Information Processing and Management 61 (2024) 103564

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Not all fake news is semantically similar: Contextual semantic representation learning for multimodal fake news detection

Liwen Peng [a,b], Songlei Jian [b,*], Zhigang Kan [a,b], Linbo Qiao [a,b], Dongsheng Li [a,b]

[a] *National Key Laboratory of Parallel and Distributed Computing, China*
[b] *College of Computer, National University of Defense Technology, Changsha Hunan 410073, China*

ARTICLE INFO

Keywords:
Fake news detection
Multimodal learning
Social network
Representation learning
Deep learning

ABSTRACT

Multimodal fake news detection, which aims to detect fake news across vast amounts of multimodal data in social networks, greatly contributes to identifying potential risks on the Internet. Although numerous fake news detection methods have been proposed and achieved some progress in recent years, almost all existing methods rely solely on global semantic features to detect fake news while ignoring that fake news is not consistently semantically similar. To fill the gap between news semantic feature space and fake news decision space, we propose a novel method, i.e., Contextual Semantic representation learning for multimodal Fake News Detection (CSFND), by introducing the context information into the representation learning process. Specifically, CSFND implements an unsupervised context learning stage to acquire the local context features of news, which are then fused with the global semantic features to learn the contextual semantic representation of news. In our proposed representation space, semantically dissimilar fake news is explicitly isolated and distinguished from real news separately. Moreover, CSFND devises a contextual testing strategy aimed at distinguishing between fake and real news within the data having similar semantics, wherein the learned decision boundaries are impervious to the semantic characteristics. Extensive experiments conducted on two real-world multimodal datasets demonstrate that CSFND significantly outperforms ten state-of-the-art competitors in detecting fake news and outperforms the best baselines on two datasets by 2.5% on average in terms of Accuracy.

1. Introduction

Fake news spreading on social networks hurts the credibility of official news and even causes crimes worldwide (Alam et al., 2022; Shu et al., 2017). Detecting fake news by extracting features from multimodal data has attracted increasing attention and achieved remarkable progress in recent years (Jing et al., 2023; Yu et al., 2022; Zheng et al., 2022). In general, most existing multimodal fake news detection methods aim to detect fake news relying solely and totally on semantic features inherent to news textual and visual contents (Qian et al., 2021; Wu et al., 2021). However, not all fake news is semantically similar, which means the distribution of news global semantic features is inconsistent with the fake news decision space, making the decision boundary between fake and real news hard to learn. Typically, decision boundaries refer to lines or surfaces that separate different classes in the data space. In representation learning, the representation space in which we learn the decision boundaries to classify data representations is known as the decision space.

[*] Corresponding author.
E-mail addresses: pengliwen13@nudt.edu.cn (L. Peng), jiansonglei@nudt.edu.cn (S. Jian), kanzhigang13@nudt.edu.cn (Z. Kan), qiaolinbo@nudt.edu.cn (L. Qiao), dsli@nudt.edu.cn (D. Li).

https://doi.org/10.1016/j.ipm.2023.103564
Received 26 July 2023; Received in revised form 1 October 2023; Accepted 27 October 2023
0306-4573/© 2023 Elsevier Ltd. All rights reserved.

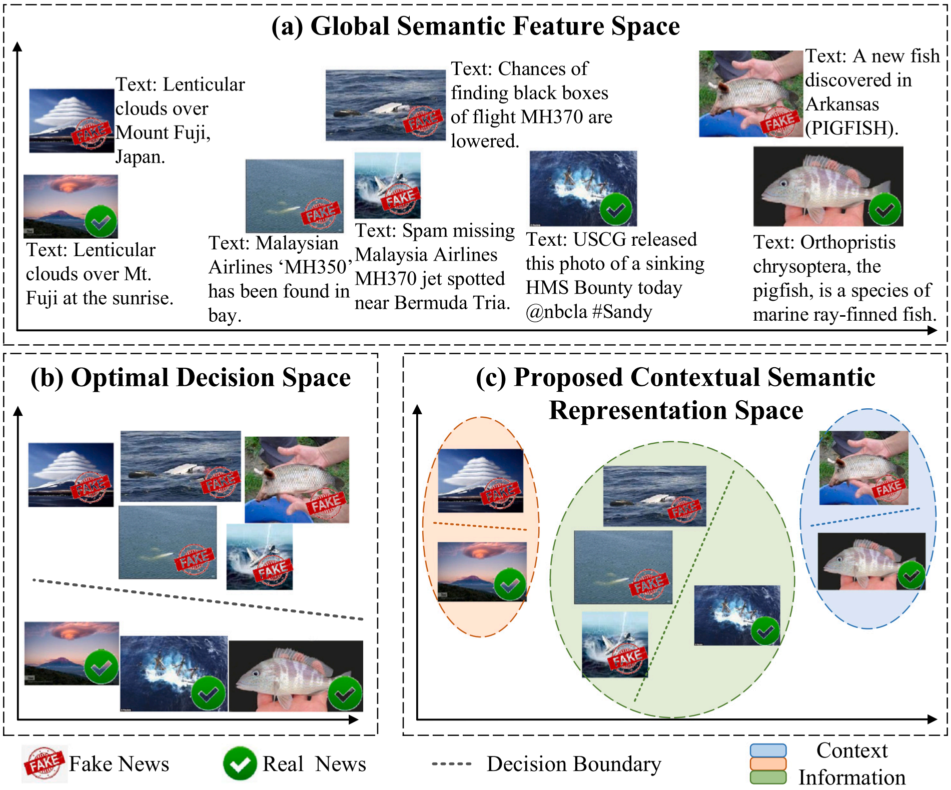

(a) Global Semantic Feature Space
Text: Lenticular clouds over Mount Fuji, Japan.
Text: Chances of finding black boxes of flight MH370 are lowered.
Text: A new fish discovered in Arkansas (PIGFISH).
Text: Lenticular clouds over Mt. Fuji at the sunrise.
Text: Malaysian Airlines 'MH350' has been found in bay.
Text: Spam missing Malaysia Airlines MH370 jet spotted near Bermuda Tria.
Text: USCG released this photo of a sinking HMS Bounty today @nbcla #Sandy
Text: Orthopristis chrysoptera, the pigfish, is a species of marine ray-finned fish.
(b) Optimal Decision Space
(c) Proposed Contextual Semantic Representation Space
FAKE
Fake News
Real News
Decision Boundary
Context Information

Fig. 1. Some real-world fake and real news on Twitter. The news distribution in the (a) global semantic feature space is inconsistent with it in the (b) optimal decision space. However, in our proposed (c) contextual semantic representation space, the context information can bridge the semantic and decision space gap.

We present some real-world news from the most popular social media platform Twitter.[1] as an example to illustrate the global semantic feature space, optimal decision space, and our proposed contextual semantic representation space in Fig. 1 In the (a) global semantic feature space, the distribution demonstrates the natural semantic clusters of the news. While in the (b) optimal decision space, the news is divided into fake and real classes, and the fake news, which is semantically dissimilar, is pushed together, whose distribution is quite different from that in the global semantic feature space. The distribution disparity and inconsistency between the global semantic feature space and optimal decision space make it hard for detection methods to fit the correct boundary between fake and real news.

To address the inconsistency problem, we propose a novel detection method, the Contextual Semantic representation Learning for multimodal Fake News Detection, short for CSFND. CSFND leverages the context information extracted from news to bridge the gap between the global semantic feature space and the optimal decision space. Specifically, CSFND comprises two stages: the unsupervised context learning stage and the supervised contextual detection stage. The unsupervised context learning stage aims to learn the local context features of news, which reflect the news's intrinsic semantic cluster structure. After that, in the supervised contextual detection stage, CSFND selectively combines the learned local context features and global semantic features of news and maps the news into the contextual semantic representation space, in which the local boundaries between fake and real classes within each semantic context are much easier to learn, as shown in Fig. 1(c). Then, the learned textual and visual representations of news are fused to obtain the multimodal representation. Several local fake news detectors are trained to distinguish fake and real news concerning context information differences. Finally, in the inference part, the test news data are assigned to the news groups with similar context information. Then, the corresponding local fake news detector is applied to detect whether the test news is fake or real.

The main contributions of this paper are summarized below.

- We reveal a new observation in the fake news detection task indicating a huge distribution gap between the semantic space and the decision space, which is ignored by almost all existing multimodal fake news detection methods.
- Based on the observation, we propose a novel multimodal fake news detection method, i.e., CSFND, which incorporates context information to learn the contextual semantic representations to alleviate the inconsistency between the semantic space and the decision space and precisely detect fake news within the context information.

[1] https://twitter.com/

- CSFND is an inductive method allowing for the dynamic handling of unseen and newly introduced news. By assigning new data to appropriate news group containing existing data with similar context information, CSFND can effectively distinguish the truthfulness of the new data based on the trained local fake news detector.

Extensive experiments on two real-world multimodal datasets verify the effectiveness of CSFND in detecting fake news compared with SOTA methods. The ablation study demonstrates the contribution of our proposed components. The visualization shows that the learned contextual semantic representations keep the local context features of news and reflect clear boundaries between fake news and real news.

2. Related work

This section briefly reviews the works related to the fake news detection task from unimodal and multimodal methods.

