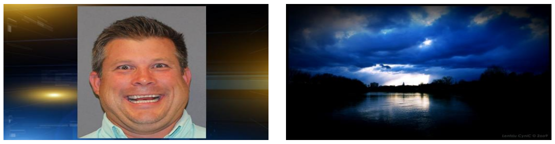
ACM Reference Format:
Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, and Tun Lu, Li Shang. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022 (WWW '22), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3485447.3511968

1 INTRODUCTION

Online social media has become the primary platform for daily information sharing among people. Studies have shown that over three billion people consider Facebook and Twitter as their primary daily information sources [21]. While people enjoy the convenience of online social media, the lack of systematic efforts to verify the credibility of online posts has led to the wide and fast spread of fake news across social platforms [19, 38, 40]. To tackle this problem, fake news detection has received increasing research attention in recent years [3, 4, 14, 22, 23, 34, 35].

Online social content, such as microblog, has quickly evolved from text only to multimodality, often containing both text and images. While early works on fake news detection focused on text-only content analysis, cross-modal content analysis can offer complementary benefits to assist with fake news detection [24, 27, 28]. For instance, recent works aim to fuse multimodal content information to boost the performance of fake news detection [1, 17]. However, the prior works have not explicitly considered the inherent ambiguity across different content modalities, and thus lead to inferior performance.

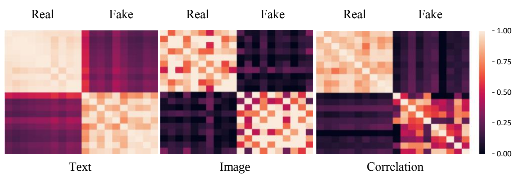
Figure 1 illustrates the potential impact of cross-modal ambiguity on fake news detection. The fake news example (left figure) tells a fictional death story but includes a smiling person's image. The text and image present strong cross-modal ambiguity, and multimodal feature fusion, which captures such cross-modal information gap, can help improve classification accuracy. In contrast, the real news example (right figure) expresses sad emotion with a blue image



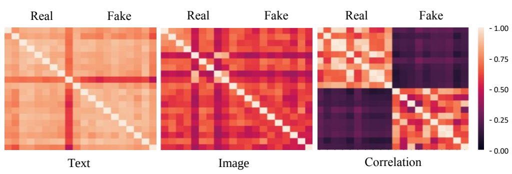
Fake news: "An employee of the Jefferson County morgue died this morning after being accidentally cremated by one of his coworkers."

Real news: "You left in peace, left me in pieces."

Figure 1: Illustrations of cross-modal ambiguity.



(a) Cross-modal correlation may be unhelpful or even harmful when text and image alone are sufficient.



(b) Cross-modal correlation can present extra insights when text and image alone are insufficient.

Figure 2: Illustration of the importance of ambiguity-aware cross-modal correlation using the Weibo dataset [15]. Each cell of the heat map represents the cosine similarity between the representations of each text or image pair.

included. The unimodalities are emotionally consistent and are sufficient to determine the news credibility, while cross-modal fusion features are unnecessary or even introduce noise to the classification task. Statistically, as shown in Figure 2, our empirical studies using the Weibo dataset released by [15] show that cross-modal information may be unhelpful or even harmful when unimodal fake news detectors are sufficient and agree with each other (42.9% posts). On the other hand, cross-modal information is crucial when unimodal fake news detectors are insufficient (11.9% posts). Together, the aforementioned two cases account for 54.8% posts of the Weibo dataset. Therefore, the multimodal fake news detection methods should be aware of the ambiguity between different modalities and adaptively aggregate discriminative cross-modal features with unimodal features to perform better multimodal classification.

In this paper, we first formulate the cross-modal ambiguity learning problem from an information-theoretic perspective, using the distributional divergence between different unimodal features to quantify their ambiguity. Then, we propose CAFE — an ambiguity-aware multimodal fake news detection method. CAFE consists of 1) a cross-modal alignment module, which can transform the heterogeneous unimodal features into a shared semantic space with an auxiliary semantic regularization task, 2) a cross-modal ambiguity learning module, which can estimate the ambiguity between different modalities via evaluating the Kullback-Leibler (KL) divergence between the distributions of unimodal features, and 3) a cross-modal fusion module, which can capture the cross-modal correlations by learning the semantic interactions between different modalities to provide complementary features for fake news detection. CAFE can improve fake news detection accuracy by adaptively aggregating unimodal features and cross-modal correlations, i.e., relying on unimodal features when cross-modal ambiguity is weak and referring to cross-modal correlations when cross-modal ambiguity is strong. Experimental studies on two widely used datasets (Twitter and Weibo) demonstrate that CAFE can outperform state-of-the-art fake news detection methods by 2.2-18.9% and 1.7-11.4% in terms of accuracy, respectively.

The main contributions of this work are as follows:

- We formulate the cross-modal ambiguity learning problem, a key challenge to multimodal fake news detection, and present a KL divergence based method to quantify the ambiguity between text and image by estimating the divergence of their feature distributions.
- We propose CAFE — an ambiguity-aware multimodal fake news detection method to adaptively aggregate unimodal features and cross-modal correlations, governed by the learnt ambiguity score.
- We perform experiments on two widely used datasets — Twitter and Weibo. The results demonstrate that CAFE can outperform state-of-the-art fake news detection methods by 2.2-18.9% and 1.7-11.4% in terms of accuracy on the two datasets, respectively.

The rest of this paper is organized as follows. Section 2 discusses the related works on multimodal fake news detection. Section 3 formulates the cross-modal ambiguity learning problem. Section 4 details the proposed multimodal fake news detection method and presents a KL divergence based method for cross-modal ambiguity learning. Section 5 presents and discusses the experimental results. We conclude the work in Section 6.

