See How You Read? Multi-Reading Habits Fusion Reasoning for Multi-Modal Fake News Detection

Lianwei Wu[1,2,3], **Pusheng Liu**[1], **Yanning Zhang**[1*]

[1]National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, China
[2]Research & Development Institute of Northwestern Polytechnical University in Shenzhen, China
[3]Chongqing Science and Technology Innovation Center of Northwestern Polytechnical University, China
wlw@nwpu.edu.cn, lps@mail.nwpu.edu.cn, ynzhang@nwpu.edu.cn

Abstract

The existing approaches based on different neural networks automatically capture and fuse the multimodal semantics of news, which have achieved great success for fake news detection. However, they still suffer from the limitations of both shallow fusion of multimodal features and less attention to the inconsistency between different modalities. To overcome them, we propose multi-reading habits fusion reasoning networks (MRHFR) for multi-modal fake news detection. In MRHFR, inspired by people's different reading habits for multimodal news, we summarize three basic cognitive reading habits and put forward cognition-aware fusion layer to learn the dependencies between multimodal features of news, so as to deepen their semantic-level integration. To explore the inconsistency of different modalities of news, we develop coherence constraint reasoning layer from two perspectives, which first measures the semantic consistency between the comments and different modal features of the news, and then probes the semantic deviation caused by unimodal features to the multimodal news content through constraint strategy. Experiments on two public datasets not only demonstrate that MRHFR not only achieves the excellent performance but also provides a new paradigm for capturing inconsistencies between multi-modal news.

Introduction

The rapid development of social media has not only brought great convenience to knowledge sharing and communication, but also caused the widespread spread of massive fake news, which has posed realistic threats to politics (Osmundsen et al. 2021), public health (Diseases 2020), etc. Especially, multimodal fake news is inherently more infectious, spreading deeper and farther, and causing more damage than unimodal (textual) fake news (Nielsen and McConville 2022). Therefore, under such severe scenarios, how to automatically detecting multimodal fake news already become a crucial issue.

The existing approaches to fake news detection could be roughly divided into two categories based on news content, i.e., unimodal-based and multimodal-based. The unimodal fake news detection task has developed in two stages: 1) **Feature engineering stage** focuses on extracting manually the surface features (e.g., the number of repetitions of punctuation) from text content (Castillo, Mendoza, and Poblete 2011) and collects simple platform features from meta data (Wu et al. 2016); and 2) **Automatic detection stage** strives to construct reasonable neural networks to learn semantic (Hu et al. 2021; Wu et al. 2021a), emotional (Zhang et al. 2021), stance-based (Wu et al. 2019; Xie et al. 2021), stylistic (Wu et al. 2021b) features around news text content, and capture comment-based (Shu et al. 2019), and propagation-based (Shu et al. 2020) features around meta data, which has achieved satisfactory performance and gained considerable development. At present, multimodal fake news detection has received more and more attention with the emergence of forged images or fake news with text and images, which is primarily absorbed in two perspectives: 1) **Consistent alignment** endeavours to promote the associations of multimodal features through building entity alignment (Li et al. 2021), relationship alignment (Zhou, Wu, and Zafarani 2020), and semantic alignment (Chen et al. 2022) for detection; and 2) **Interaction fusion** first extracts textual and visual features, and then integrates the two types of features through simple early fusion or late fusion strategies to detect fake news (Wu et al. 2021c; Dhawan et al. 2022).

However, although these approaches have advanced the progress of multimodal fake news detection, which still possess several dilemmas: 1) **In terms of feature fusion**, existing methods generally employ superficial fusion strategies such as concatenation, addition, or simply neural networks to integrate the features of different modalities, which is difficult to capture the internal dependencies between them; and 2) **In terms of interaction alignment**, the great majority of methods emphasize mainly on capturing the similarity semantics between different modalities through alignment mechanism, but ignore the acquirement of extensive inconsistent semantics. These inconsistent semantics include abundant credibility-indicator features, which may be the key to improve the performance of fake news detection task.

To overcome the challenges, we propose **M**ulti-**R**eading **H**abits **F**usion **R**easoning networks (MRHFR) by strengthening both deep fusion and coherence reasoning. In detail, we know that there is generally massive interaction when people read a multimodal news, which is sufficient for the deep fusion of textual and visual information. Inspired by people's reading habits facing multimodal news, we summarize three basic reading habits and construct cognition-aware fusion layer (CFU) to model them so as to learn the dependencies between multi-

*Corresponding author.

modal features of the news. Particularly, three reading habits we summarized are: (a) **Read-Text&Glimpse-Image**: When faced with a news with funny text but plain images, most of audiences prefer to peruse text and glance at images roughly; (b) **Glimpse-Text&Read-Image**: When faced with a news with eye-catching images and plain text, most of audiences tend to read images carefully and glance at text roughly; and (c) **Read-Text&Read-Image**: There is also a considerable part of audiences who has always been earnest, they like to read both text and images carefully. To model these three habits, in our CFU layer, we use the initial embeddings of unimodal content as a glance behavior and the encoding of unimodal information as a careful reading behavior, and then design cognition-aware interaction block to enhance the interaction of behaviors in each reading habit, so as to learn the dependencies between multimodal features of news, so as to deepen their semantic-level fusion. To capture inconsistent semantics of multimodal news, we build coherence constraint reasoning layer (CCR) from two perspectives, which first measures the inconsistency of external comments on different modal content of news, and then investigates the semantic deviation caused by unimodal features to the multimodal features of news content by constraint strategy. Extensive experiments confirm the superiority of MRHFR. Its contributions are summarized as follows:

- A new paradigm of cognition-aware fusion inspired by audiences' frequently-used reading habits for multimodal news is proposed, which establishes deep fusion between multimodal features and explores the inconsistency between them for fake news detection.
- Explored coherence constraint reasoning layer could not only infer the coherence between comments and news, but also evaluate the semantic deviation between unimodal content and multimodal content of news.
- We empirically reveal that MRHFR significantly outperforms several state-of-the-art baselines on two competitive datasets.

Related Work

According to the different modal forms of news content, we divide fake news detection into unimodal and multimodal.