2.1. Unimodal methods

Numerous works investigating the fake news detection task focus on the textual modality of news by analyzing the post texts, user profiles, social metadata, and retweets of news (Huang et al., 2023; Jiang et al., 2022; Lu et al., 2022; Wang et al., 2022). In the beginning, linguistic and statistical characteristics of text content, such as count of opinion words, sentiment, and stylistic features, are used to detect fake news (Kwon et al., 2013; Popat et al., 2016). However, with the evolution and progress in fake news creation, these hand-designed characteristics are insufficient to distinguish fake news that leveled up (Shu et al., 2017). Therefore, to accurately identify news, researchers have resorted to leveraging machine learning and deep neural networks to extract more comprehensive and generalized fake news features (Allein et al., 2023; Kochkina et al., 2023; Zhang et al., 2023). Within this context, adopting Recurrent Neural Networks (RNN) or attention mechanism (Bahdanau et al., 2015) has proven to be instrumental in learning fake news representations in a time series (Chen et al., 2018; Luvembe et al., 2023). Recently, many works construct graphs using social metadata or the news propagation path to effectively and accurately identify fake news (Song et al., 2021; Xu et al., 2022; Yang et al., 2020).

By mining features from the textual modality of news, these neural network-based methods have achieved remarkable performance in detecting fake news.

2.2. Multimodal methods

As images appear increasingly on social networks, which attract more attention and provide extra information, researchers exploit image content to help detect fake news (Alam et al., 2022). Through the extraction of both the textual and visual features of news, the multimodal methods exhibit a superior ability in detecting fake news compared to methods solely using textual features (Wei et al., 2022; Xue et al., 2021; Zheng et al., 2022; Zhou et al., 2020). Intuitively, the features of each modality can be extracted using large-scale textual and visual pre-trained models and concatenated as multimodal representations of news, which are then put into a binary classifier for detecting fake news (Singhal et al., 2020, 2019). Further, attention and transformer (Vaswani et al., 2017) is widely employed to extract correlated characteristics or integrating cross-modal features between textual and visual modalities (Jin et al., 2017; Qian et al., 2021; Wu et al., 2021). MVAE (Khattar et al., 2019) utilizes the variational autoencoder to reconstruct the text and image content of news and applies a binary classifier on the latent representation to detect fake news. SAFE (Zhou et al., 2020) measures the within-modal relationship of both modalities and cross-modal similarity of news to distinguish fake news. Moreover, by introducing real-world knowledge graphs, researchers draw on extra information to help determine whether the news is true or false (Wang et al., 2020). CAFE (Chen et al., 2022) designs a cross-modal ambiguity learning module to tackle the inherent ambiguity across different content modalities.

In recent years, some researchers have shared similar concerns with our thought. The different characteristics of news related to distinct topics or events (Castelo et al., 2019; Hu et al., 2021; Min et al., 2022) and news in specific domains (Nan et al., 2021; Silva et al., 2021; Zhu et al., 2022) are considered in the fake news detection process. In addition, the different topics to which users and publishers of news belong are also considered in the judgment of fake news (Bazmi et al., 2023). However, these works focus only on analyzing the text content to detect fake news, which is insufficient for the multimodal fake news detection task. For the multimodal news, only works EANN (Wang et al., 2018) and MKEMN (Zhang et al., 2019) have noticed the influence of news events. However, the authors in EANN and MKEMN extract the event-invariant features of news to exclude the impact of the difference other than making use of it. In this work, instead, we take advantage of the inconsistency and propose an effectively multimodal fake news method dealing with news text and image content. Furthermore, we introduce the context information of news, a general concept that the topic, event, or domain essentially refers to within the broader realm., to eliminate the inconsistency between the news semantic feature space and optimal decision space.

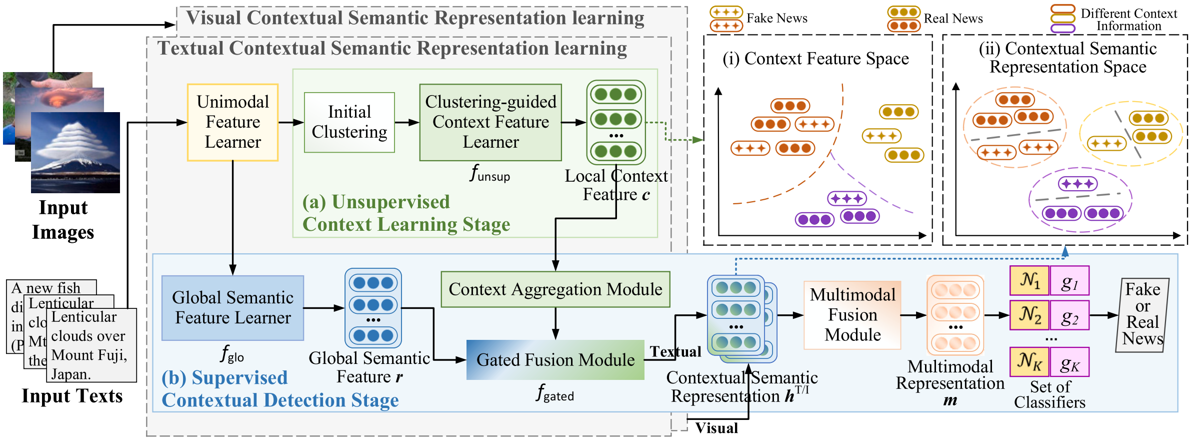

Visual Contextual Semantic Representation learning
Textual Contextual Semantic Representation learning
Input Images
Input Texts
A new fish
Lenticular
Lenticular clouds over Mount Fuji, Japan.
Unimodal Feature Learner
Initial Clustering
Clustering-guided Context Feature Learner
f_{unsup}
Local Context Feature $\boldsymbol{c}$
(a) Unsupervised Context Learning Stage
Global Semantic Feature Learner
f_{glo}
Global Semantic Feature $\boldsymbol{r}$
Context Aggregation Module
Gated Fusion Module
f_{gated}
(b) Supervised Contextual Detection Stage
Textual
Visual
Contextual Semantic Representation $\boldsymbol{h}^{T/I}$
Multimodal Fusion Module
Multimodal Representation $\boldsymbol{m}$
$\mathcal{N}_1$ g_1
$\mathcal{N}_2$ g_2
$\mathcal{N}_K$ g_K
Set of Classifiers
Fake or Real News
Fake News
Real News
Different Context Information
(i) Context Feature Space
(ii) Contextual Semantic Representation Space

Fig. 2. The overall structure of CSFND. The (a) unsupervised context learning stage is to extract the context information of news, which is illustrated in (i) local context feature space. In the (b) supervised contextual detection stage, we learn the contextual semantic representation, illustrated in (ii) contextual semantic representation space, in textual and visual modalities, clearly reflecting the local boundaries between fake and real news. Based on the fused multimodal representation, a set of classifiers is applied to detect fake news concerning different context information. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Problem statement

Typically, fake news refers to a news story or message disseminated via media, bearing false information irrespective of the underlying methods and intentions (Sharma et al., 2019). The fake news detection task aims to determine whether a news article from social media is fake or real, typically formalized as a binary classification problem. Further, the objective of multimodal fake news detection task is to distinguish fake and real news by analyzing the news characteristics from multiple modalities, including the textual content T and visual content I of the news. Specifically, given the training set of news $\mathcal{D}_{\text{tr}} = \{(T_i, I_i, y_i)\}_{i=1}^{N_{\text{tr}}}$, where N_{tr} is the training data size, (T_i, I_i) is a text-image pair attached to one piece of news, and $y_i \in \{1, 0\}$ indicates the label of this news, showing whether it is fake ($y_i = 1$) or real ($y_i = 0$). In this paper, we dedicate to constructing a detection model that can accurately predict the label of news concerning its text and image contents: $F(T, I) \to \hat{y} \in \{0, 1\}$.

4. The proposed method

In this section, we describe the proposed method CSFND in detail. The overall structure of CSFND is shown in Fig. 2. Specifically, the unimodal feature learner module aims to extract modality-specific features of each modality (Section 4.1). Then, in the unsupervised context learning stage, we introduce the context information of news into the representation learning process with the help of clustering techniques (Section 4.2). The supervised contextual detection stage involves the procedure of representation learning, the design of the loss functions, and the prediction of test data labels (Section 4.3). We finally illustrate the training and inference procedures of CSFND by giving the pseudo codes (Section 4.4).

4.1. Unimodal feature learner

Since the data characteristics of different modalities are diverse, we first employ two unimodal feature learners on the text and image contents of news to extract modality-specific textual and visual unimodal features, respectively. We denote the unimodal textual features extracted from the text content T as $\boldsymbol{s} \in \mathbb{R}^{d^T}$ and the unimodal visual features extracted from the image content I as $\boldsymbol{v} \in \mathbb{R}^{d^I}$. With the development of deep neural networks, the large-scale pre-trained language models and vision models have attracted increasing attention (Yang, Feng, et al., 2019). They have been proven effective in capturing semantic features and improved performance in various downstream tasks related to text and image (Khan et al., 2019; Qiu et al., 2020). Therefore, we utilize pre-trained models, such as BERT (Devlin et al., 2019) and XLNet (Yang, Dai, et al., 2019) for texts, VGG (Simonyan & Zisserman, 2015) and ResNet (He et al., 2016) for images, to serve as unimodal feature learners, extracting the modality-specific unimodal features. Thereafter, without loss of generality, we introduce the following of our method by taking the visual modality as a showcase until the multimodal fusion module.