2 RELATED WORK

2.1 Unimodal Approach

A large body of fake news detection works focused on unimodal information, one line of works relied on the text content analysis [5, 7–10, 24, 25, 32], the second line of works aggregated user profiles and their responses to identify fake news [18, 35, 36], and the third line of works considered image content only in posts [12, 16, 20]. Recently, Qian *et al.* built a text-based method to capture semantic information from article text, and proposed a generative model of user responses to assist fake news detection [25]. From the same perspective of utilizing user responses, Yang *et al.* aggregated user profiles and their responses to each targeted post via a graph-based detector to identify fake news [36]. Beside text content, recent works on fake news detection have also considered image content in posts [12, 16]. Even though visual features have been extensively studied in computer vision tasks [20], there are limited works applying visual features in the context of fake news detection. One potential challenge comes from the semantic gap between information-rich content and symbolic pixel values. Gupta *et al.* and Jin *et al.* both claimed that the spreading pattern of image content across a social platform exhibits discriminating features, which are suitable for fake image detection [12, 16].

2.2 Multimodal Approach

More recently, several methods were proposed to leverage cross-modal discriminative patterns to improve the accuracy of fake news detection [17, 37]. The early work [15] developed a fusion method that jointly considers image, text, and social context features for fake news detection. To learn cross-modal correlations, a variable autoencoder [17] was proposed to reconstruct textual representations and visual representations by learning probabilistic latent variable model, and then quantify the cross-modal correlation between text and image. The proposed work demonstrates good performance but with high computational cost. The EANN method [31] leverages textual and visual information via feature concatenation, and then utilizes a multi-task learning framework for event classification and fake news detection simultaneously. The event-classification helps remove post-specific information from the fake news detection and keep post-invariant rumor-discriminative features for accurate fake news detection. The MKEMN method [37] combines aligned embeddings of text, image and knowledge to learn multimodal representations of each post for multimodal fake news detection. The SAFE method [39] defines the relevance between news textual and visual information as a cosine similarity modification, which is fed into a classifier to detect fake news. Similarly, Xue *et al.*[33] proposed to capture the similarity of multimodal data, semantic features of texts and images and incorporate error level analysis algorithm to capture physical features of the visual modalities.

Existing works on multimodal fake news detection represent individual unimodal information separately, and the cross-modal semantic gap could limit their capability to effectively exploit cross-modal feature correlation. Furthermore, existing works on cross-modal feature fusion do not explicitly consider the ambiguity across different modalities and may fail to effectively leverage the cross-modal information as demonstrated in our case studies.

3 CROSS-MODAL AMBIGUITY LEARNING PROBLEM DEFINITION

In this section, we formulate the key problem to multimodal fake news detection — cross-modal ambiguity learning. Given a multimodal dataset $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, each sample $(\mathbf{x}, y) \in \mathcal{D}$ contains multiple unimodality information denoted as $(\mathbf{x}, y) = \{\{x^u\}_n, y\}$, where x^u denotes the information from the u-th modality, $\{\cdot\}_n$ denotes the collection of information from all n modalities and y is the label of $\mathbf{x}$. For instance, x^u could be text, image, video, etc., in the multimodal classification tasks.

A multimodal classification task aims to learn a rich set of features from input $\mathbf{x}$, i.e., unimodal features and cross-modal correlations in a unified representation space, and then map it onto the most probable label y in the label space. One unique characteristic of multimodal classification task lies in the inherent cross-modal ambiguity, which affects the efficacy of cross-modal correlation, hence the classification performance. To better understand the problem, we formally define the cross-modal ambiguity as follows:

Definition 1. *Given each data sample $(\mathbf{x}, y) \in \mathcal{D}$, the cross-modal ambiguity $a_{\mathbf{x}}^{i,j}$ between the i-th modality and the j-th modality is defined as a distance measure between the i-th and the j-th modality, i.e., $D(x^i, x^j)$.*

As shown in the aforementioned motivating example, cross-modal ambiguity measures the information gap between unimodal information. When cross-modal ambiguity is weak, i.e., the information gap across unimodalities is small, unimodal information alone is sufficient for accurate fake news detection. In contrast, when cross-modal ambiguity is strong with significant information gap across uni-modalities, unimodal information alone is insufficient, and cross-modal correlation can provide important and complementary information to the classification task. Therefore, cross-modal ambiguity learning is an important measure to decide when unimodal information is sufficient and when cross-modal information is essential.

In this work, we tackle the cross-modal ambiguity learning problem. More specifically, given an online news dataset, corresponding with two modalities: 1) text, denoted as x^t and 2) image, denoted as x^v, our goal is to learn the cross-modal ambiguity $a_{\mathbf{x}}^{t,v}$ for each news article. To simplify the notation, we omit the superscript and use $a_{\mathbf{x}}$ to denote the ambiguity for the rest of this paper.

The cross-modal ambiguity can be measured by the similarity, or the distance, between unimodal distributions using methods such as KL divergence and Wasserstein distance [29]. In this work, we propose a KL divergence based method for cross-modal ambiguity learning.

4 PROPOSED METHOD

In this paper, we propose CAFE to tackle the problem of multimodal fake news detection via cross-modal ambiguity learning. As shown in Figure 3, CAFE consists of: 1) *modal-specific encoder*, which encodes the unimodal information into embeddings via modality-specific encoders; 2) *cross-modal alignment*, which transforms the original unimodal embeddings into a shared space via an auxiliary cross-modal learning task; 3) *cross-modal ambiguity learning*, which estimates the ambiguity between unimodal features by learning from the distributional divergence of unimodal features. 4) *cross-modal fusion*, which fuses the aligned unimodal features into the cross-modal feature to facilitate the classification when cross-modal ambiguity is strong. 5) *classifier*, which first obtains the multimodal representations by concatenating unimodal embeddings and cross-modal correlations, governed by the cross-modal ambiguity, and then makes the final predictions.