Unimodal Fake News Detection

Most of the existing methods are based on unimodal detection, which are mainly separated into text-based, vision-based, and metadata-based. **Text-based.** Text-based fake news detection has obtained sufficient development. The early studies concentrate on statistical features around the text content in artificial ways, such as the number of punctuation (Parikh and Atrey 2018), the proportion of negative words (Guo et al. 2019), etc. However, these artificial ways are time-consuming and labor-intensive, which are difficult to meet the needs of massive data. To address it, automatic fake news detection emerges, which relies on deep neural networks based on CNN (Verma et al. 2021), RNN (Shu et al. 2019), attention (Wu et al. 2020), and graph (Hu et al. 2021) architectures to gain semantic, emotional, stylistic, and stance-based features to identify fake news. **Vision-based.** Besides text content, several works also consider image information in news (Qi et al. 2019; Abdali et al. 2021). The methods generally capture spatial-domain features by pre-training models such as VGG-19 and frequently-domain features by CNN-based networks to identify fake or forged images. **Metadata-based.** The features in fake news include not only content features, but also rich social context features, i.e., metadata. The metadata-based methods strive to capture comment-based (Choi and Ko 2021), user profile-based (Dou et al. 2021), platform-based (Qi et al. 2019), and propagation structure-based (Shu et al. 2020) features for detection. Specifically, comment-based methods design interactive mechanisms to earn valuable features between comments and news (Setty and Rekve 2020). User profile-based methods are suitable for fake news with abundant users (Dou et al. 2021). Social platform-based methods often appear in cross-platform fake news detection task (Qi et al. 2019). Propagation structure-based methods are always time-sensitive and apply to early detection (Shu et al. 2020).

Multimodal Fake News Detection

The majority of studies for multimodal fake news detection are absorbed in two aspects: **Consistent Alignment.** The mismatch of different modal information in news is a common false type, which includes image-text divergence, video-descriptive text disparity, etc. To discover these vital credibility-indicator clues, current studies generally focus on similarity comparison (Zhou, Wu, and Zafarani 2020), semantic matching (Xue et al. 2021), entity alignment (Li et al. 2021), and other alignment strategies (Qi et al. 2021) for detection. Nevertheless, such consistent methods are difficult to explore inconsistent information between multimodal features. **Interaction Fusion.** In the task, interaction fusion mechanism could be roughly divided into two categories: **Early Fusion** (Singhal et al. 2019; Boulahia et al. 2021; Xue et al. 2021) also known as feature-level fusion, refers to the information fusion of different modalities in the early stage of the model by adopting concatenation or addition operations. After the fusion, the features equally output to the downstream for learning. **Late Fusion** (Meel and Vishwakarma 2021b,a; Singhal et al. 2021), also known as decision-level fusion, depends on the results obtained by each modality data individually and is fused at the final stage of task learning, which usually applies summation, maximum, average, or dot product operations as fusion strategies. Nevertheless, they have the following defects: 1) The level of feature fusion is shallow; 2) They lack correlation and interaction between different types of features. To this end, we construct multi-reading habits fusion reasoning networks from both deep feature interaction fusion and the capture of inconsistency among multimodal features in news for fake news detection.

The Proposed Model

Our MRHFR aims to learn deeply multimodal fusion representations of news and explore the inconsistent semantics among different modalities of news for fake news detection. As shown in Figure 1, MRHFR consists of four major layers:

Feature Representations

The inputs of MRHFR are multimodal news (i.e., text and image content) and its series of comment content. For multimodal news, the text content is represented as a text sequence with l_T tokens $\mathbf{T} = \{t_1, ..., t_{l_T}\}$, $\mathbf{T} \in \mathbb{R}^{l_T \times d}$, where each token $t_i \in \mathbb{R}^d$ is a

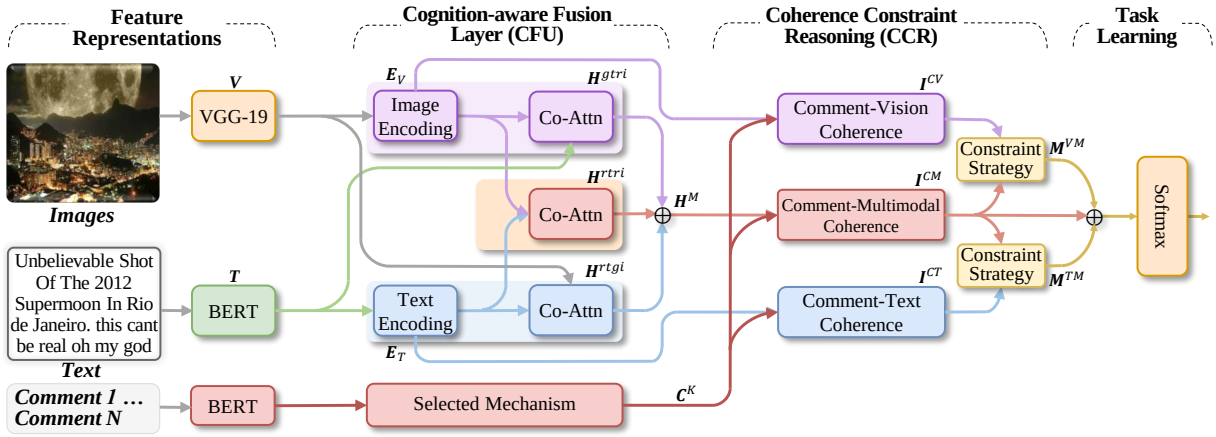


Figure 1: The architecture of MRHFR. CFU enhances the deep fusion between multimodal news features by considering three reading habits for multimodal news, i.e., Glimpse-Text&Read-Image $\mathbf{H}^{gtri}$, Read-Text&Read-Image $\mathbf{H}^{rtri}$, and Read-Text&Glimpse-Image $\mathbf{H}^{rtgi}$. CCR measures the inconsistency from two perspectives, i.e., learning the inconsistency between comments and news by coherence blocks, and exploring the semantic deviation caused by unimodal features to multimodal news content by constrain strategy.