4.2. Unsupervised context learning stage

Considering the essence of multimedia news, which is textual discussions or pictures of specific topics in real life, the natural semantic clusters are a kind of context information that can bridge the gap between semantic space and decision space. Therefore,

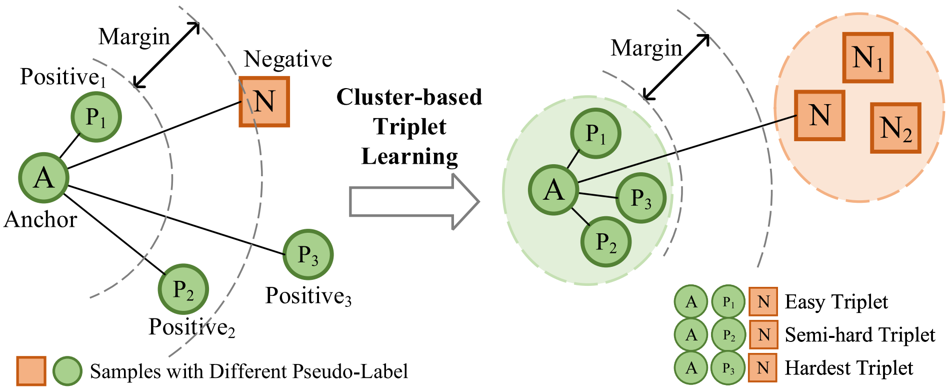


Fig. 3. After cluster-based triplet learning, the positive samples (color in green, having the same pseudo-label with the anchor sample) are closer to the anchor sample than the negative samples (color in orange, having the opposite pseudo-label with the anchor sample). As shown on the left, the easy triplet (A, P_1, N) that satisfies the distance constraint in the cluster-based triple learning training objective provides no contribution to the model training. The semi-hard triplet (A, P_2, N) and hardest triplet (A, P_3, N) are useful for training. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

we learn the contexts in the news by incorporating the clustering process. As shown in Fig. 2(a), we first get the cluster pseudo-label of news through the initial clustering network and then learn the local context features of news in an unsupervised manner.

Specifically, in clustering initialization, we employ the K-Means algorithm (Sculley, 2010) to cluster the data into K clusters and get the cluster pseudo-label $\mathcal{O} = \{o_i\}_{i=1}^{N_{\mathrm{tr}}}$, which indicates the similarity of news context information. We denote $o_i \in \{0, 1, \dots, K\}$ as the cluster pseudo-label of image $\boldsymbol{v}_i$. Moreover, we define the set containing images having similar context information, i.e., the same cluster pseudo-label, with image $\boldsymbol{v}_i$ as $\mathcal{N}_{o_i} = \{\boldsymbol{v}_j\}_{j=1}^{N_{\mathrm{tr}}}, o_j = o_i$.

Since the appearance of triplet learning, it has shown the effectiveness of learning data characteristics based on different distance relationships (Schroff et al., 2015; Wang et al., 2019). Thus we use triplet learning based on the cluster pseudo-label of news to learn news local context features. Given the image triplets $(\boldsymbol{v}_a, \boldsymbol{v}_p, \boldsymbol{v}_n)$, where $\boldsymbol{v}_a$ denotes the anchor sample, we define its positive sample $\boldsymbol{v}_p$ as the image having the same pseudo-label with $\boldsymbol{v}_a$, that is, $\boldsymbol{v}_p \in \mathcal{N}_{o_a}$. Meanwhile, the negative sample $\boldsymbol{v}_n$ has a different pseudo-label with $\boldsymbol{v}_a$, that is, $\boldsymbol{v}_n \in \mathcal{D}_{\mathrm{tr}} - \mathcal{N}_{o_a}$. As shown in Fig. 3, we encourage the local context feature of $\boldsymbol{v}_a$ to be closer to images with similar context information, i.e., the positive sample $\boldsymbol{v}_p$, and further to images with dissimilar context information, i.e., the negative sample $\boldsymbol{v}_n$. Thus we calculate the loss for the unsupervised learning stage as follows:

$$L_{\mathrm{unsup}} = \sum^{\mathcal{T}} \left[\|\boldsymbol{c}_a - \boldsymbol{c}_p\|_2^2 - \|\boldsymbol{c}_a - \boldsymbol{c}_n\|_2^2 + \alpha_{\mathrm{unsup}}\right]_+, \tag{1}$$

where α_{unsup} is a margin that keeps away the distance between negative and positive samples, $\mathcal{T} = \forall((\boldsymbol{v}_a, \boldsymbol{v}_p, \boldsymbol{v}_n)) \in \mathcal{D}_{\mathrm{tr}}$ is the set of valid triplets, and $[x]_+ = \max(x, 0)$. Moreover, $\boldsymbol{c}_{a/p/n} \in \mathbb{R}^d$ is the local context feature of image $\boldsymbol{v}_{a/p/n}$ obtained by: $\boldsymbol{c} = f_{\mathrm{unsup}}(\boldsymbol{v})$, where f_{unsup} is the clustering-guided context feature learner implemented by neural networks. By minimizing Eq. (1), the local context features of the anchor sample and positive sample will be α_{unsup} closer than that of the anchor and negative sample, which means the learned local context features can distinguish news with different context information. As shown in Fig. 2(i), the local context features of news with similar context information (features with the same color) are close in the feature space and far away from news with dissimilar context information (features with different colors).

Further, we design a warm-up strategy for the unsupervised context learning stage. In Fig. 3, concerning the distance difference of anchor, positive and negative samples in triplet, there are three types of triplets: (1) easy triplet (A, P_1, N) that has $L_{\mathrm{unsup}} = 0$ in Eq. (1); (2) semi-hard triplet (A, P_2, N) with a relatively small $L_{\mathrm{unsup}} > 0$; (3) hardest triplet (A, P_3, N) with a large $L_{\mathrm{unsup}} > 0$. Since most of the triplets in the dataset are easy triplets that would not contribute to the model training, we omit them in our training process. While the semi-hard triplets can help the model training and smooth the training procedure, we first use the semi-hard triplets to warm up the training and then use both the semi-hard and hardest triplets until convergence. In addition, for the textual modality, the local context features of news text contents are learned the same as above, except the textual and visual modality parameters are trained independently. Furthermore, the parameters in the unsupervised context learning stage stay fixed in the later learning processing.

Cluster the fused multimodal representations v.s. Cluster the text contents and image contents separately. It is worth noting that the unsupervised clustering process of our method is carried out on textual and visual modality, respectively, rather than conducting clustering on the fused multimodal data. We detail the reasons for doing so below. In a general multimodal learning task, the image and text modalities usually contain some information that is not available in the other modality, i.e., complementary information. Therefore the multimodal methods are committed to fully extracting the data information from the two modalities, obtaining a fused multimodal representation that can adequately represent the characteristics of the data, and then carrying out subsequent comprehension or classification operations based on this multimodal representation. But the situation is somewhat different for the specific task of multimodal fake news detection. In fact, most fake news contains inconsistent text and image content, sometimes even the opposite. Because of this, a critical point in detecting fake news is to uncover inconsistencies between the text and image content in news. Therefore, we cannot fuse the text and image data first and then perform feature extraction operations such as clustering on the fused multimodal representations. Rather, we cluster and extract features for each modality first, uncover the differences between the two modalities, and then analyze the cross-modal relationships to obtain the fused multimodal representations.

4.3. Supervised contextual detection stage

As shown in Fig. 2(b), we first extract the global semantic features of news. Meanwhile, we implement the context aggregation module to aggregate the local context features of news having similar context information. Then we use the gated fusion module to learn the contextual semantic representations, which reflect the distinction between fake and real news concerning different context information, as in Fig. 2(ii). Through the same unimodal representation learning process, we obtain the contextual semantic representations of news in the textual modality. Then, the two representations from textual and visual modalities are fused as the multimodal representation of news. Finally, we design the contextual testing strategy to detect fake news by a set of local classifiers concerning different context information.

4.3.1. Contextual semantic representation learning

Given the image $\boldsymbol{v}$ from one piece of news, we extract its global semantic features by: $\boldsymbol{r} = f_{\text{glo}}(\boldsymbol{v})$, where f_{glo} is the global semantic feature learner implemented by neural networks, and $\boldsymbol{r} \in \mathbb{R}^d$. Intuitively, to introduce the context information into news representation, we can fuse the learned local context feature $\boldsymbol{c}$ with the global semantic feature $\boldsymbol{r}$. However, the contextual characteristics of news are strongly correlated to other news with similar context information, not just to itself. Thus in CSFND, we design the context aggregation module for learning the aggregated context features of news based on news clusters with similar context information. In the experiments, we demonstrate that fusing aggregated context features can detect fake news more effectively than fusing the local context features into news representation.

Given the image $\boldsymbol{v}_i$ with pseudo-label o_i, $\mathcal{N}_{o_i}$ contains news having the same pseudo-label as $\boldsymbol{v}_i$. We learn the aggregated context feature $\boldsymbol{c}_i^{\text{agg}} \in \mathbb{R}^d$ of $\boldsymbol{v}_i$ based on the different importance of news having similar context information with $\boldsymbol{v}_i$ as follows:

$$\boldsymbol{c}_i^{\text{agg}} = \sum_{\boldsymbol{v}_j \in \mathcal{N}_{o_i}, j \neq i} \alpha_{ij} \boldsymbol{c}_j, \tag{2}$$

where $\boldsymbol{c}_j$ is the local context feature of $\boldsymbol{v}_j$ and

$$\alpha_{ij} = \frac{exp(\text{Score}(\boldsymbol{c}_i, \boldsymbol{c}_j))}{\sum_{\boldsymbol{v}_l \in \mathcal{N}_{o_i}} exp(\text{Score}(\boldsymbol{c}_i, \boldsymbol{c}_l))}. \tag{3}$$

The function $\text{Score}(\boldsymbol{c}_i, \boldsymbol{c}_j)$ is calculated by:

$$\text{Score}(\boldsymbol{c}_i, \boldsymbol{c}_j) = v^{\text{agg}} \tanh(\boldsymbol{W}^{\text{agg}}[\boldsymbol{c}_i; \boldsymbol{c}_j]), \tag{4}$$

where v^{agg} and $\boldsymbol{W}^{\text{agg}}$ are parameters to learn, and $[\boldsymbol{c}_i; \boldsymbol{c}_j]$ means the concatenation of these two.