4.1 Modal-specific Encoder

We represent the text and images associated with each news article by vectors to entangle key explanatory factors of variation behind the data [2]. Since the modal-specific encoders are not the focus of this work, we adopt the off-the-shelf techniques. More specially, for each news $\mathbf{x}$, we leverage pre-training techniques to encode its text x^t and image x^v into unimodal embedding e^t and e^v, respectively.

4.1.1 Text Encoder. Given a text x^t with a set of words, we adopt a pretrained BERT model [11] to obtain its embedding e^t. The textual embedding $e^t \in \mathbb{R}^{256}$ is obtained by using a fully connected layer to transform the temporal textual attributes extracted by BERT.

4.1.2 Image Encoder. Given an image x^v, we adopt a popular pretrained method — ResNet-34 [13] to learn meaningful representations from images. The final visual embedding $e^v \in \mathbb{R}^{512}$ is obtained

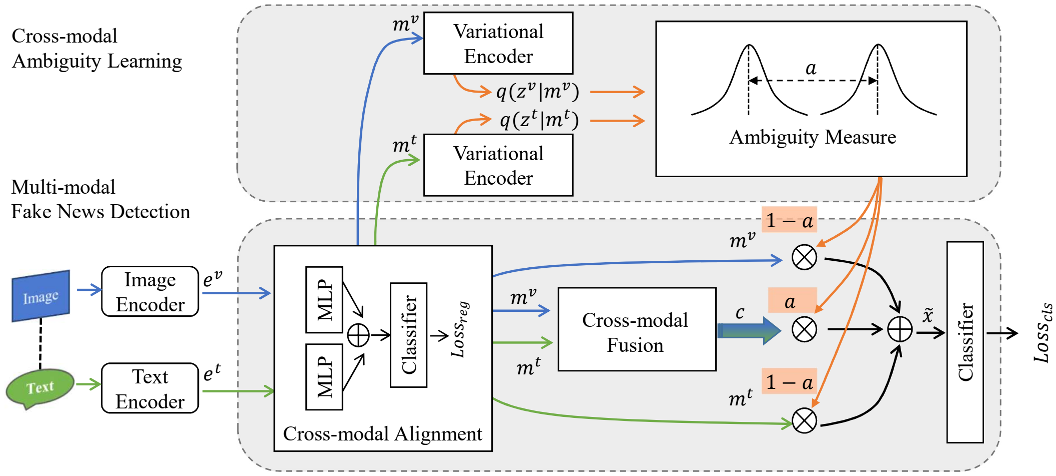


Figure 3: The architecture of the proposed CAFE method. For news with different levels of ambiguity, the proposed cross-modal ambiguity learning module can adaptively aggregate the unimodal features and cross-modal correlations to improve fake news classification. We set the weight of cross-modal correlation as a and the weight of unimodal features as $1-a$, so that the classifier will rely more on cross-modal correlation when a is large, i.e., stronger ambiguity appears.

by using a fully connected layer (Linear) to transform the regional features captured by ResNet-34 to fit our task.

4.2 Cross-modal Alignment

Features from different modalities may have huge semantic gaps, so that we need to align the features from different modalities by transforming the unimodal embeddings into a shared space. To this end, we propose to solve an auxiliary correlation learning task to help achieve cross-modal feature alignment. More specifically, we design a binary classification task to identify whether a pair of textual and visual embeddings shares a common semantics or not, which terms as *Semantic Regularization*.

Given each text-image pair, we first define the semantic correlation is positive or negative, i.e., labeled by 1 or 0, respectively. In this work, the semantic correlation of a text-image pair is defined as positive if the textual and visual embeddings are from the same piece of real news, and negative if the textual and visual embeddings are from different pieces of real news. Then, we randomly sample positive text-image pairs and negative text-image pairs to generate a synthetic dataset $\mathcal{D}_2$ for the auxiliary correlation learning task.

Building upon the previous unimodal embeddings e^t and e^v, the proposed cross-modality alignment module consists of a modality-specific multilayer perceptron (MLP) and a modality-shared layer to jointly learn the shared semantics. Then, the joint embeddings are fed to an average pooling layer, which is followed by a full-connected layer as a binary (positive or negative) classifier. The entire module is trained with positive and negative pairs using the cosine embedding loss with margin d as follows:

$$\mathcal{L}_{reg} = \begin{cases} 1 - \cos\left(e^t, e^v\right) & \text{if } y_2 = 1. \\ \max\left(0, \cos\left(e^t, e^v\right) - d\right) & \text{if } y_2 = 0. \end{cases} \tag{1}$$

$\cos(\cdot)$ is the normalized cosine similarity and we set the margin d as 0.2 due to empirical studies. The above objective is to maximize the cosine similarity of embeddings between positive text-image pairs, and minimize it between negative pairs, up to a specified margin. With the gradients from back-propagation, the semantic regularization can automatically force heterogeneous multimodal embeddings into a shared semantic space.

Finally, we jointly train the cross-modality alignment module to produce the semantically aligned unimodal representations m^t and m^v as the input of the *cross-modal ambiguity learning* module and the *cross-modal fusion* module.

4.3 Cross-modal Ambiguity Learning

Following the definition of cross-modal ambiguity, we propose an ambiguity learning method via evaluating the Kullback-Leibler (KL) divergence between unimodal distributions approximated by two modal-specific variational autoencoders. The learned ambiguity score is then used to adaptively control the contribution of cross-modal features and unimodal features in fake news detection. Therefore, when unimodal features present strong ambiguity, the cross-modal fake news detector should pay more attention to cross-modal features, and vice versa.