d-dimensional vector learned from pre-trained BERT (Devlin et al. 2018). Then, we utilize pre-trained VGG-19 (Simonyan and Zisserman 2014) to learn visual features of image content from spatial domain. We gain the outputs V_g of the second last layer of VGG-19 on ImageNet dataset and pass them into a fully-connected layer $\sigma(\cdot)$ to convert to the final dimension with length l_v. Finally, the visual representations $\mathbf{V} \in \mathbb{R}^{l_v \times d}$ of the news are learned as follows:

$$\mathbf{V} = \sigma(\mathbf{W}V_g) \tag{1}$$

Selected Mechanism For comments, we know that there are multiple comments under a news and each comment is generally text sequence. Thus, the representations of each comment are same as these of news text, which are represented by BERT, i.e., $\mathbf{C}_i \in \mathbb{R}^{l_{C_i} \times d}$, where l_{C_i} is the length of the i-th comment. To select the top representative comments, we design selected mechanism to extract top-K comments, which calculates the difference between each comment and other comments in an automated manner. To do this, our selected mechanism optimizes one inter-sequential attention matrix $\mathbf{U} \in \mathbb{R}^{N \times N}$, where N is the number of comments under one news. The entry (m, n) of $\mathbf{U}$ holds differences between comment m and comment n ($1 \leq m, n \leq N$, and $m \neq n$), which could be formalized as:

$$u_m = \varphi(\mathbf{W}_m\mathbf{C}_m + \mathbf{b}_m) \tag{2}$$

$$u_n = \varphi(\mathbf{W}_n\mathbf{C}_n + \mathbf{b}_n) \tag{3}$$

$$\mathbf{U}[m,n] = \frac{u_m \odot u_n}{\sum_{i=1}^{N} \mathbf{exp}(u_i \odot u_n)} \tag{4}$$

where $\varphi(\cdot)$ is the activation function. All $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters, and $\odot$ denotes dot product operator. Thus, we finally select top-K representative comments with high difference $\mathbf{C}^K$.

Cognition-aware Fusion Layer (CFU)

To explore the relationships between text and image content of news, we design cognition-aware fusion layer (CFU) by considering people's reading habits. According to the differences in people's attention to multimodal information, we summarize three reading habits, i.e., Read-Text&Glimpse-Image, Glimpse-Text&Read-Image, and Read-Text&Read-Image, where 'Read' means the behavior of reading carefully and 'Glimpse' refers to a glance behavior. In CFU layer, we adopt the initial embeddings of unimodal information as 'a glance' behavior while its deep-seated encoding as a 'read carefully' behavior. Therefore, CFU first constructs different unimodal encoding blocks, and then co-attention block is designed to model three types of reading interaction of people reading multimodal information.

Text Encoding Block. We utilize self-attention networks as text encoding block to explicitly learn the dependencies between any two tokens and learn the inner structure features of the text sequence:

$$\mathbf{H} = \mathbf{Attention}(\mathbf{Q}_1, \mathbf{K}_1, \mathbf{V}_1) \tag{5}$$

$$= \mathbf{softmax}(\frac{\mathbf{Q}_1\mathbf{K}_1^T}{\sqrt{d_k}})\mathbf{V}_1 \tag{6}$$

where $\mathbf{Q}_1$, $\mathbf{K}_1$, $\mathbf{V}_1$ are query, key, and value matrix, respectively. In our settings, $\mathbf{Q}_1$=$\mathbf{K}_1$=$\mathbf{V}_1$=$\mathbf{T}$, and d_k equals to $d/2$. To widely learn richer context information of text from different perspectives, multi-head attention mechanism projects the query, key, and value m times through different linear projections, and then executes them in parallel. Finally, the processed results are integrated and projected to gain a new representation. Thus, it is formalized as:

$$\mathrm{head}_i = \mathbf{Attention}(\mathbf{Q}_1\mathbf{W}_q, \mathbf{K}_1\mathbf{W}_k, \mathbf{V}_1\mathbf{W}_v) \tag{7}$$

$$\mathbf{E}_T = \mathbf{MultiHead}(\mathbf{Q}_1, \mathbf{K}_1, \mathbf{V}_1)$$

$$= \mathbf{Concat}(\mathrm{head}_1, ..., \mathrm{head}_m)\mathbf{W}_e \tag{8}$$

where all $\mathbf{W} \in \mathbb{R}^{d \times d_e}$ are trainable parameters and d_e is d/m. $\mathbf{E}_T \in \mathbb{R}^{l_T \times d}$ is the encoding of news text.

Image Encoding Block. To capture eye-catching semantics in image content of news, we apply CNN-based focusing

network to extract its frequent features. Specifically, the image $\mathbf{V}$ is first converted from spatial space to frequent space by discrete cosine transform (DCT) (Qi et al. 2019), and outputs 64 vectors $\mathbf{V}_0$, $\mathbf{V}_1$, ..., $\mathbf{V}_{63}$. Then, we feed these vectors to CNN with different window sizes ([1, 1], [1, 3], [3, 3], and [5, 5]), and then concatenate them to obtain eye-catching vectors with different scales $\mathbf{E}_V \in \mathbb{R}^{l_v \times d}$.

Cognition-aware Interaction Block To model interaction behaviors in each reading habit, we build co-attention block (Co-Attn) to learn the dependencies between multimodal information. Specifically, in the case of Read-Text&Glimpse-Image habit, the inputs of Co-Attn are $<\mathbf{E}_T, \mathbf{V}>$:

$$\hat{\mathbf{H}}_T = \mathbf{Norm}(\mathbf{E}_T + \mathbf{softmax}(\frac{\mathbf{V}}{\sqrt{d}}\mathbf{V})) \quad (9)$$

$$\hat{\mathbf{H}}_V = \mathbf{Norm}(\mathbf{V} + \mathbf{softmax}(\frac{\mathbf{V}\mathbf{E}_T}{\sqrt{d}}\mathbf{E}_T)) \quad (10)$$

$$\mathbf{H}_T^{rtgi} = \mathbf{Norm}(\hat{\mathbf{H}}_T + \mathbf{FFN}(\hat{\mathbf{H}}_V)) \quad (11)$$

$$\mathbf{H}^{rtgi} = \mathbf{concat}(\mathbf{H}_T^{rtgi}, \mathbf{H}_V^{rtgi}) \quad (12)$$

where **Norm** and **FFN** are the normalization method and feed forward network as in Vaswani et al. (2017). $\mathbf{H}^{rtgi}$ is the fused semantics of the interaction block aiming at Read-Text&Glimpse-Image. Here, the fused semantics aiming at Glimpse-Text&Read-Image and Read-Text&Read-Image are $\mathbf{H}^{gtri}$ and $\mathbf{H}^{rtri}$.