To selectively fuse the context information with the global semantic features of news, we utilize the gated fusion module to learn the contextual semantic representation $\boldsymbol{h}_i \in \mathbb{R}^d$ of $\boldsymbol{v}_i$:

$$\boldsymbol{h}_i = f_{\text{gated}}(\boldsymbol{r}_i, \boldsymbol{c}_i^{\text{agg}}). \tag{5}$$

Since its appearance, the Gated Recurrent Unit (GRU) has been good at extracting critical information from different parts of the input data (Cho et al., 2014; Litjens et al., 2017). Therefore, we implement the gated fusion module f_{gated} by GRU. Overall, the learned contextual semantic representation contains both context information and semantic feature of the news. As shown in Fig. 2(ii), news with similar context information is tightly gathered and far away from news with different context information. Meanwhile, fake news and real news are well distinguished, with a clear boundary in each context region. In addition, we obtain the contextual semantic representation $\boldsymbol{h}^{\text{T}} \in \mathbb{R}^d$ in textual modality with the same data processing procedure, except that the learned parameters in the textual modality are trained independently of those in the visual modality.

4.3.2. Multimodal fusion

Furthermore, we use a multimodal fusion module to flexibly absorb features from textual and visual modalities and obtain the multimodal representation $\boldsymbol{m}$ of news. Inspired by the large-scale multimodal pre-training models (Lu et al., 2019; Su et al., 2020; Wu et al., 2021), we implement a cross-modal attention network based on the standard multi-head self-attention module (Vaswani et al., 2017) to fuse features from multiple modalities. Given the learned textual representation $\boldsymbol{h}^{\text{T}}$ and visual representation $\boldsymbol{h}^{\text{I}}$, the query matrix $\boldsymbol{Q}$ of the cross-attention network is derived from the visual modality $\boldsymbol{H}^{\text{I}}$, and the key $\boldsymbol{K}$ and value $\boldsymbol{V}$ matrices are computed based on the textual representation $\boldsymbol{H}^{\text{T}}$: $\boldsymbol{Q} = \boldsymbol{H}^{\text{I}}\boldsymbol{W}^{Q}, \boldsymbol{K} = \boldsymbol{H}^{\text{T}}\boldsymbol{W}^{K}, \boldsymbol{V} = \boldsymbol{H}^{\text{T}}\boldsymbol{W}^{V}$. Then we get the output of one head attention:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\boldsymbol{Q}\boldsymbol{K}^T / \sqrt{d_k})\boldsymbol{V}. \tag{6}$$

The rest of the cross-attention network proceeds as the standard multi-head self-attention, including the residual add with the feed-forward network. Then, we get the attention-pooled textual representation $\boldsymbol{H}_{T\leftarrow I}$ conditioned on the visual features. Similarly, the attention-pooled visual representation $\boldsymbol{H}_{I\leftarrow T}$ is calculated using textual representation to compute the query matrix and visual representation for the key and value matrices. Finally, we obtain the fused multimodal representation $\boldsymbol{m} \in \mathbb{R}^{2*d}$ by concatenating the two attention-pooled cross-modal visual and textual representations and feeding them through a linear layer:

$$\boldsymbol{m} = \boldsymbol{W}^{\text{fuse}}[\boldsymbol{H}_{T\leftarrow I}; \boldsymbol{H}_{I\leftarrow T}] + \boldsymbol{b}^{\text{fuse}}. \tag{7}$$

4.3.3. Context-maintained training loss

In order to distinguish fake news concerning different context information, we design the context-maintained training losses to train our model, including the context-based triplet loss and intra-context distance loss. The context-based triplet loss aims to separate the fake and real news within news clusters having similar context information under the constraint of the cluster pseudo-labels obtained in Section 4.2. Specifically, given the multimodal representation $\boldsymbol{m}_a$ of one piece of fake news and $\mathcal{N}_{o_a}$ consisting of news having the same cluster pseudo-label with $\boldsymbol{m}_a$, we refer to the fake news in $\mathcal{N}_{o_a}$ as the positive sample $\boldsymbol{m}_p$ and the real news as the negative sample $\boldsymbol{m}_n$. If $\boldsymbol{m}_a$ is real news, we denote the real news as its positive samples and fake news as negative samples. Then, we calculate the context-based triplet loss as follows:

$$L_{\text{con}} = \sum_{o_a=1}^{K} \sum^{\mathcal{T}_a} \left[\|\boldsymbol{m}_a - \boldsymbol{m}_p\|_2^2 - \|\boldsymbol{m}_a - \boldsymbol{m}_n\|_2^2 + \alpha_{\text{con}} \right]_+. \tag{8}$$

The $\mathcal{T}_a = \forall((\boldsymbol{m}_a, \boldsymbol{m}_p, \boldsymbol{m}_n)) \in \mathcal{N}_{o_a}$ consists of valid triplets. The α_{con} is a margin, and K is the cluster number.

Moreover, to strengthen the intra-context relationship, we implement the intra-context distance loss as follows:

$$L_{\text{intra}} = \frac{1}{N_{\text{tr}}} \sum_{k=1}^{K} \sum_{<i,j>}^{\mathcal{N}_k} \|\boldsymbol{m}_i - \boldsymbol{m}_j\|_2^2, \tag{9}$$

where i and j are news having the same pseudo-label, i.e., $o_i = o_j = k$, and $< i, j >$ represents the pairs in $\mathcal{N}_k$ that satisfy $i \neq j$. In fact, the intra-context distance loss aims to compel semantically similar news closer within the representation space, amplifying the correlation amongst news bearing identical context information. Specifically, for each cluster, we calculate the sum of the distances between all pairs of news articles. By minimizing the distance summations of all clusters, news representations with similar semantic features will be closer in their cluster, and news in different clusters will be further away from others in the representation space.

Generally, in multimodal learning, researchers tend to treat one modality as the primary modality containing the leading information and the other as the complementary modality to assist the learning in the primary modality (Baltrušaitis et al., 2019). However, for different multimodal datasets, which modality contains the leading information and should be treated as the primary modality is uncertain. Unfortunately, there is no standard rule for choosing the primary modality. Thus in this work, we design a strategy to select the primary modality for different datasets. Intuitively, the more precise the characteristics of the data in the modality, the higher the quality of the data, indicating that the more it should be treated as the primary modality. Conversely, if the data in the modality is not distinct and contains much noise, it should be treated as a complementary modality providing partly valid information to assist with the task.

Specifically, we cluster modality-specific data features and evaluate the clusters' goodness in each modality to select the primary modality. We think the data characteristics are more precise and explicit in the modality having denser and better-separated clusters, which will be treated as the primary modality. In this paper, we cluster the textual and visual local context features to choose the primary modality, as they reflect the leading context information of data. Then we use the Calinski Harabasz score (Caliński & Harabasz, 1974), which is the ratio of the sum of between-clusters dispersion and inter-cluster dispersion for all clusters, to judge the clustering result and thus determine the primary modality for different datasets. In the model training, we use the pseudo-label in the primary modality to calculate the above two losses.

After that, we employ a fake news detector consisting of a fully connected layer followed by the softmax function on the multimodal representation of news to strengthen the model's ability to predict input news' fake or real label $\hat{y} = \text{FC}(\boldsymbol{m})$. We use the cross-entropy to calculate the prediction loss:

$$L_{\text{pred}} = -\mathbb{E}_{(d,y)\sim(D_{\text{tr}},Y)}[y log(\hat{y}) + (1-y) log(1-\hat{y})], \tag{10}$$

where Y is the set of ground truth labels of news.

Overall, the objective of our method is as follows:

$$L_{\text{all}} = \lambda_{\text{con}} L_{\text{con}} + \lambda_{\text{intra}} L_{\text{intra}} + L_{\text{pred}}. \tag{11}$$

We set the weight of L_{pred} to 1 as the baseline weight and optimize the weights of the other two losses, L_{con} and L_{intra}, to balance the individual terms of the overall loss function. By reducing the three hyper-parameters that need optimization to two, we decrease the computational load and search complexity and improve the search efficiency. Besides, the fake news detection loss L_{pred} is the model's critical training goal and directly correlates with our purpose. We assign its weight to 1, thus making this loss function predominant, intuitively reflecting the relative significance of each loss function within the model.

4.3.4. Contextual testing

In testing, we implement a set of binary classifiers to detect fake news rather than constructing a single classifier for all data. Concerning different context information, using multiple classifiers allows for learning simple and clear classification boundaries within each cluster. In contrast, employing a single classifier yields a challenging boundary that is difficult to fit between all the fake and real news. We first calculate the cluster centers learned by the training data in the unsupervised context learning stage and then determine the pseudo-labels of the testing data by measuring the distance between the testing data and the cluster centers. In this way, all test data is divided into existing clusters containing the training news. After that, for each $\mathcal{N}_k$ containing training and testing data with the same pseudo-label, a binary classifier g_k, $k \in 1, 2, \ldots, K$ is trained by the training data in $\mathcal{N}_k$, and then the classifier is used to predict whether the test data in $\mathcal{N}_k$ is fake or real:

$$\hat{y} = g_k(\boldsymbol{m}), \boldsymbol{m} \in \mathcal{N}_k, \tag{12}$$

where $\boldsymbol{m}$ is the multimodal representation of news.

The inductive learning. Practically, the fake news detection strategy proposed in this paper is inductive and can flexibly deal with new data from unseen semantic clusters. Firstly, once the networks of CSFND are well-trained, we obtain the multimodal representations and cluster pseudo-labels of all the training data. We then calculate the clustering centers of each cluster based on the multimodal representations and cluster pseudo-labels, each cluster comprising semantically similar training data. Secondly, we feed the new data into the well-trained network to acquire its multimodal representation. Then, we compute the distance between this representation and the centers of all clusters drawn from the training data and allocate the new data to the closest cluster. Regardless of whether the semantic characteristics of the new data have appeared in the training data, it is certain that by calculating its multimodal representation and the distance to all cluster centers, we can allocate the new data to an appropriate existing cluster. Finally, we implement a corresponding binary classifier for each cluster composed of training data and new data. This classifier, trained by labeled training data from each cluster, can predict the labels of the newly assigned data in the same cluster. In this way, our method has the flexibility to handle and predict any new data points that have not been seen in the training set. Further, for new clusters/events, as our clusters are determined by context information similarity rather than the exact event label of news, our model can find the appropriate clusters for the unseen events most similar to the existing clusters learned by the training data and thus process new clusters/events.