The unimodal features are fixed for each given input sample, so that it is challenging to know their distributions. To tackle this problem, we model the unimodal features from a generative perspective, i.e., the unimodal features (m^t or m^v) are sampled from a latent space $\mathbb{R}^d$ with isotropic Gaussian priors. Also, we assume the distributional divergence between unimodal features represent the information gap between unimodalities, i.e., we can use the divergence over feature space to approximate their ambiguity.

Specially, the variational posterior for an unimodal observation can be denoted as: $q(z|m) = \mathcal{N}(z|\mu(m), \sigma(m))$, in which the mean μ and variance σ can be obtained from the modal-specific encoder. More formally, for each data sample $\mathbf{x}_i$ with aligned textual feature m_i^t and image feature m_i^v, the variational posteriors of the two modalities can be defined as follows:

$$q\left(z_i^t|m_i^t\right) = \mathcal{N}\left(z_i^t \mid \mu\left(m_i^t\right), \sigma\left(m_i^t\right)\right), \tag{2}$$

$$q\left(z_i^v|m_i^v\right) = \mathcal{N}\left(z_i^v \mid \mu\left(m_i^v\right), \sigma\left(m_i^v\right)\right). \tag{3}$$

Considering the distribution over the entire dataset, we have

$$\begin{aligned} q\left(z^t\right) &= \mathbb{E}_{\mathrm{Pr}_{\mathrm{data}}(m^t)}[q\left(z^t|m^t\right)] = \frac{1}{N}\sum_{i=1}^{N} q\left(z_i^t|m_i^t\right), \\ q\left(z^v\right) &= \mathbb{E}_{\mathrm{Pr}_{\mathrm{data}}(m^v)}[q\left(z^v|m^v\right)] = \frac{1}{N}\sum_{i=1}^{N} q\left(z_i^v|m_i^v\right). \end{aligned} \tag{4}$$

Then, the ambiguity of different modalities in data sample $\mathbf{x}_i$ can be measured by the averaged KL divergence between unimodal distributions as follows:

$$a_i^1 = \left(\frac{\mathrm{D}_{KL}\left(q\left(z_i^t||m_i^t\right)||q\left(z_i^v||m_i^v\right)\right)}{\mathrm{D}_{KL}\left(q\left(z^t\right)||q\left(z^v\right)\right)}\right), \tag{5}$$

$$a_i^2 = \left(\frac{\mathrm{D}_{KL}\left(q\left(z_i^v||m_i^v\right)||q\left(z_i^t||m_i^t\right)\right)}{\mathrm{D}_{KL}\left(q\left(z^v\right)||q\left(z^t\right)\right)}\right), \tag{6}$$

$$a_i = \mathrm{sigmoid}\left(\frac{1}{2}\left(a_i^1 + a_i^2\right)\right). \tag{7}$$

Here $\mathrm{D}_{KL}(\cdot)$ denotes the KL divergence, and the ambiguity score a_i is computed as the symmetrized KL divergence obtained by averaging the normalized value of $\mathrm{D}_{KL}\left(q\left(z_i^t||m_i^t\right)||q\left(z_i^v||m_i^v\right)\right)$ and $\mathrm{D}_{KL}\left(q\left(z_i^v||m_i^v\right)||q\left(z_i^t||m_i^t\right)\right)$. sigmoid$(\cdot)$ is the activation function used to map the ambiguity scores to be between 0 and 1.

A small ambiguity score indicates that the two unimodal distributions are close to each other. Thus, we utilize the ambiguity score a_i as the weight to govern the fusion of unimodal features and cross-modal features in both training and inference, i.e., the cross-modal ambiguity learning help adaptively leverage cross-modal feature and drop out unimodal features when the ambiguity is large, and vice versa.

4.4 Cross-modal Fusion

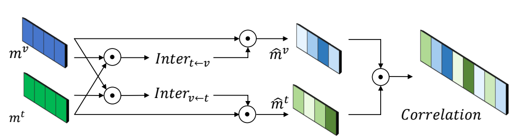


Figure 4: Architecture of the proposed cross-modal fusion module.

Cross-modal correlations can capture the semantic interactions between different modalities to provide complementary features for fake news detection, especially when text and image alone provide contradict predictions on the same news. To this end, we design the *cross-modal fusion* module to learn such ambiguity-aware cross-modality correlations.

Given the aligned unimodal representations m^t and m^v from the *cross-modal alignment* module, we first obtain the inter-modal attention weights $InterC$ by calculating the association weights between unimodal representations, which can help aggregate information from text features to each of the visual features, and vice versa. After normalizing the raw feature map by the square root of the dimension size and passing it over a softmax function, we obtain two sets of inter-modal weight maps as follows:

$$InterC_{t\leftarrow v} = \mathrm{softmax}\left([m^t][m^v]^T/\sqrt{dim}\right). \tag{8}$$

$$InterC_{v\leftarrow t} = \mathrm{softmax}\left([m^v][m^t]^T/\sqrt{dim}\right). \tag{9}$$

Since the correlation between all textual features and one visual feature can be regarded as the weighted sum of the textual features and vice versa, we can obtain the explicit correlation map by updating the original unimodal embedding vector as follows:

$$\hat{m}^t = InterC_{T\leftarrow I} \times m^t. \tag{10}$$

$$\hat{m}^v = InterC_{I\leftarrow T} \times m^v. \tag{11}$$

Previous works in multimodal fake news detection have used simple concatenation as the fusion approach [15, 17], which may fail to capture complex cross-modal interactions [1]. In contrast, we fuse the textual features and visual features by their interaction matrix c, which is formally defined as an outer product between $\hat{m}^t$ and $\hat{m}^v$ as follows:

$$c = \hat{m}^t \otimes \hat{m}^v. \tag{12}$$

$\otimes$ denotes outer product. The final correlation matrix c is flattened into a vector.