Finally, we integrate the three reading habits to form the comprehensive fusion representations of multimodal news $\mathbf{H}^M = \mathbf{concat}(\mathbf{H}^{rtgi}, \mathbf{H}^{gtri}, \mathbf{H}^{rtri})$.

Coherence Constraint Reasoning (CCR)

To explore inconsistency of different modalities of news, we design CCR layer from two perspectives, which first explores inconsistency between external comments and multimodal semantics of news by coherence reasoning block, and then proposes association constraint strategy to capture semantic deviation caused by coherence semantics of unimodal to whole multimodal news.

Coherence Reasoning Block Take the coherence alignment between comments and multimodal news (comment-multimodal coherence) as an example, we introduce the block in detail. Considering representative comments $\mathbf{C}^K$ and the fused semantics $\mathbf{H}^M$ in different semantic spaces, we first project them into a d_c-dimensional shared latent space:

$$\mathbf{F}^C = \mathbf{tanh}(\mathbf{W}_c\mathbf{C}^K + \mathbf{b}_c) \quad (13)$$

$$\mathbf{F}^M = \mathbf{tanh}(\mathbf{W}_m\mathbf{H}^M + \mathbf{b}_m) \quad (14)$$

where $\mathbf{F}^C$ and $\mathbf{F}^M$ are comment semantics and fused multimodal semantics in shared spaces, respectively.

Then, we promote coherence alignment between the two types of semantics, which adopt comment semantics $\mathbf{Q}_c = \mathbf{W}_q\mathbf{F}^C$ as query and multimodal semantics $\mathbf{K}_M = \mathbf{W}_k\mathbf{F}^M$ as key. Hence, their coherence representations $\mathbf{I}^{CM}$ could be formulated as:

$$\mathbf{A}_{CM} = \mathbf{softmax}(\mathbf{Q}_c\mathbf{K}_M^\top), \quad \hat{\mathbf{I}}^{CM} = \mathbf{F}^C + \mathbf{A}_{CM}\mathbf{F}^M \quad (15)$$

where $\mathbf{A}_{CM}$ is the query attended mask. Next, a max-pooling operation is conducted to obtain an aggregated vector of comment-multimodal coherence $\mathbf{I}^{CM}$:

$$\mathbf{I}^{CM} = \mathbf{max\text{-}pooling}(\hat{\mathbf{I}}^{CM}) \quad (16)$$

Following Eq. (13-16), the same procedure is applied to capture comment-text coherence $\mathbf{I}^{CT}$ and comment-vision coherence $\mathbf{I}^{CV}$.

Association Constraint Strategy We devise association constraint strategy to measure the semantic deviation between unimodal information and multimodal information of news, which evaluates associations between comment-text coherence (or comment-vision coherence) and comment-multimodal coherence. Take the association of $\mathbf{I}^{CT}$ and $\mathbf{I}^{CM}$ as an example:

$$\mathbf{M}_{i,j}^{TM} = \mathbf{cos}(\mathbf{I}_i^{CT}, \mathbf{I}_j^{CM}) \quad (17)$$

where $\mathbf{M}_{i,j}^{CM} \in \mathbb{R}^{l_{CT} \times l_{CM}}$, l_{CT} and l_{CM} are the length of $\mathbf{I}^{CT}$ and $\mathbf{I}^{CM}$, respectively. All synthesis matrices are stacked to obtain text-multimodal synthesis associations $\mathbf{M}^{TM}$. In this way, we obtain vision-multimodal synthesis associations $\mathbf{M}^{VM}$ between $\mathbf{I}^{CV}$ and $\mathbf{I}^{CM}$.

Subsequently, the two types of synthesis associations are passed to MLP for earning holistic semantic deviation $\mathbf{M}^{all}$.

$$\mathbf{M}^{all} = \mathbf{MLP}(\mathbf{concat}(\mathbf{M}^{TM}, \mathbf{M}^{VM})) \quad (18)$$

Therefore, the support degree of news is expressed as the integration between the coherence of three modalities of news and their holistic semantic deviation:

$$\mathbf{IM} = \mathbf{concat}(\mathbf{I}^{CV}, \mathbf{I}^{CM}, \mathbf{I}^{CT}, \mathbf{M}^{all}) \quad (19)$$

Task Learning

Finally, we classify the authority of the news by deploying a fully-connected block with activation function for a training sample with ground-truth label y:

$$p = \mathbf{softmax}(\mathbf{W}_p\mathbf{IM} + \mathbf{b}_p), \quad \text{loss} = -\sum y\mathbf{log}p \quad (20)$$

Experiments

Datasets and Evaluation Metrics

To validate the superiority of MRHFR, we experiment on two competitive datasets collected from Twitter and Weibo platforms, respectively. Twitter dataset is released for evaluating multimodal task at MediaEval (Khattar et al. 2019), while Weibo dataset is assembled from Jin et al. (2017) for multimodal fake news detection. The tweets in each dataset consist of texts, attached images/videos, and social context. In our work, we also crawl through the comments under each news. We focus more on text and image data, so we filter the tweets with videos and those without texts or images. We divide the same data subset scheme as the benchmark on the two datasets.

Evaluation Metrics. We apply accuracy as a standard metric for multimodal news detection. Owing to two datasets suffering from class imbalance, accuracy alone is difficult to achieve sufficient fairness. Therefore, in our experiments, we also expand precision, recall, and F1-measure as complementary metrics.