4.4. Algorithm of CSFND

In order to deliver a clearer understanding of our model training and inference process, we demonstrate the flow of CSFND in this section.

Algorithm 1 presents the training procedure of CSFND. We first extract the unimodal features of each modality in Step 1. Steps 3-9 are the unsupervised context learning stage, which is carried out separately in the text and image modalities. Thus in step 9, we get the local context features $\boldsymbol{c}$ of the two modalities. For each training batch, the contextual semantic representations $\boldsymbol{h}^{\mathrm{I}}$ and $\boldsymbol{h}^{\mathrm{T}}$ in visual and textual modalities are learned in Step 12. Step 13 is the multimodal fusion procedure. Then, the overall training loss is computed in Step 14. Step 15 performs back propagation to optimize the network parameters. After several model training epochs, we obtain the trained CSFND model networks.

Algorithm 1 Training of CSFND

```
Input: Data in training set - $\{(T, I, y)\} \in \mathcal{D}_{\mathrm{tr}}$, cluster number $K$, epoch of unsupervised stage, epoch of supervised stage
Output: CSFND network - $\Pi$
 1: Extract unimodal textual/visual features $\boldsymbol{s}/\boldsymbol{v}$ of $T/I$
 2: for modality in [visual, textual] do
 3:     Cluster $\boldsymbol{v}/\boldsymbol{s}$ into $K$ clusters and get pseudo-label $\mathcal{O}$
 4:     for $i = 1$ to $unsuper\_epoch$ do
 5:         Sample valid triplets $(\boldsymbol{v}_a, \boldsymbol{v}_p, \boldsymbol{v}_n)$
 6:         Compute loss $L_{\mathrm{unsup}}$ by Eq. (1)
 7:         Optimize parameters of $L_{\mathrm{unsup}}$
 8:     end for
 9:     Compute $\boldsymbol{c} \leftarrow$ fix parameters of the unsupervised networks
10: end for
11: for $j = 1$ to $super\_epoch$ do
12:     Get contextual semantic representation $\boldsymbol{h}^{\mathrm{I}}$ and $\boldsymbol{h}^{\mathrm{T}}$ by Eq. (5)
13:     Get multimodal representation $\boldsymbol{m}$ by Eqs. (6)–(7)
14:     Calculate loss $L_{\mathrm{all}}$ by Eqs. (8)–(11)
15:     Optimize parameters of $L_{\mathrm{all}}$
16: end for
17: return Trained network $\Pi$
```

Algorithm 2 Inference of CSFND

```
Input: Training set $\mathcal{D}_{\mathrm{tr}}$, trained CSFND network $\Pi$, testing set $\mathcal{D}_{\mathrm{te}}$, cluster number $K$
Output: Predicted labels $\hat{y}$ of test data
 1: Get the multimodal representation $\boldsymbol{m}$ of train data by $\Pi$
 2: Compute the $K$ cluster centers of train data by $\boldsymbol{m}$
 3: for data $i$ in test set do
 4:     Get the multimodal representation $\boldsymbol{m}_i$ of $i$ by $\Pi$
 5:     Compute the distance between $\boldsymbol{m}_i$ and $K$ cluster centers
 6:     Assign $i$ into the nearest cluster $k_i$
 7: end for
 8: for $k = 1$ to $K$ do
 9:     Build a classifier $g_k$
10:     Train $g_k$ using the train data in cluster $k$
11:     Use $g_k$ to predict the label of test data in cluster $k$
12: end for
13: return $\hat{y}$ of all test data
```

Table 1
The statistics of Weibo and Twitter datasets.

	Weibo	Twitter
# of Fake news	4211	7979
# of Real news	3642	6467
# of Images	7851	476

The inference process of our CSFND network is illustrated in Algorithm 2. Given the data with labels in the training set $\mathcal{D}_{\mathrm{tr}}$, the data in the testing set $\mathcal{D}_{\mathrm{te}}$ and the trained network Π, the inference process aims to predict the fake and real labels $\hat{y}$ of all the data in the test set. In Steps 1 and 2, we use the multimodal representations and the cluster pseudo-labels of the training set data to compute the cluster centers of the K clusters. Then, by calculating the distance of each test data from the centers of the K clusters, we assign each test set data to its nearest cluster in Steps 3-6. Finally, in Steps 7-11, we train a corresponding binary classifier g_k for each cluster k containing training and test data. The trained classifier g_k is used to predict the labels of the test data in cluster k. In this way, we obtain fake and real labels for all test data.

5. Experiments

In this section, we conduct extensive experiments to verify the effectiveness of our method in detecting fake news. We first describe the detailed experimental setups in Sections 5.1–5.3 and then analyze the experiment results in Sections 5.4–5.7.

5.1. Datasets

We evaluate the performance of our method compared with other baselines on two widely used real-world multimodal fake news detection datasets, i.e., Weibo (Jin et al., 2017) and Twitter (Boididou et al., 2016). News data in the Weibo and Twitter datasets are collected from the most popular social media websites worldwide, Weibo[2] and Twitter, respectively. Each news data contains text content and corresponding image content. Remarkably, the Twitter dataset contains fewer distinct images than distinct texts, as some news articles present different texts with identical attached images. For this case, we process the news articles following the baseline methods. For each news article with distinct text content, we compose the text content and its attached image as a news article, forming a single input data for our model. Consequently, some news items have the same image, but since their text content differs, they are still different input data. As the other works do (Khattar et al., 2019; Wu et al., 2021), we remove the text-only and image-only news from the datasets, and the videos are out of consideration in the experiments. Also, we keep the same data split scheme as the benchmark on these two datasets Simple statistics of the datasets are summarized in Table 1.

5.2. Baselines

We compare our method with two types of baselines:

[2] https://weibo.com/

- *Unimodal methods*: To validate the effectiveness of the text and image contents of news for detecting fake news, we select the language and vision pre-trained models BERT (Devlin et al., 2019), XLNet (Yang, Dai, et al., 2019), and VGG-19 (Simonyan & Zisserman, 2015) as the unimodal comparison methods. It is because these models are used as unimodal feature extractors in most existing multimodal fake news detection methods. Meanwhile, the pre-trained models perform well in extracting textual and visual semantic features. In the experiments, we use the models to extract the textual (BERT and XLNet) or visual (VGG-19) unimodal representations of news, which are then put into a fully connected layer followed by a softmax function to predict whether the news is fake or real.
- *Multimodal methods*: We compare ten multimodal fake news detection methods, including classical fake news detection methods and SOTA approaches. (1) **att-RNN** (Jin et al., 2017) fuses textual, visual and social context features by attention. In our experiments, we remove the component that processes social metadata for a fair comparison. (2) **EANN** (Wang et al., 2018) applies an event discriminator on the concatenated multimodal representation to exclude the event-specific features of news. For a fair comparison, we remove the discriminator in the experiments. (3) **MVAE** (Khattar et al., 2019) uses the variational autoencoder to learn the latent representations of text and image, where a binary classifier is used to detect fake news. (4) **SpotFake** (Singhal et al., 2019) concatenates unimodal features extracted by BERT and VGG-19 as news' multimodal representations. (5) **MKEMN** (Zhang et al., 2019) uses a convolutional operation to fuse the text, image, and external retrieved knowledge. In the experiments, we remove the part of retrieved knowledge. (6) **SpotFake+** (Singhal et al., 2020) extends SpotFake by extracting the textual features using XLNet, and then the concatenated multimodal features are also feed into a binary classifier. (7) **SAFE** (Zhou et al., 2020) investigates the within-modal relationship of textual and visual modalities and the cross-modal similarity between two modalities to detect fake news. (8) **HMCAN** (Qian et al., 2021) designs the multimodal attention networks based on the self-attention (Vaswani et al., 2017) to fuse the textual and visual features. For news without images, HMCAN generates dummy images to construct text-image pairs. For a fair comparison, in the experiments, we remove the news without images, as the other benchmarks do (Jin et al., 2017; Khattar et al., 2019; Zhou et al., 2020). (9) **MCAN** (Wu et al., 2021) extracts the spatial-domain and frequency-domain features of the images, which are then fused with the textual features by multiple cross-modal co-attention blocks. (10) **CAFE** (Chen et al., 2022) adaptively aggregates the unimodal features of text and image and the cross-modal correlations to detect fake news.

We keep the same parameter settings reported in their papers for all the baselines. Moreover, we report the overall classification Accuracy (**Acc**) and the Precise (Pre), Recall (Rec) and **F1** scores for the fake and real news detection, respectively.

5.3. Implementation details

For a fair comparison, we use BERT (Devlin et al., 2019) and VGG-19 (Simonyan & Zisserman, 2015) as the textual and visual unimodal feature learners in our method CSFND for both datasets, as do most baselines (Singhal et al., 2019; Wu et al., 2021). For the Twitter dataset with English, we use the bert-base-cased model, and for Weibo with Chinese, we use the bert-base-chinese model (Wolf et al., 2020).[3]

For text, each sentence is padded or truncated to have the same number of tokens. The token number N_{token} are 160 and 25 for Weibo and Twitter, respectively. We use BERT to get the embeddings of each token with a dimension of 768. Then, the token embeddings are concatenated as the unimodal textual feature with $d^T = N_{token} * 768$. For images, the unimodal visual feature is the output of the penultimate linear layer of VGG-19 with a dimension of $d^I = 4096$. In our experimental implementation and the reproductions of baseline methods, the parameters of the pre-trained models employed in all these methods, including BERT, XLNet, VGG-19, and ResNet, are frozen, ensuring a fair comparison.