4.5 Classifier

The input of the classifier is obtained by adaptively concatenating two sets of embeddings: the unimodal representations from the *cross-modal alignment* module and the cross-modality correlations from the *cross-modal fusion* module, which is governed by the cross-modal ambiguity score $a_{\mathbf{x}}$ from the *cross-modal ambiguity learning* as follows:

$$\tilde{\mathbf{x}} = (a_{\mathbf{x}} \times c) \oplus \left((1-a_{\mathbf{x}}) \times m^t\right) \oplus \left((1-a_{\mathbf{x}}) \times m^v\right), \tag{13}$$

where $\oplus$ represents the concatenation operation. Then, we feed the final representation $\tilde{\mathbf{x}}$ into a fully-connected network to predict the label $\tilde{y}_{cls}$ as:

$$\tilde{y_1} = \mathrm{softmax}\left(MLP\left(\tilde{\mathbf{x}}\right)\right). \tag{14}$$

Since fake news detection is a binary classification task, we apply the cross-entropy loss $\mathcal{L}_1$ over all labeled pairs between the ground-truth y_1 and the predicted scores $\tilde{y_1}$ as follows:

$$\mathcal{L}_{cls} = y_1 log\left(\tilde{y_1}\right) + \left(1-\tilde{y_1}\right) log\left(1-y_1\right). \tag{15}$$

Next, we discuss the optimization strategy for the proposed method. The auxiliary semantic regularization task aims to bridge the semantic gaps between textual features and image features which may not be totally helpful for the classification task, so we limit its effect by placing a weight $\beta \in (0,1)$ on its loss function. By combining the loss functions from the main classification task and the auxiliary learning task, the final loss function for CAFE is defined as follows:

$$\mathcal{L} = \mathcal{L}_{cls} + \beta\mathcal{L}_{reg}. \tag{16}$$

The training of CAFE is accomplished via stochastic gradient decent by looping over each of the two tasks as presented in Algorithm 1. More specifically, we adopt an alternative optimization procedure for training the model of CAFE, in which we first train the auxiliary task and then train the main task in each epoch until the loss converges.

Algorithm 1 Model training of CAFE.

Input: Datasets: $\mathcal{D}_1$ for the main task, $\mathcal{D}_2$ for the auxiliary task
Output: Model parameters: Θ_1 for the main task, Θ_2 for the auxiliary task

while not converge **do**
  for the auxiliary task **do**
    Sample minibatch from $\mathcal{D}_2$.
    Compute loss using $\mathcal{L}_{reg}\left(e^t, e^v, y_2\right)$.
    Update parameters in Θ_2 by Adam.
  end for
  for the main task **do**
    Sample minibatch from $\mathcal{D}_1$.
    Compute loss using $\mathcal{L}_{cls}\left(m^t, m^v, y_1\right)$.
    Update parameters in Θ_1 by Adam.
  end for
end while

5 EXPERIMENTS

5.1 Experimental Setup

Datasets. We use two real-world datasets collected from social medias. The datasets are described as follows:

1) The *Twitter* dataset was released for MediaEval Verifying Multimedia Use task [6]. Given the focus on the text and image, following existing works we filter the tweets with videos attached. In experiments, we keep the same data split scheme as the benchmark [6], which is also the same as all the compared methods. The training set contains 6, 840 real news and 5, 007 fake news and the test set contains 1, 406 posts.

2) The *Weibo* dataset was released by Jin *et al.*[15], which has been widely used in prior multimodal fake news detection works. The real ones were collected from Xinhua News Agency, an authoritative news source of China. The fake ones were gathered by crawling the official fake news debunking system of Weibo over a time span from May 2012 to January 2016. In experiments, the training set contains 7, 532 news, including 3, 749 fake news and 3, 783 non-fake news; the test sets contains 1, 996 posts.

Baseline Methods. To comprehensively evaluate the proposed method, we consider both **unimodal** and **multimodal** fake news detection methods in the comparison:

U1: CAR [7], which combines RNN with attention mechanism to capture important textual features to detect text-only fake news;

U2: VS [16], which explores visual and statistical features of image content to detect fake news;

M1 RA [15], which utilizes an LSTM network and attention mechanism to model text and social context. This work focuses on fake news detection with text and image, so we remove social information from multimodal baselines for a fair comparison;

M2: EANN [31], which consists of two related tasks: event discrimination and fake news detection. To detect fake news, we use the multimodal feature extractor and the fake news detector. Meanwhile, the configure of EANN is set as the official implementation;

M3: MVAE [17], which uses a variational autoencoder with a binary classifier to model representations between text and images for fake news detection;

M4: MKEMN [37], which regards text, image and retrieved knowledge embeddings as stacked channels and makes a fusion via a convolutional operation;

M5: SAFE [39], which uses a pre-trained image to text model to transform the image into text, and then measures the similarity to detect fake news;

M6: MVNN [33], which incorporates textual semantic features, visual tampering features and similarity of textual and visual information computed by the cosine similarity in fake news detection.

Implementation Details. In the textual encoder, we set the length of the input text to at most 200 words. Then, we adopt a pre-trained BERT model [11] to encode each text into embedding with 256 dimensions. In the visual encoder, the size of the input image is 224×224, and we use the features from ResNet-34 [13] pre-trained on ImageNet dataset as the visual embedding. In the cross-modal alignment module, we implement the modal-specific MLPs using three fully-connected layers with 64 hidden units in each layer. When estimating the cross-modal ambiguity, the modal-specific variational encoders are implemented by fully-connected layers. In the cross-modal fusion module, the interaction map c between two modalities is flattened into a vector with dimension 64×64. The margin d in Equation 1 is set to 0.2 and the hyper-parameter β in Equation 16 is set to 0.5 in all experiments. We keep the same data splits when comparing among all the methods. If a news article contains multiple images, we randomly select one image. In the ablation study, we retrained each variant of the proposed method by only removing the corresponding component. We use the batch size of 64 and train the model using Adam with an initial learning rate of 10^{-4} for 50 epochs with early stopping. Also the early stopping is used to avoid overfitting. ReLU is used as the default activation function unless otherwise specified. In order to get optimal parameters for our model, we use Adam as the optimizer.