Experimental Settings

The length l_T of text sequence of news on Twitter and on Weibo is 30 and 160, respectively. The length l_v of visual representations and the length l_{C_i} of each comment are set to 1024

Dataset	Methods	Accuracy	Fake News			True News		
			Precision	Recall	F1	Precision	Recall	F1
Twitter	SVM-TS	0.529	0.488	0.497	0.496	0.565	0.556	0.561
	CNN	0.549	0.508	0.597	0.549	0.598	0.509	0.550
	GRU	0.634	0.581	0.812	0.677	0.758	0.502	0.604
	TextGCN	0.703	0.808	0.365	0.503	0.680	0.939	0.779
	SAFE	0.766	0.777	0.795	0.786	0.752	0.731	0.742
	Att-RNN	0.664	0.749	0.615	0.676	0.589	0.728	0.651
	EANN	0.648	0.81	0.498	0.617	0.584	0.759	0.66
	MVAE	0.745	0.801	0.719	0.758	0.689	0.777	0.73
	MCAN	0.809	0.889	0.765	0.822	0.732	0.871	0.795
	HMCAN	0.897	0.971	0.801	0.878	0.853	0.979	0.912
	Ours	**0.921**	**0.976**	**0.828**	**0.896**	**0.876**	**0.981**	**0.926**
Weibo	SVM-TS	0.640	0.741	0.573	0.646	0.651	0.798	0.711
	CNN	0.740	0.736	0.756	0.744	0.747	0.723	0.735
	GRU	0.702	0.671	0.794	0.727	0.747	0.609	0.671
	TextGCN	0.787	0.975	0.573	0.727	0.712	0.985	0.827
	SAFE	0.763	0.833	0.659	0.736	0.717	0.868	0.785
	Att-RNN	0.772	0.854	0.656	0.742	0.720	0.889	0.795
	EANN	0.782	0.827	0.697	0.756	0.752	0.863	0.804
	MVAE	0.824	0.854	0.769	0.809	0.802	0.875	0.837
	MCAN	0.899	0.913	0.889	0.901	0.884	0.909	0.897
	HMCAN	0.885	0.920	0.845	0.881	0.856	0.926	0.890
	Ours	**0.907**	**0.939**	**0.869**	**0.903**	**0.879**	**0.931**	**0.904**

Table 1: Results of comparison different baselines and our proposed MRHFR on the two datasets.

and 100. The embedding size d of text is set as 768. The K of top-K comments is 5. When training on Twitter dataset due to overfitting, the parameters of VGG-19 and BERT are frozen. In self-attention networks, attention heads and blocks are set to 6 and 4, respectively, and the dropout of multi-head attention is 0.5. In addition, the model is trained for 120 epochs with a learning rate of 0.001, and the mini-batch size is 256. We recommend to achieve it in MindSpore framework.

Performance Comparison

Comparative Baselines We compare MRHFR with several state-of-the-art baselines, including unimodal and multimodal:

Unimodal Models: SVM-TS (Ma et al. 2015) adopts specific rules around textual features and linear SVM classifier to detect fake news. **CNN** (Yu et al. 2017) applies convolutional neural networks to capture the dependencies of text sequence for detection. **GRU** (Ma et al. 2016) is used to earn long-term context of news text to identify fake news. **TextGCN** (Yao, Mao, and Luo 2019) relies on the graph convolutional network to learn word-level and document-level representations for detection.

Multimodal Models: Att-RNN (Jin et al. 2017) is RNN with attention mechanism, which integrates textual and visual features for rumor detection. **EANN** (Wang et al. 2018) relies on adversarial networks to derive event-invariant features for helping fake news detection. In our experiments, we remove the event discriminator for a fair comparison. **MVAE** (Khattar et al. 2019) captures shared features between textual and visual information by a variational autoencoder. **SAFE** (Zhou, Wu, and Zafarani 2020) extracts textual and visual features of news as well as their relationships by a similarity-aware multimodal model. **MCAN** (Wu et al. 2021c) is multimodal co-attention networks to extract features from textual and visual information for fake news detection. **HMCAN** (Qian et al. 2021) designs hierarchical multimodal attention networks to learn both inter-/intra-modality relationships between textual and visual information for detection.

Overall Performance

The results are shown in Table 1, we observe that:

- Neural network methods (e.g., CNN, GRU) perform better performance than SVM-TS relying on hand-crafted features, which confirms the superiority of automatic feature ways and deep learning. TextGCN learning word-level and document-level embeddings achieves more excellent performance than CNN and GRU, which illustrates the effectiveness of integrating different level representations. Att-RNN and EANN also obtain better performance than CNN and GRU, which shows that applying multimodal information is beneficial to detection.
- HCAN and HMCAN designing co-attention networks display superior performance than MVAE and EANN, which indicates the effectiveness of capturing coherence semantics between multimodal features.
- Our MRHFR outperforms consistently all state-of-the-art baselines on the two datasets, showing from 0.2% to 2.7% improvements. We analyze two reasons: 1) Modeling multi-reading habits mechanism to fuse fake news are more effective than simple interaction between multimodal features, and 1) Exploring the inconsistency from two perspectives of both comments-to-multimodal news and internal semantic deviation among multimodal news could capture more credibility-indicator features for fake news detection.

Ablation Analysis

Effectiveness of Each Component To investigate the effectiveness of each component in MRHFR, we ablate our model into

Dataset	Methods	Accuracy	Fake News			True News		
			Precision	Recall	F1	Precision	Recall	F1
Twitter	-Text	0.741	0.801	0.621	0.700	0.701	0.810	0.752
	-Image	0.865	0.920	0.761	0.833	0.816	0.933	0.871
	-Comments	0.884	0.935	0.791	0.857	0.823	0.941	0.878
	-CFU	0.853	0.912	0.750	0.823	0.801	0.918	0.856
	-CCR	0.871	0.923	0.767	0.838	0.816	0.921	0.865
	-Constraint	0.892	0.940	0.795	0.861	0.832	0.950	0.887
	MRHFR	0.921	0.976	0.828	0.896	0.876	0.981	0.926
Weibo	-Text	0.705	0.798	0.681	0.735	0.672	0.790	0.726
	-Image	0.846	0.890	0.824	0.856	0.828	0.886	0.856
	-Comments	0.862	0.895	0.831	0.862	0.835	0.894	0.863
	-CFU	0.831	0.878	0.810	0.843	0.813	0.872	0.841
	-CCR	0.852	0.881	0.814	0.846	0.821	0.882	0.850
	-Constraint	0.874	0.903	0.837	0.869	0.842	0.903	0.871
	MRHFR	0.907	0.939	0.869	0.903	0.879	0.931	0.904