In CSFND, the dimensions of the local context feature $\boldsymbol{c}$, global semantic feature $\boldsymbol{r}$, and the aggregated context feature $\boldsymbol{c}^{\text{agg}}$ of textual and visual modality are all set to $d = 128$. Our learned contextual semantic representation $\boldsymbol{h}$ is 128-dimensional. The dimension of multimodal representation is $2 * d = 256$. Moreover, learner f_{unsup} consists of two linear layers with the size of 512 and 128. Learner f_{glo} is a linear layer with a dimension of $d = 128$. The activate function for them is ReLU. The training epoch is 100, with early stopping on the validation set. Our algorithms are trained by the Adam optimizer (Kingma & Ba, 2014) with a batch size of 128. The learning rates for Weibo and Twitter are 0.001 and 0.0005. The optimal hyper-parameters are determined by grid searching on the validation set, and the selection criterion is accuracy. We get $\alpha_{\text{unsup}} = 0.5, \alpha_{\text{con}} = 0.2$ and $\lambda_{\text{con}} = 0.6, \lambda_{\text{intra}} = 0.2$. The cluster number K are 17 and 33 for Weibo and Twitter, respectively.

In testing, each classifier is one linear layer followed by tanh with independent parameters. For each classifier, we record the number of true positive, false negative, true negative, and false positive predictions for fake and real news, respectively. Then we sum the numbers of all classifiers to get the number of predictions of the whole test set, which are used to calculate the classification Accuracy of our method and Precise, Recall and F1 results for fake and real news detection.

5.4. Results and analysis

Table 2 displays the results of CSFND and all the baselines on Weibo and Twitter datasets. We can observe that, in terms of the Accuracy (**Acc**) and **F1** scores for both fake and real news detection, our method outperforms all the baselines on the two datasets.

[3] https://github.com/huggingface

Table 2
Fake news detection results of different methods on Weibo and Twitter datasets.

Model	Weibo							Twitter						
	Acc	Fake news			Real news			Acc	Fake news			Real news		
		Pre	Rec	F1	Pre	Rec	F1		Pre	Rec	F1	Pre	Rec	F1
Bert (2019)	0.845	0.858	0.833	0.845	0.833	0.857	0.844	0.642	0.666	0.766	0.711	0.602	0.474	0.526
XLNet (2019)	0.842	0.859	0.826	0.842	0.827	0.859	0.843	0.613	0.644	0.737	0.687	0.557	0.445	0.493
VGG-19 (2015)	0.647	0.640	0.700	0.668	0.657	0.591	0.621	0.767	0.829	0.753	0.787	0.704	0.785	0.740
att-RNN (2017)	0.772	0.854	0.656	0.742	0.720	0.889	0.795	0.664	0.749	0.615	0.676	0.589	0.728	0.651
EANN (2018)	0.782	0.827	0.697	0.756	0.752	0.863	0.804	0.648	0.810	0.498	0.617	0.584	0.759	0.660
MVAE (2019)	0.824	0.854	0.769	0.809	0.802	0.875	0.837	0.745	0.801	0.719	0.758	0.689	0.777	0.730
SpotFake (2019)	0.869	0.877	0.859	0.868	0.861	0.879	0.870	0.771	0.784	0.744	0.764	0.769	0.807	0.787
MKEMN (2019)	0.814	0.823	0.799	0.812	0.723	0.819	0.798	0.715	0.814	0.756	0.708	0.634	0.774	0.660
SpotFake+ (2020)	0.870	0.887	0.849	0.868	0.855	0.892	0.873	0.790	0.793	0.827	0.810	0.786	0.747	0.766
SAFE (2020)	0.763	0.833	0.659	0.736	0.717	0.868	0.785	0.766	0.777	0.795	0.786	0.752	0.731	0.742
HMCAN (2021)	0.790	0.803	0.758	0.780	0.778	0.821	0.799	0.759	0.705	0.745	0.724	0.804	0.770	0.786
MCAN (2021)	0.873	**0.944**	0.794	0.863	0.817	**0.952**	0.879	0.796	0.785	**0.891**	0.835	**0.819**	0.669	0.736
CAFE* (2022)	0.840	0.855	0.830	0.842	0.825	0.851	0.837	0.806	0.807	0.799	0.803	0.805	0.813	0.809
CSFND	**0.895**	0.899	**0.895**	**0.897**	**0.892**	0.896	**0.894**	**0.833**	**0.899**	0.799	**0.846**	0.763	**0.878**	**0.817**

The bold values represent the best results. For methods that do not open their source codes, we display the results reported in their papers with *.

Table 3
Fake news detection results of CSFND and its four ablated versions with the improvement rates of CSFND compared to its variants per dataset. Positive rates are boldfaced.

Model	Weibo			Twitter		
	Accuracy	F1 of fake	F1 of real	Accuracy	F1 of fake	F1 of real
w/o UNSPR	0.852 (+**5.0%**)	0.853 (+**5.2%**)	0.851 (+**5.1%**)	0.658 (+**26.6%**)	0.676 (+**25.1%**)	0.638 (+**28.1%**)
ONE_CLS	0.874 (+**2.4%**)	0.875 (+**2.5%**)	0.873 (+**2.4%**)	0.729 (+**14.3%**)	0.749 (+**13.0%**)	0.705 (+**15.9%**)
w/o AGG	0.848 (+**5.5%**)	0.848 (+**5.8%**)	0.848 (+**5.4%**)	0.741 (+**12.4%**)	0.747 (+**13.3%**)	0.736 (+**11.0%**)
w/ AVG	0.866 (+**3.3%**)	0.875 (+**2.5%**)	0.856 (+**4.4%**)	0.797 (+**4.5%**)	0.821 (+**3.0%**)	0.766 (+**6.7%**)
CSFND	0.895	0.897	0.894	0.833	0.846	0.817

Firstly, the unimodal methods' results demonstrate that both news text and image contents help detect fake news. Further, for Weibo, the textual methods (BERT, XLNet) perform better than the visual method, while on Twitter, the visual method VGG-19 gains a higher accuracy. It means that in different datasets, the modality containing the leading information is different, textual modality in Weibo and visual modality in Twitter. As illustrated in Section 4.3.3, in our method, the selected primary modalities of Weibo and Twitter are textual and visual, respectively, consistent with the detection results of unimodal methods on these two datasets.

Further, the multimodal methods SpotFake, SpotFake+, MCAN, and our method CSFND perform better than the unimodal methods in both datasets, which proves the effectiveness of extracting features from multiple modalities for detecting fake news. The higher detection results of CSFND than EANN and MKEMN indicate that the effective usage of context information (events of news) helps distinguish fake news. EANN and MKEMN learn event-irrelevant features, which may lose some essential fake news characteristics in the extracted news features. Comparing most detection metrics, especially the Accuracy and F1 scores, on both datasets, our method CSFND outperforms all the other baselines. It demonstrates that CSFND, considering the inconsistency between the global semantic feature space and the optimal decision space, can better detect fake news according to different context information.

5.5. Ablation study

To validate the effectiveness of the proposed components in our method, we carry out the ablation study shown in Table 3. The bottom row represents the entire model of our method, denoted as CSFND.

5.5.1. Effect of the unsupervised context learning stage

The sub-model w/o UNSPR is the reduced model without the unsupervised context learning stage. We fuse the extracted textual and visual global semantic features as the multimodal representation of news, which is fed into a classifier to detect fake news. Comparing the results with CSFND, we can observe that the unsupervised context learning stage significantly contributes to detecting fake news, which means introducing the context information into the news representation learning is helpful.

5.5.2. Effect of the contextual testing strategy

In ONE_CLS, we predict the labels of all test data by the detector trained in Eq. (10). The significant improvement of CSFND compared to ONE_CLS shows the necessity of applying multiple classifiers concerning different context information in testing, which is much better than using one classifier to learn a complex boundary between all the fake and real news.

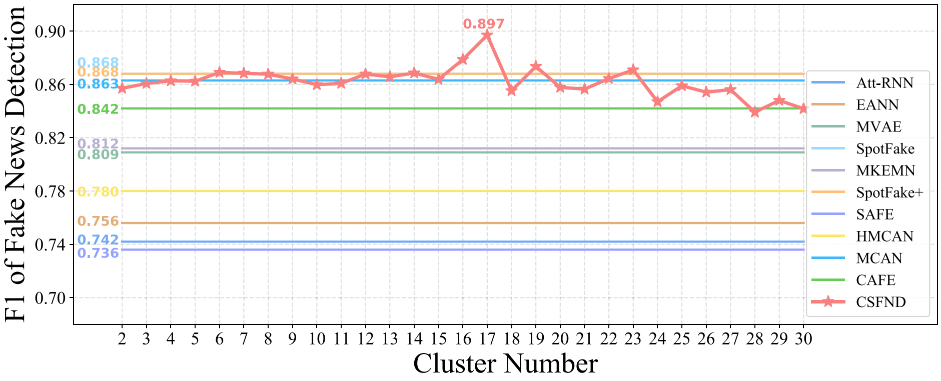


Fig. 4. Impact of the key hyper-parameter cluster number K for the F1 score of fake news detection on Weibo dataset of our method CSFND (the line with color red). The lines with other colors represent the F1 scores achieved by state-of-the-art multimodal fake news methods. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.5.3. Effect of the context aggregation module

With the context aggregation module removed, the sub-model w/o AGG gets the contextual semantic representation by fusing the local context features and global semantic features. In contrast, the sub-model w/ AVG averages the local context features of news having similar context information, which are then fused with the global semantic features. The higher results of w/ AVG than w/o AGG show the necessity to extract features from news clusters with similar context information. Their lower results compared to CSFND indicate the effectiveness of our context aggregation module.