5.2 Overall Performance

Table 1 presents the performance comparison between CAFE and the other six methods. As shown in the table, CAFE outperforms all the compared methods on every dataset in terms of *Acc* and $F1$. Specifically, CAFE achieves the highest accuracy of 80.6% and 84.0% on two real-world datasets, respectively. We also draw the following observations:

Among unimodal methods, text-based method performs better in accuracy and recall, while image-based method performs better in precision. This indicates that text and image can provide different discriminability in fake news detection and aggregating these unimodal information can potentially help to improve fake news detection.

The multimodal methods outperform the unimodal methods in all datasets, confirming the advantage of leveraging multimodal information in fake news detection. Among the multimodal methods, RA and EANN perform worst because both methods learn unimodality features separately and ignore the semantic gap across modalities resulting in different embedding spaces and less effective fusion. The performance of MKEMN varies significantly among different datasets. MKEMN regards different modalities as stacked channels without considering the heterogeneity issue, bonding its

Table 1: Performance comparison between CAFE and the two unimodal and six multi-modal baseline methods.

	Method	Acc	Rumor			Non Rumor		
			P	R	F_1	P	R	F_1
Twitter	CAR	0.637	0.574	0.690	0.682	0.724	0.602	0.617
	VS	0.617	0.635	0.644	0.639	0.639	0.630	0.634
	RA	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.810	0.498	0.617	0.584	0.759	0.660
	MAVE	0.745	0.801	0.719	0.758	0.689	0.777	0.730
	MKEMN	0.715	0.814	0.756	0.708	0.634	0.774	0.660
	SAFE	0.762	**0.831**	0.724	0.774	0.695	0.811	0.748
	MCNN	0.784	0.778	0.781	0.779	0.790	0.787	0.788
	CAFE	**0.806**	0.807	**0.799**	**0.803**	**0.805**	**0.813**	**0.809**
Weibo	CAR	0.745	0.705	0.765	0.750	0.756	0.725	0.740
	VS	0.726	0.732	0.712	0.722	0.720	0.74	0.73
	RA	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.795	0.806	0.795	0.800	0.752	0.793	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MKEMN	0.814	0.823	0.799	0.812	0.723	0.819	0.798
	SAFE	0.816	0.818	0.815	0.817	0.816	0.818	0.817
	MCNN	0.823	**0.858**	0.801	0.828	0.787	0.848	0.816
	CAFE	**0.840**	0.855	**0.830**	**0.842**	**0.825**	**0.851**	**0.837**

performance on the data distribution. MVNN achieves the best performance among all baselines due to the adoption of cross modality correlation captured by the cosine similarity between textual and visual features. However, its correlation information does not focus on news with strong cross-modal ambiguity, and fails to explicitly leverage the cross-modal correlation, causing inferior performance.

CAFE outperforms all these state-of-the-art methods in all three datasets mainly due to the following reasons: 1) the auxiliary correlation learning task in CAFE can produce discriminative unimodal features, ensure well aligned semantic space across different modalities and adaptively utilize these aligned features to assistant the main task; 2) the cross-modality ambiguity learning module can accurately estimate the ambiguity between different modalities, which can weigh the importance between unimodal features and cross-modal features given different levels of ambiguity; 3) the main fake news detection task in CAFE can adaptively aggregate complementary unimodal representations and cross-modal correlations to perform accurate classification, i.e., alleviating the noises introduced by cross-modal information when unimodal detection agrees with each other and incorporating discriminative cross-modal features to assist when unimodal detection fails.

5.3 Ablation Study

To further investigate the effectiveness of each component in CAFE, we conduct three sets of experiments.

5.3.1 Effectiveness of Each Component. The first study analyzes the impact of each proposed component in CAFE for fake news detection. More specifically, the compared variants of CAFE are implemented as follows: CAFE w/o R: We remove the cross-modal alignment module and use unimodal embeddings to learn the correlations; CAFE w/o A: We remove the cross-modal ambiguity learning module and treat the cross-modal correlations and unimodal representations as equally important; CAFE w/o C: We remove the cross-modal fusion module and replace it with simply concatenating m^t and m^v.

Table 2: Ablation study on the architecture design of CAFE on two datasets.

Method	Data	*Acc*	*Pre*	*Rec*	*F1*
CAFE w/o R	Twitter	0.791	0.834	0.744	0.787
	Weibo	0.830	0.875	0.801	0.837
CAFE w/o A	Twitter	0.786	0.767	0.790	0.779
	Weibo	0.829	0.831	0.826	0.828
CAFE w/o C	Twitter	0.806	0.807	0.799	0.803
	Weibo	0.827	0.863	0.805	0.833
CAFE	Twitter	0.806	0.807	0.799	0.803
	Weibo	0.840	0.855	0.830	0.842

From the results shown in Table 2, we have the following observations: 1) CAFE w/o A yields poor performance, proving that unimodal features and cross-modal features are not equally important and cross-modal ambiguity learning is essential in cross-modal fake news detection; 2) CAFE w/o R yields poor performance, proving that aligning features across different modalities can also help to improve the performance significantly; and 3) Compared to CAFE, we observe CAFE w/o C yields weaker performance indicating that cross-modal features learned by the proposed cross-modal fusion module are more effective than simply concatenating unimodal features as cross-modal features.

5.3.2 Cross-modal Ambiguity Learning Analysis. In this paper, we formulate the key problem to multimodal fake news detection — cross-modal ambiguity learning and present a computation method based on the KL divergence. Therefore, the second set of experiments is to evaluate different alternative methods for cross-modal ambiguity learning. Following the common assumption that the unimodal detectors are linear to the unimodal features, we compute

Table 3: Performance comparison of different distance measurement methods in ambiguity learning methods.