Table 2: Ablation analysis of our proposed model on Twitter and Weibo datasets.

different layers. We employ **-Text**, **-Image**, **-Comments**, **-CFU**, **-CCR**, **-Constraint** to respectively refer to the removal of the following components: the BERT module aiming at news text, the VGG-19 module aiming at news image, cognition-aware fusion layer, coherence constraint reasoning layer, and association constraint strategy. As shown in Table 2, we observe that:

- The removal of different layers suffers from varying degrees of degradation, which embodies the effectiveness of each component. -Text and -Image obtain weaker performance than our MRHFR, which confirms that only utilizing unimodal information is not conducive to detection.
- MRHFR without CFU layer is subjected to a significant reduction, reflecting that modeling cognition behaviors of recognizing fake news to promote the tight fusion between multimodal information contributes to improving performance.
- As a part of CCR layer, association constraint strategy displays at most 2.9% reduction on the two datasets, which is reflective of the significance of measuring the semantic deviation between unimodal features and the multimodal news.

Superiority of CFU Layer To further analyze the superiority of our CFU layer in modeling people's reading habits aiming at multimodal news in a detailed manner, we ablate various reading habits from MRHFR: 1) **-RTGI**, **-GTRI**, **-RTRI** denote respectively the removal of the co-attention blocks aiming at Read-Text&Glimpse-Image $\mathbf{H}^{rtgi}$, Glimpse-Text&Read-Image $\mathbf{H}^{gtri}$, and Read-Text&Read-Image $\mathbf{H}^{rtri}$; 2) **-RTGI(image)** and **-RTRI(image)** represent the image features are removed from $\mathbf{H}^{rtgi}$ and $\mathbf{H}^{rtri}$, respectively; and 3) **-GTRI(text)** and **-RTRI(text)** mean the text features are separated from $\mathbf{H}^{gtri}$ and $\mathbf{H}^{rtri}$. From Figure 2, we observe that:

- Every component in CFU layer plays a significant role in boosting the performance of MRHFR. Especially, the removal of **-H**^{rtgi}, **-H**^{gtri}, and **-H**^{rtri} greatly reduce the model performance, which reflects that the three kinds of reading cognition behaviors we summarized contribute to improving the multi-modal feature fusion of fake news.
- In the last four methods, the removal of each modal feature (text or image) decreases the model performance, showing from 1.1% to 4.9% performance degradation, which not only demonstrates the effectiveness of different modals, but also presents the ability of CFU layer in multimodal fusion.

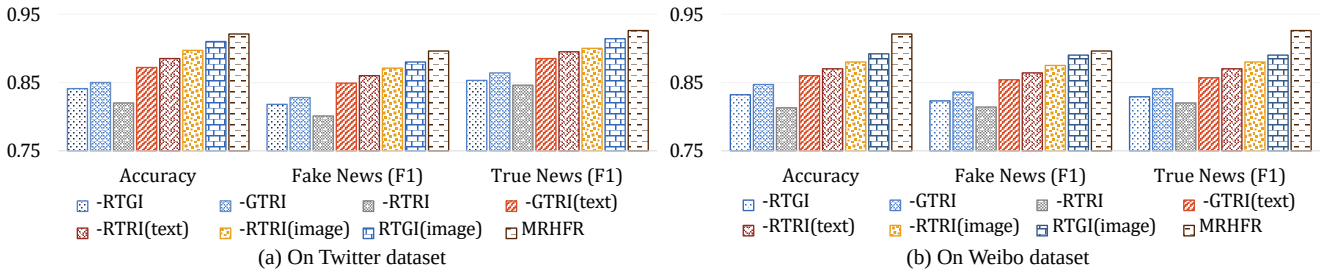
The Analysis of Inconsistency Learning We evaluate alternative approaches for inconsistency learning in MRHFR. Specifically, we respectively replace our association constraint strategy with **KL-divergence**, **Euclidean distance**, **Orthogonality constraints** (Bousmalis et al. 2016), and **RAcoherence** (Zhang et al. 2020). **-Constraint** is introduced in the above subsection. As shown in Figure 3, we observe that:

Compared with -Constraint, all four variants present superior performance on the two datasets, which depicts that capturing the semantic deviation between unimodal content and multimodal content of news is important for multi-modal fake news detection. Furthermore, our MRHFR outperforms the four alternative approaches on the two datasets. The reason is that our strategy could capture the inconsistency representations among the modalities of news by calculating the correlation matrix of two modalities, while the other four methods focus more on the correlations of the two types of features (i.e., single numerical value), lacking effective measurement of the distribution of different features. This demonstrates the superiority of our strategy in measuring the semantic deviation.

Case Study

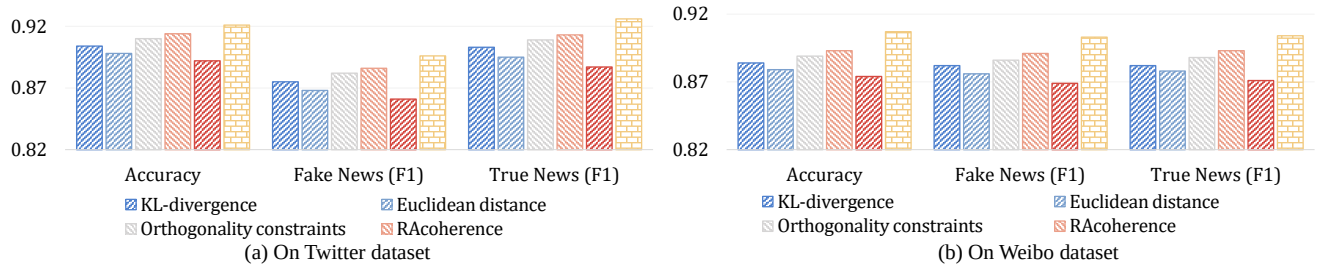
To intuitively describe the adequacy of multimodal feature fusion of MRHFR and the capture of inconsistent features between different modalities, we visualize the outputs of CFU and CCR layers.