5.6. Sensitivity analysis

In this section, we investigate the influence of different settings of the key hyper-parameters of our method CSFND, including the cluster number K, the weights λ_{con} and λ_{intra} of the overall loss function.

5.6.1. Impact of the cluster number K

As shown in Fig. 4, the F1 scores of fake news detection remain stable as the cluster number K varies (the line with color red). To provide a clearer understanding of the F1 scores obtained by setting different values of K, we also display the F1 scores achieved by SOTA methods in recent years in Fig. 4 for comparison (the lines with colors not red). By introducing the clustering process into the fake news detection procedure, our method achieves F1 results on par with the highest fake news detection methods available. Further, for some values of the cluster number, our method achieves F1 scores higher than all other methods, which means the effective use of context information helps to improve the performance of fake news detection.

5.6.2. Impact of the weights λ_{con} and λ_{intra}

In this section, we investigate the impact of the different choices of weights λ_{con} and λ_{intra} on the overall loss function in Eq. (11). In our method, the overall loss function L_{all} is composed of three terms: context-based triplet loss L_{con}, intra-context distance loss L_{intra} and fake news prediction loss L_{pred}. We assign the weight of L_{pred} to 1 as the baseline weight. Then, we vary λ_{con} and λ_{intra} from 0.0 to 1.0 with an interval of 0.1 to mirror the relative significance of the three losses within the model. The optimal hyper-parameters are determined according to the Accuracy results. Fig. 5 illustrates the variations in classification Accuracy under the joint influence of these two parameters on the Twitter dataset. The highest Accuracy result is achieved when the value of λ_{intra} is 0.2 and the value of λ_{con} is 0.6. On the Weibo dataset, such parameter settings also yield better results. Thus, we set the weights λ_{con} and λ_{intra} to 0.6 and 0.2 in our experiments.

5.7. Visualization

To further analyze the effectiveness of our method in learning the contextual semantic representations of news and distinguishing the fake and real news, we qualitatively visualize the representations of the test data in Weibo learned by CSFND and SpotFake with t-SNE (Laurens & Hinton, 2008) in Fig. 6.

Firstly, sub-figures 6(a) and 6(b) present the local context features and contextual semantic representations learned by CSFND. In 6(a), the local context features of news with similar context information are gathered closely (different colors represent different context information), which means our unsupervised context learning stage is effective in capturing the context information of news. Then, in 6(b), the learned contextual semantic representations can reflect the semantic similarity between news as well as separate the fake and real news (denoted by symbols • and +). We can see that news representations with similar context information are mapped together. Meanwhile, the fake and real news within each context region are divided separately and can be easily distinguished. Taking the data with the color purple as an example, we can see that the purple data in 6(a) are close and far away

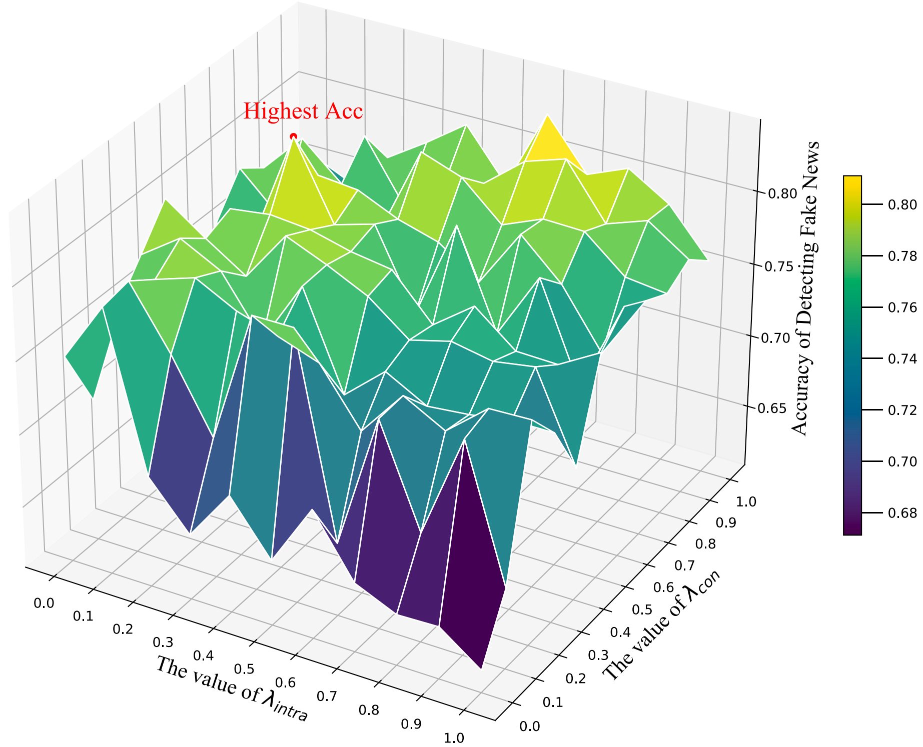
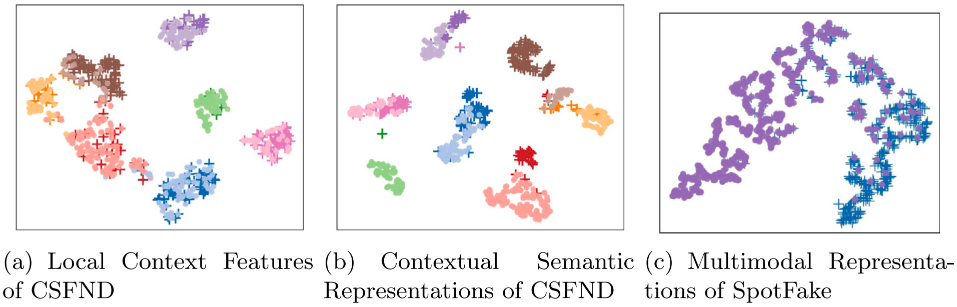


Fig. 5. Impact of the key hyper-parameters λ_{intra} and λ_{con} for the Accuracy of fake news detection on Twitter dataset.



(a) Local Context Features of CSFND
(b) Contextual Semantic Representations of CSFND
(c) Multimodal Representations of SpotFake

Fig. 6. Visualizations of the learned representations of test data on Weibo of CSFND and SpotFake. Symbols • and + represent fake and real news, respectively. In (a) and (b), different colors denote the group of data with different context information. In (c), different colors represent fake and real news. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

from data with different colors. Then, in 6(b), in the compact region formed by the purple data, the fake and real news are separated clearly with the local decision boundary easily to determine.

Further, we display the learned multimodal representations of SpotFake in Fig. 6(c). Taking into account the unimodal feature extraction models of the baselines, we select SpotFake for comparison. In 6(c), SpotFake tries to gather all the fake news together without considering that the fake news is not all semantically similar. Moreover, it is complex and inconsistent with the natural characteristics of news, trying to use one boundary to separate all the fake and real news. In contrast, in 6(b), CSFND learns news representations based on different context information, successes in bridging the gap between the semantic space and decision space. In our proposed representation space, it is easier to distinguish fake and real news in each context region by learning several local boundaries. For example, in 6(b), the fake and real news with the color orange and brown can be accurately distinguished by their corresponding contextual classifiers in our method concerning their different context information. However, for other methods not considering the context information, like SpotFake in 6(c), these news are mixed and it is hard to find the decision boundary dividing the fake and real news well. For example, the real news (symbol +) with the color orange may be detected as fake (symbol •) as their representations are closer to the fake news (symbol • of data in orange and brown) and far away from the real news (symbol + of data in brown).

6. Discussions

6.1. Implications

First, in detecting fake news, this paper reveals the inconsistency problem between the semantic space and the decision space of fake and real news, which most existing multimodal fake news detection methods ignore. Existing methods directly classify fake and real news in the semantic space of all the news, suffering from the problem that the classification boundary in the global semantic space is complex and challenging to fit. The experiments carried out in this paper, particularly the visualizations of the representations learned by CSFND and SpotFake, demonstrate that this inconsistency problem exists and that addressing this problem can improve the performance of fake news detection.

Second, to address the inconsistency problem, this paper proposes a contextual semantic representation space where several concise classification boundaries can be learned in local news clusters, considering different context information. Specifically, we design the unsupervised context learning stage and use clustering to introduce the context information into the news representation learning process. The extensive experiments conducted in this paper demonstrate that the inconsistency problem can be addressed by clustering semantically similar news and detecting fake news within the clusters. CSFND outperforms other comparative methods in detecting fake news on both Weibo and Twitter datasets, suggesting that our proposed method can improve the performance of detecting fake news.

6.2. Limitations

To address the inconsistency between the global semantic space and the decision space of news, we design unsupervised clustering-based context learning to introduce the context information into the news representation learning process. Further, we classify fake and real news in the learned contextual semantic representation space concerning different context information, i.e., detect fake news within each clusters. However, there is still much to be explored to address the inconsistency problem, such as designing other strategies besides clustering to exploit the inconsistent information or alleviating the impact of the inconsistency by transfer learning.

In addition, in our implementation, considering the effectiveness and efficiency, we choose the basic K-Means (Sculley, 2010) algorithm for clustering and determine the appropriate number of clusters by grid search on the validation set. However, the cluster number can be determined in a more flexible way, and alternative clustering algorithms can be explored, such as distribution-based, density-based, hierarchical-based, etc. We plan to investigate the problems further in future works.