Method	Data	*Acc*	*Pre*	*Rec*	*F1*
CAFE-COS	Twitter	0.793	0.823	0.753	0.787
	Weibo	0.837	0.848	0.829	0.838
CAFE-DIS	Twitter	0.784	0.801	0.753	0.776
	Weibo	0.834	0.843	0.828	0.835
CAFE-KL	Twitter	0.806	0.807	0.799	0.803
	Weibo	0.840	0.855	0.830	0.842

Table 4: Performance comparison between different cross-modal fusion methods.

Method	Data	*Acc*	*Pre*	*Rec*	*F1*
CAFE-CAT	Twitter	0.789	0.801	0.756	0.778
	Weibo	0.828	0.863	0.805	0.833
CAFE-CNN	Twitter	0.794	0.801	0.763	0.782
	Weibo	0.832	0.843	0.825	0.834
CAFE	Twitter	0.806	0.807	0.799	0.803
	Weibo	0.840	0.855	0.830	0.842

the distance between unimodal features to approximate their ambiguity. Then we produce two CAFE variants, both of them directly obtain unimodal features (m^t and m^v) by the modal-specific encoders but with different unimodal distance measurement methods, where **CAFE-COS** and **CAFE-DIS** represent cosine distance and Euclidean distance as the distance metrics, respectively.

Table 3 shows the performance of different distance measurement methods for ambiguity learning on fake news detection. We can observe that: all three variants of CAFE present good performances, demonstrating that ambiguity learning is important for multi-modal fake news detection. Specifically, CAFE-KL performs better than CAFE-COS and CAFE-DIS. The reason is that CAFE-KL performs direct regression over the space of discretely sampled output distributions via the KL divergence, while CAFE-COS and CAFE-DIS compute ambiguity using fixed unimodal representation without characterizing the uncertainty of the distributions.

5.3.3 Cross-modal Fusion. The third group of experiments is to evaluate the performance of different cross-modal fusion strategies. Following previous works of cross-modal fusion [22, 26, 30], we propose two CAFE variants: CAFE-CAT, which concatenates the aligned unimodal representations extracted from alignment module; and CAFE-CNN, which adopts a convolutional neural network to slide through the aligned unimodal representations for cross-modal fusion. As shown in Table 4, we can observe that: the performance degradation of CAFE-CAT indicates that concatenating unimodality without modeling cross-modal interactions is insufficient for multimodal representation. CAFE-CNN tends to obtain locally confined semantic interactions due to the limited size of the convolution kernel, while CAFE is able to explore such interactions more globally, and thus achieves better performance.

5.4 Quantitative analysis

Discriminative features, i.e., features with strong similarity among intra-class news and large difference among inter-class news, are essential to classification problems. In this case study, we demonstrate the capability of the proposed method in terms of learning the cross-modal correlation to support accurate fake news detection. Specifically, we use heatmaps to visualize the correlation patterns between inter-class and intra-class news. We select 20 news, including 10 pieces of fake news and 10 pieces of real news, and then extract the corresponding correlations from CAFE, CAFE w/o R and CAFE w/o A, respectively.

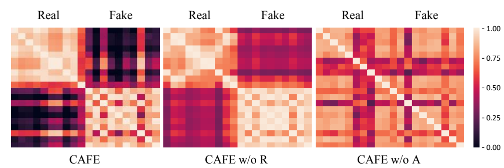


Figure 5: The result of quantitative analysis. CAFE presents clear inter-class difference and intra-class similarity, while CAFE w/o A and CAFE w/o R yield poor capability to learn inter-class difference.

Figure 5 compares the discriminative capability of the cross-modal features extracted from the aforementioned alternatives. We can observe that the features learned by CAFE present clear inter-class difference and intra-class similarity. Compared to CAFE, CAFE w/o A and CAFE w/o R exhibit significant performance degradation. Note that, with the support of deep neural networks, CAFE w/o A is able to learn the semantic correlation between different content modalities, which however may not be directly beneficial to fake news classification as demonstrated by the blurred boundary between real and fake news. On the other hand, with the support of the proposed cross-modal ambiguity learning module, CAFE can learn the discriminative cross-modal features which are explicitly beneficial to the cases when unimodalities present strong ambiguity, and thus improve multimodal fake news detection accuracy.

6 CONCLUSION

Cross-modal ambiguity is crucial in multimodal fake news detection. In this paper, we first formulate the cross-modal ambiguity learning task. Then, we propose CAFE, a cross-modal ambiguity learning based method for multimodal fake news detection. Different from prior works, CAFE is capable of adaptively aggregating discriminative cross-modal correlation features and unimodal features based on the inherent cross-modal ambiguity, addressing the misclassifications caused by the disagreement between different modalities. Experimental studies on two widely used datasets (Twitter and Weibo) demonstrate that CAFE outperforms prior arts in multimodal fake news detection, with accuracy improvements of 2.2-18.9% and 1.7-11.4%, respectively.

ACKNOWLEDGMENTS

We are grateful to Liang Hu for his valuable comments and suggestions in the early days of this work. This work is supported by the National Natural Science Foundation of China (NSFC) under the Grants nos. 61932007, 61902075 and 61572459.

REFERENCES

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828.

[3] Gaurav Bhatt, Aman Sharma, Shivam Sharma, Ankush Nagpal, Balasubramanian Raman, and Ankush Mittal. 2018. Combining Neural, Statistical and External Features for Fake News Stance Identification. In *The World Wide Web Conference 2018*. ACM, 1353–1357.

[4] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 549–556.

[5] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor Detection on Social Media with Bi-Directional Graph Convolutional Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (2020), 549–556.