Visualization of Features Captured by CFU To vividly evaluate the superiority of our model in multimodal fusion, we compare features learned by traditional co-attention networks and our CFU layer. Taking one specific sample in Twitter as an example, from Figure 4, we observe that: Co-attention networks only pay attention to the superficial shared semantics (like "flood" and "man" in Figure 4(a)) between text content and image content, while in our CFU layer, it not only focuses on ordinary shared semantics ("flood" and "Indian man"), but also concentrates on learning deeper association semantics



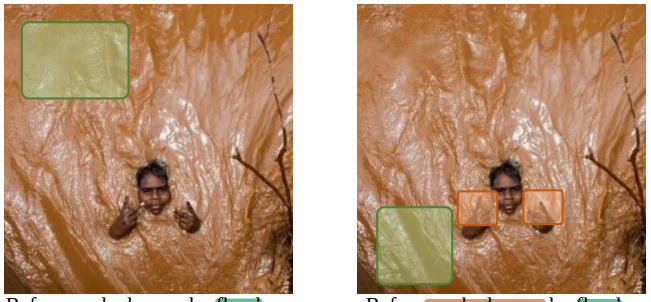
(a) On Twitter dataset (b) On Weibo dataset

Figure 2: Performance comparison between different ablated blocks in cognition-aware fusion layer.



(a) On Twitter dataset (b) On Weibo dataset

Figure 3: Performance comparison between different measurement methods in inconsistency learning.



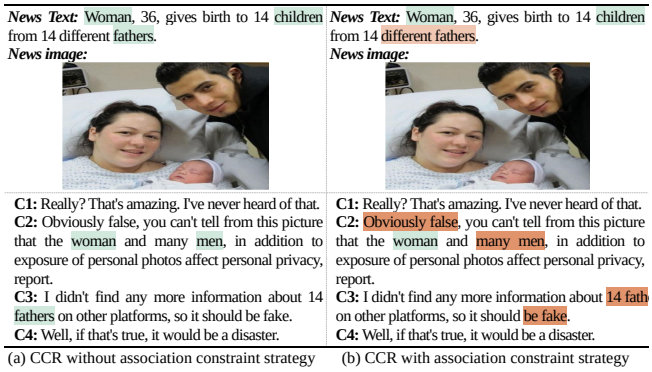
Before washed away by flood, an Indian man calmly gave the last gesture to a photographer.

(a) Features learned by traditional co-attention networks

Before washed away by flood, an Indian man calmly gave the last gesture to a photographer.

(b) Features learned by our CFU layer

Figure 4: The visualization of features captured by co-attention networks and CFU layer



(a) CCR without association constraint strategy (b) CCR with association constraint strategy

Figure 5: The visualization of features captured by CCR with/without association constraint strategy.

("last gesture" and "wash away" in Figure 4(b)) between text and image content of news. These fully reflect the superiority of our CFU layer in the aspect of deep multimodal fusion.

Visualization of Features Captured by CCR To visually verify the ability of our model to capture inconsistent semantics, we compare the features learned by CCR with/without association constraint strategy (i.e., CCR/CCR-cons). As shown in Figure 5, we found that CCR-cons focuses more on coherence semantics between different types of features, like "women", "children", and "fathers" in news and comments. However, our CCR not only focuses on coherence semantics "women" and "children", but also concentrates on inconsistent semantics between different modalities "different fathers", and even emphasizes the difference semantics "14 fathers fake" and "obviously false many men" between news and comments, which vividly depicts the ability of CCR to explore semantic deviation.

Conclusion

In this work, we propose multi-reading habits fusion reasoning networks to tackle the challenges of fusing deeply multimodal features and exploring inconsistent information between them for fake news detection. We design cognition-aware fusion layer inspired by people's three reading habits to learn and integrate multimodal features of news. Furthermore, coherence constraint reasoning layer is developed, which devotes to measure the inconsistency of different modalities of news and discover the semantic deviation between unimodal and multimodal features. Experimental results demonstrate the effectiveness of our model. In the future, we plan to extend our CFU layer by integrating more interdisciplinary knowledge (like social psychology and social cognition) for detection.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants U19B2037, U22B2036, and 62202381, in part by Shenzhen Science and Technology Program and Guangdong Basic and Applied Basic Research Foundation (2021A1515110717), General Program of Chongqing Natural Science Foundation (No. CSTB2022NSCQ-MSX1284), the Fundamental Research Funds for the Central Universities (D5000220185), Sponsored by CAAI-Huawei MindSpore Open Fund, the National Postdoctoral Innovative Talents Support Program for L. Wu. We would like to thank the anonymous reviewers for their constructive comments.

References

Abdali, S.; Gurav, R.; Menon, S.; Fonseca, D.; Entezari, N.; Shah, N.; and Papalexakis, E. E. 2021. Identifying Misinformation from Website Screenshots. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 2–13.

Boulahia, S. Y.; Amamra, A.; Madi, M. R.; and Daikh, S. 2021. Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. *Machine Vision and Applications*, 32(6): 1–18.

Bousmalis, K.; Trigeorgis, G.; Silberman, N.; Krishnan, D.; and Erhan, D. 2016. Domain separation networks. *Advances in neural information processing systems*, 29.

Castillo, C.; Mendoza, M.; and Poblete, B. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*, 675–684.

Chen, Y.; Li, D.; Zhang, P.; Sui, J.; Lv, Q.; Tun, L.; and Shang, L. 2022. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022*, 2897–2905.

Choi, H.; and Ko, Y. 2021. Using Topic Modeling and Adversarial Neural Networks for Fake News Video Detection. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2950–2954.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhawan, M.; Sharma, S.; Kadam, A.; Sharma, R.; and Kumaraguru, P. 2022. GAME-ON: Graph Attention Network based Multimodal Fusion for Fake News Detection. *arXiv preprint arXiv:2202.12478*.

Diseases, T. L. I. 2020. The COVID-19 infodemic. *The Lancet. Infectious Diseases*, 20(8): 875.

Dou, Y.; Shu, K.; Xia, C.; Yu, P. S.; and Sun, L. 2021. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2051–2055.