7. Conclusions

This paper reveals the inconsistency between the global semantic feature space and the optimal decision space in fake news detection. To solve the problem, we propose the method CSFND to learn the contextual semantic representations in which the local boundaries between fake and real news are easier to learn concerning different context information. Extensive experiments verify the effectiveness of CSFND in detecting fake news, and the qualitative visualization results prove the ability to relieve the inconsistency problem.

CRediT authorship contribution statement

Liwen Peng: Conceptualization, Methodology, Software, Validation, Investigation, Writing – original draft. **Songlei Jian:** Conceptualization, Methodology, Investigation, Writing – review & editing, Supervision, Funding acquisition. **Zhigang Kan:** Software, Validation, Investigation, Writing – review & editing. **Linbo Qiao:** Writing – review & editing, Resources, Funding acquisition. **Dongsheng Li:** Writing – review & editing, Funding acquisition.

Data availability

Data will be made available on request.

Acknowledgments

This work is in part supported by the National Natural Science Foundation of China under Grants 62025208, 61932001, and 62002371, Foundation of National University of Defense Technology, China under Grant ZK21-17, and Xiangjiang Laboratory Foundation under Grant 22XJ01012.

References

Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D., Martino, G. D. S., Shaar, S., Firooz, H., & Nakov, P. (2022). A survey on multimodal disinformation detection. In *Proceedings of the 29th international conference on computational linguistics* (pp. 6625–6643).

Allein, L., Moens, M.-F., & Perrotta, D. (2023). Preventing profiling for ethical fake news detection. *Information Processing & Management*, *60*(2), Article 103206.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International conference on learning representations*.

Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(2), 423–443.

Bazmi, P., Asadpour, M., & Shakery, A. (2023). Multi-view co-attention network for fake news detection by modeling topic-specific user and news source credibility. *Information Processing & Management*, *60*(1), Article 103146.

Boididou, C., Papadopoulos, S., Dang Nguyen, D. T., Boato, G., Riegler, M., Petlund, A., & Kompatsiaris, I. (2016). Verifying multimedia use at MediaEval 2016. In *MediaEval 2016 workshop*.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*(1), 1–27.

Castelo, S., Almeida, T., Elghafari, A., Santos, A., Pham, K., Nakamura, E., & Freire, J. (2019). A topic-agnostic approach for identifying fake news pages. In *Companion proceedings of the 2019 world wide web conference* (pp. 975–980).

Chen, T., Li, X., Yin, H., & Zhang, J. (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* (pp. 40–52).

Chen, Y., Li, D., Zhang, P., Sui, J., Lv, Q., Tun, L., & Shang, L. (2022). Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM web conference 2022* (pp. 2897–2905).

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 conference on empirical methods in natural language processing* (pp. 1724–1734).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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)* (pp. 4171–4186).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, L., Yang, T., Zhang, L., Zhong, W., Tang, D., Shi, C., Duan, N., & 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)* (pp. 754–763).

Huang, Y., Gao, M., Wang, J., Yin, J., Shu, K., Fan, Q., & Wen, J. (2023). Meta-prompt based learning for low-resource false information detection. *Information Processing & Management*, *60*(3), Article 103279.

Jiang, G., Liu, S., Zhao, Y., Sun, Y., & Zhang, M. (2022). Fake news detection via knowledgeable prompt learning. *Information Processing & Management*, *59*(5), Article 103029.

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

Jing, J., Wu, H., Sun, J., Fang, X., & Zhang, H. (2023). Multimodal fake news detection via progressive fusion networks. *Information Processing & Management*, *60*(1), Article 103120.

Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2019). A survey of the recent architectures of deep convolutional neural networks. CoRR abs/1901.06032. arXiv:1901.06032.

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. In *The world wide web conference* (pp. 2915–2921).

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. In *International conference on learning representations*.

Kochkina, E., Hossain, T., Logan, R. L., Arana-Catania, M., Procter, R., Zubiaga, A., Singh, S., He, Y., & Liakata, M. (2023). Evaluating the generalisability of neural rumour verification models. *Information Processing & Management*, *60*(1), Article 103116.

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. In *2013 IEEE 13th international conference on data mining* (pp. 1103–1108).

Laurens, V. D. M., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(2605), 2579–2605.

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, *42*, 60–88.

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd international conference on neural information processing systems* (pp. 13–23).

Lu, M., Huang, Z., Li, B., Zhao, Y., Qin, Z., & Li, D. (2022). SIFTER: A framework for robust rumor detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 429–442.

Luvembe, A. M., Li, W., Li, S., Liu, F., & Xu, G. (2023). Dual emotion based fake news detection: A deep attention-weight update approach. *Information Processing & Management*, *60*(4), Article 103354.

Min, E., Rong, Y., Bian, Y., Xu, T., Zhao, P., Huang, J., & Ananiadou, S. (2022). Divide-and-conquer: Post-user interaction network for fake news detection on social media. In *Proceedings of the ACM web conference 2022* (pp. 1148–1158).

Nan, Q., Cao, J., Zhu, Y., Wang, Y., & Li, J. (2021). MDFEND: Multi-domain fake news detection. In *Proceedings of the 30th ACM international conference on information and knowledge management* (pp. 3343–3347).

Popat, K., Mukherjee, S., Strötgen, J., & Weikum, G. (2016). Credibility assessment of textual claims on the web. In *Proceedings of the 25th ACM international on conference on information and knowledge management* (pp. 2173–2178).

Qian, S., Wang, J., Hu, J., Fang, Q., & 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* (pp. 153–162).

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. CoRR abs/2003.08271. arXiv:2003.08271.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE conference on computer vision and pattern recognition* (pp. 815–823).

Sculley, D. (2010). Web-scale k-means clustering. In *Proceedings of the 19th international conference on world wide web* (pp. 1177–1178).

Sharma, K., Qian, F., Jiang, H., Ruchansky, N., Zhang, M., & Liu, Y. (2019). Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology*, *10*(3).

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22–36.

Silva, A., Luo, L., Karunasekera, S., & Leckie, C. (2021). Embracing domain differences in fake news: Cross-domain fake news detection using multi-modal data. In *Thirty-Fifth AAAI conference on artificial intelligence, AAAI 2021, thirty-third conference on innovative applications of artificial intelligence, IAAI 2021, the eleventh symposium on educational advances in artificial intelligence* (pp. 557–565).

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *3rd International conference on learning representations*.

Singhal, S., Kabra, A., Sharma, M., Shah, R. R., Chakraborty, T., & Kumaraguru, P. (2020). SpotFake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *The thirty-fourth AAAI conference on artificial intelligence* (pp. 13915–13916).

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

Song, C., Shu, K., & Wu, B. (2021). Temporally evolving graph neural network for fake news detection. *Information Processing & Management*, *58*(6), Article 102712.

Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2020). VL-BERT: Pre-training of generic visual-linguistic representations. In *International conference on learning representations*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 6000–6010).

Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & 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* (pp. 849–857).

Wang, Y., Qian, S., Hu, J., Fang, Q., & Xu, C. (2020). Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 international conference on multimedia retrieval* (pp. 540–547).

Wang, T., Xu, X., Yang, Y., Hanjalic, A., Shen, H. T., & Song, J. (2019). Matching images and text with multi-modal tensor fusion and re-ranking. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 12–20).

Wang, S., Xu, X., Zhang, X., Wang, Y., & Song, W. (2022). Veracity-aware and event-driven personalized news recommendation for fake news mitigation. In *Proceedings of the ACM web conference 2022* (pp. 3673–3684).

Wei, Z., Pan, H., Qiao, L., Niu, X., Dong, P., & Li, D. (2022). Cross-modal knowledge distillation in multi-modal fake news detection. In *ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing* (pp. 4733–4737).

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., ... Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: System demonstrations* (pp. 38–45).

Wu, Y., Zhan, P., Zhang, Y., Wang, L., & Xu, Z. (2021). Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021* (pp. 2560–2569).

Xu, W., Wu, J., Liu, Q., Wu, S., & Wang, L. (2022). Evidence-aware fake news detection with graph neural networks. In *Proceedings of the ACM web conference 2022* (pp. 2501–2510).

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

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems* (pp. 5753–5763).

Yang, S., Feng, D., Qiao, L., Kan, Z., & Li, D. (2019). Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 5284–5294).

Yang, X., Lyu, Y., Tian, T., Liu, Y., Liu, Y., & Zhang, X. (2020). Rumor detection on social media with graph structured adversarial learning. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence* (pp. 1417–1423).

Yu, C., Ma, Y., An, L., & Li, G. (2022). BCMF: A bidirectional cross-modal fusion model for fake news detection. *Information Processing & Management*, *59*(5), Article 103063.

Zhang, H., Fang, Q., Qian, S., & Xu, C. (2019). Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM international conference on multimedia* (pp. 1942–1951).

Zhang, Q., Yang, Y., Shi, C., Lao, A., Hu, L., Wang, S., & Naseem, U. (2023). Rumor detection with hierarchical representation on bipartite ad hoc event trees. *IEEE Transactions on Neural Networks and Learning Systems*, 1–13.

Zheng, J., Zhang, X., Guo, S., Wang, Q., Zang, W., & Zhang, Y. (2022). MFAN: Multi-modal feature-enhanced attention networks for rumor detection. In *Proceedings of the thirty-first international joint conference on artificial intelligence* (pp. 2413–2419).

Zhou, X., Wu, J., & Zafarani, R. (2020). SAFE: Similarity-aware multi-modal fake news detection. In *Advances in knowledge discovery and data mining* (pp. 354–367).

Zhu, Y., Sheng, Q., Cao, J., Nan, Q., Shu, K., Wu, M., Wang, J., & Zhuang, F. (2022). Memory-guided multi-view multi-domain fake news detection. *IEEE Transactions on Knowledge and Data Engineering*, 1–14.