[6] Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. 2018. Detection and Visualization of Misleading Content on Twitter. *International Journal of Multimedia Information Retrieval* 7, 1 (2018), 71–86.

[7] Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. 2018. Call Attention to Rumors: Deep Attention Based Recurrent Neural Networks for Early Rumor Detection. In *Trends and Applications in Knowledge Discovery and Data Mining*. Springer, 40–52.

[8] Yixuan Chen, Liang Hu, and Jie Sui. 2019. Text-Based Fusion Neural Network for Rumor Detection. In *International Conference on Knowledge Science, Engineering and Management*. Springer, 105–109.

[9] Yixuan Chen, Jie Sui, Liang Hu, and Wei Gong. 2019. Attention-Residual Network with CNN for Rumor Detection. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. ACM, 1121–1130.

[10] Leon Derczynski and Arkaitz Zubiaga. 2020. Detection and Resolution of Rumors and Misinformation with NLP. In *Proceedings of the 28th International Conference on Computational Linguistics: Tutorial Abstracts*. 22–26.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

[12] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking Sandy: Characterizing and Identifying Fake images on Twitter during Hurricane Sandy. In *Proceedings of the 22nd International Conference on World Wide Web*. ACM, 729–736.

[13] 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.

[14] Zhenyu He, Ce Li, Fan Zhou, and Yi Yang. 2021. Rumor Detection on Social Media with Event Augmentations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA, 2020–2024.

[15] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 795–816.

[16] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2017. Novel Visual and Statistical Image Features for Microblogs News Verification. *IEEE Transactions on Multimedia* 19, 3 (2017), 598–608.

[17] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal Variational Autoencoder for Fake News Detection. In *The World Wide Web Conference*. ACM, 2915–2921.

[18] Ling Min Serena Khoo, Hai Leong Chieu, Zhong Qian, and Jing Jiang. 2020. Interpretable Rumor Detection in Microblogs by Attending to User Interactions. *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (2020), 8783–8790.

[19] An Lao, Chongyang Shi, and Yayi Yang. 2021. Rumor Detection with Field of Linear and Non-Linear Propagation. In *Proceedings of the Web Conference 2021*. 3178–3187.

[20] Yali Li, Shengjin Wang, Qi Tian, and Xiaoqing Ding. 2015. A survey of recent advances in visual feature detection. *Neurocomputing* 149 (2015), 736–751.

[21] Tanushree Mitra, Graham P. Wright, and Eric Gilbert. 2017. A Parsimonious Language Model of Social Media Credibility Across Disparate Events. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. 126–145.

[22] Kellin Pelrine, Jacob Danovitch, and Reihaneh Rabbany. 2021. The Surprising Performance of Simple Baselines for Misinformation Detection. In *Proceedings of the Web Conference 2021*. 3432–3441.

[23] Nikhil Pinnaparaju, Manish Gupta, and Vasudeva Varma. 2021. T3N: Harnessing Text and Temporal Tree Network for Rumor Detection on Twitter. In *Advances in Knowledge Discovery and Data Mining*. 686–700.

[24] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 518–527.

[25] Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. 2018. Neural User Response Generator: Fake News Detection with Collective User Intelligence. In *Twenty-Seventh International Joint Conference on Artificial Intelligence IJCAI-18*.

[26] 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.

[27] Shivangi Singhal, Rajiv Ratn Shah, Tanmoy Chakraborty, Ponnurangam Kumaraguru, and Shin'ichi Satoh. 2019. SpotFake: A Multi-modal Framework for Fake News Detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*. IEEE, 39–47.

[28] Sebastian Tschiatschek, Adish Singla, Manuel Gomez-Rodriguez, Arpit Merchant, and Andreas Krause. 2018. Fake News Detection in Social Networks via Crowd Signals. In *The World Wide Web Conference 2018*. ACM, 517–524.

[29] SS Vallender. 1974. Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability & Its Applications* 18, 4 (1974), 784–786.

[30] Sunny Verma, Chen Wang, Liming Zhu, and Wei Liu. 2019. Deepcu: Integrating both common and unique latent information for multimodal sentiment analysis. In *International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization.

[31] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 849–857.

[32] Rui Xia, Kaizhou Xuan, and Jianfei Yu. 2020. A State-independent and Time-evolving Network for Early Rumor Detection in Social Media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 9042–9051.

[33] Junxiao Xue, Yabo Wang, Yichen Tian, Yafei Li, Lei Shi, and Lin Wei. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management* 58, 5 (2021), 102610.

[34] Fan Yang, Shiva K. Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia (Ben) Hu. 2019. XFake: Explainable Fake News Detector with Visualizations. In *The World Wide Web Conference 2019 (WWW '19)*. 3600–3604.

[35] Xiaoyu Yang, Yuefei Lyu, Tian Tian, Yifei Liu, Yudong Liu, and Xi Zhang. 2020. Rumor Detection on Social Media with Graph Structured Adversarial Learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 1417–1423.

[36] Xiaoyu Yang, Yuefei Lyu, Tian Tian, Yifei Liu, Yudong Liu, and Xi Zhang. 2020. Rumor Detection on Social Media with Graph Structured Adversarial Learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. 1417–1423.

[37] Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2019. Multi-modal Knowledge-aware Event Memory Network for Social Media Rumor Detection. In *Proceedings of the 27th ACM International Conference on Multimedia*. 1942–1951.

[38] Xueyao Zhang, Juan Cao, Xirong Li, Qiang Sheng, Lei Zhong, and Kai Shu. 2021. Mining Dual Emotion for Fake News Detection. In *Proceedings of the Web Conference 2021*. 3465–3476.

[39] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multi-modal Fake News Detection. *Advances in Knowledge Discovery and Data Mining* 12085 (2020), 354.

[40] Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)* 51, 2 (2018), 32.