Guo, C.; Cao, J.; Zhang, X.; Shu, K.; and Yu, M. 2019. Exploiting emotions for fake news detection on social media. *arXiv preprint arXiv:1903.01728*.

Hu, L.; Yang, T.; Zhang, L.; Zhong, W.; Tang, D.; Shi, C.; Duan, N.; and Zhou, M. 2021. Compare to The Knowledge: Graph Neural Fake News Detection with External Knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 754–763.

Jin, Z.; Cao, J.; Guo, H.; Zhang, Y.; and Luo, J. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*, 795–816.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, 2915–2921.

Li, P.; Sun, X.; Yu, H.; Tian, Y.; Yao, F.; and Xu, G. 2021. Entity-Oriented Multi-Modal Alignment and Fusion Network for Fake News Detection. *IEEE Transactions on Multimedia*.

Ma, J.; Gao, W.; Mitra, P.; Kwon, S.; Jansen, B. J.; Wong, K. F.; and Cha, M. 2016. Detecting rumors from microblogs with recurrent neural networks. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2016, 3818–3824.

Ma, J.; Gao, W.; Wei, Z.; Lu, Y.; and Wong, K.-F. 2015. Detect rumors using time series of social context information on microblogging websites. In *Proceedings of the 24th ACM international on conference on information and knowledge management*, 1751–1754.

Meel, P.; and Vishwakarma, D. K. 2021a. HAN, image captioning, and forensics ensemble multimodal fake news detection. *Information Sciences*, 567: 23–41.

Meel, P.; and Vishwakarma, D. K. 2021b. Multi-modal Fusion using Fine-tuned Self-attention and Transfer Learning for Veracity Analysis of Web Information. *arXiv preprint arXiv:2109.12547*.

Nielsen, D. S.; and McConville, R. 2022. MuMiN: A Large-Scale Multilingual Multimodal Fact-Checked Misinformation Social Network Dataset. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3141–3153.

Osmundsen, M.; Bor, A.; Vahlstrup, P. B.; Bechmann, A.; and Petersen, M. B. 2021. Partisan polarization is the primary psychological motivation behind political fake news sharing on Twitter. *American Political Science Review*, 115(3): 999–1015.

Parikh, S. B.; and Atrey, P. K. 2018. Media-rich fake news detection: A survey. In *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*, 436–441. IEEE.

Qi, P.; Cao, J.; Li, X.; Liu, H.; Sheng, Q.; Mi, X.; He, Q.; Lv, Y.; Guo, C.; and Yu, Y. 2021. Improving Fake News Detection by Using an Entity-enhanced Framework to Fuse Diverse Multimodal Clues. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1212–1220.

Qi, P.; Cao, J.; Yang, T.; Guo, J.; and Li, J. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE International Conference on Data Mining (ICDM)*, 518–527. IEEE.

Qian, S.; Wang, J.; Hu, J.; Fang, Q.; and Xu, C. 2021. Hierarchical multi-modal contextual attention network for fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 153–162.

Setty, V.; and Rekve, E. 2020. Truth be Told: Fake News Detection Using User Reactions on Reddit. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 3325–3328.

Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 395–405.

Shu, K.; Mahudeswaran, D.; Wang, S.; and Liu, H. 2020. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 626–637.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Singhal, S.; Dhawan, M.; Shah, R. R.; and Kumaraguru, P. 2021. Inter-modality Discordance for Multimodal Fake News Detection. In *ACM Multimedia Asia*, 1–7.

Singhal, S.; Shah, R. R.; Chakraborty, T.; Kumaraguru, P.; and Satoh, S. 2019. Spotfake: A multi-modal framework for fake news detection. In *2019 IEEE fifth international conference on multimedia big data (BigMM)*, 39–47. IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Verma, P. K.; Agrawal, P.; Amorim, I.; and Prodan, R. 2021. WELFake: word embedding over linguistic features for fake news detection. *IEEE Transactions on Computational Social Systems*, 8(4): 881–893.

Wang, Y.; Ma, F.; Jin, Z.; Yuan, Y.; Xun, G.; Jha, K.; Su, L.; and Gao, J. 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*, 849–857.

Wu, L.; Rao, Y.; Jin, H.; Nazir, A.; and Sun, L. 2019. Different Absorption from the Same Sharing: Sifted Multi-task Learning for Fake News Detection. 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)*, 4644–4653.

Wu, L.; Rao, Y.; Sun, L.; and He, W. 2021a. Evidence inference networks for interpretable claim verification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14058–14066.

Wu, L.; Rao, Y.; Zhang, C.; Zhao, Y.; and Nazir, A. 2021b. Category-controlled encoder-decoder for fake news detection. *IEEE Transactions on Knowledge and Data Engineering*.

Wu, L.; Rao, Y.; Zhao, Y.; Liang, H.; and Nazir, A. 2020. DTCA: Decision Tree-based Co-Attention Networks for Explainable Claim Verification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1024–1035.

Wu, S.; Liu, Q.; Liu, Y.; Wang, L.; and Tan, T. 2016. Information credibility evaluation on social media. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; and Xu, Z. 2021c. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2560–2569.

Xie, J.; Liu, S.; Liu, R.; Zhang, Y.; and Zhu, Y. 2021. SeRN: Stance extraction and reasoning network for fake news detection. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2520–2524. IEEE.

Xue, J.; Wang, Y.; Tian, Y.; Li, Y.; Shi, L.; and Wei, L. 2021. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5): 102610.

Yao, L.; Mao, C.; and Luo, Y. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 7370–7377.

Yu, F.; Liu, Q.; Wu, S.; Wang, L.; and Tan, T. 2017. A convolutional approach for misinformation identification. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 3901–3907.

Zhang, W.; Lam, W.; Deng, Y.; and Ma, J. 2020. *Review-Guided Helpful Answer Identification in E-Commerce*, 26202626. New York, NY, USA.

Zhang, X.; Cao, J.; Li, X.; Sheng, Q.; Zhong, L.; and Shu, K. 2021. Mining dual emotion for fake news detection. In *Proceedings of the Web Conference 2021*, 3465–3476.

Zhou, X.; Wu, J.; and Zafarani, R. 2020. Safe: similarity-aware multi-modal fake news detection (2020). *Preprint. arXiv*, 200304981.