ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240935

Semantic Similarity Driven Multi-Modal Model for Rumor Detection

Chenyang Li[a], Bo Xu[b], Meng Wang[a] and Kun He[a,*]

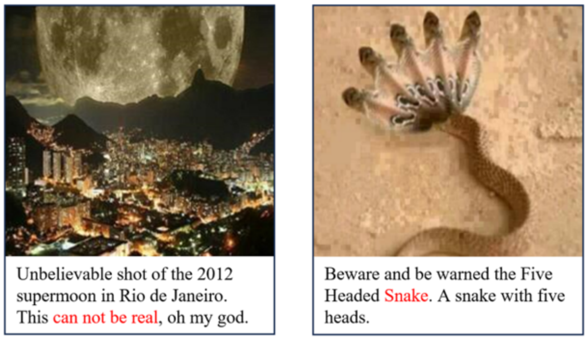
[a]Huazhong University of Science and Technology
[b]Dalian University of Technology

Abstract. The wide spread of rumors with images and texts on social media has attracted broad attention in the academy and industry. Existing models focus on utilizing powerful feature extractors to obtain multi-modal features and introducing various external knowledge. However, the intrinsic semantic similarity of different modalities is either simply ignored in most models or far from adequate in others. The insufficiency of semantic similarity information suppresses the potential of rumor detection models severely. To address this issue, we propose a novel model termed the Semantic Similarity driven Multi-modal model (SemSim) for rumor detection, which deeply captures the semantic similarity through more comprehensive fusion between different modalities and designs a new classification method consequently. Specifically, the proposed SemSim first integrates the raw image and raw text into a virtual image, which fuses information at a new view, *i.e.*, via the diffusion process inside stable diffusion models. Then SemSim captures the semantic similarity score between virtual image and raw image as the intrinsic information to drive SemSim. Besides, co-attention mechanism is employed to further perceive consistency and enhance interaction between the raw text-image pair. The fused representations via co-attention are utilized to evaluate the multi-modal feature score. In the end, SemSim balances the above two scores for final classification. Experiments on two typical real-world datasets show that SemSim can effectively detect rumors and outperform state-of-the-art methods.

1 Introduction

With the widespread utilization of social media platforms such as Twitter and Weibo, there has been a notable increase in the number of confusing rumors crafted by unscrupulous companies or individuals [26, 33, 41]. Deceptive rumors with vivid images and manipulative texts aim to distort and falsify facts, which may mislead the public [14, 31], posing a threat to the reliability and security of social media. Consequently, the effective detection of rumors from multimedia platforms has become critically important.

Figure 1 illustrates two instances of multi-modal rumors sourced from Twitter. In Figure 1(a), the moon appears unusually large, raising doubts about its authenticity. The accompanying text seems to describe super moon but provides a clue that "this cannot be real". On the other hand, in Figure 1(b), the text draws attention to a five headed snake, while the image merely shows sand arranged in the shape of a five-headed snake. From the rumor detection standpoint, the clue in the text and the anomaly in the image should be valued in Figure 1(a), whereas the inherent semantic inconsistency between the image and text should be captured in Figure 1(b). To mislead the public, rumors usually contain exaggerated text or attention-diverting fake images that lead to dissimilarity. Similarly, previous works have pointed out that multi-modal consistency information is beneficial for rumor detection [20, 26, 43, 35]. The above analysis motivates the idea to detect rumors from both information within each modality and semantic similarity between different modalities.


Unbelievable shot of the 2012 supermoon in Rio de Janeiro. This can not be real, oh my god.

Beware and be warned the Five Headed Snake. A snake with five heads.

(a) Super moon (b) Five headed snake

Figure 1: Two instances of multi-modal rumors with images and text from Twitter.

In recent years, the task of rumor detection has gained significant attention. Early deep neural network (DNN)-based models simply relied on textual information for rumor detection [15, 38, 16]. The reliance on textual cues alone limits the ability of these models to perceive visual information. To tackle this issue, various multi-modal models have been proposed for rumor detection. Wang *et al*. [32] exploit both visual and textual features and design an event discriminator for better rumor detection. Khattar *et al*. [10] introduce a VAE-based method to utilize both text and image information. While these models focus on incorporating auxiliary tasks, their performance is still unsatisfactory due to the lack of emphasis on the semantic similarity between textual and visual modalities and insufficient feature extraction. To address the need for extracting robust representations within each modality, various powerful models have been employed [6, 21, 37, 3]. Some researches leverage pre-trained models such as BERT [3] and XLNet [36] to obtain strong textual representations [24, 3, 21]. Some employ ResNet [7] and ViT [5] to obtain powerful visual representations [26, 13].

Others have also discovered that using graph neural networks like GCN [12] to capture social graph information can be advantageous for rumor detection [39]. Subsequently, various forms of external

[*] Corresponding Author. Email: brooklet60@hust.edu.cn

knowledge have been introduced to expand the horizon of the models, such as social context information [4, 42], user preference information [6] and historical information [23, 33]. However, in the real world, obtaining such external knowledge is often impractical due to limitations in data privacy and security. Even if the aforementioned information was available, the collection process would require extremely substantial manpower and material resources, which is usually unaffordable and limits the real-time capability of models severely. This challenge arises because these models rely on external knowledge rather than intrinsic information.

Some researchers have focused on exploring intrinsic information, primarily based on semantic similarity between different modalities [43, 10, 41]. And some achieve good performance without external knowledge [20, 26]. However, there is still room for improvement. These methods can be mainly categorized into four types:

- Co-attention mechanism-based methods. Most of recent models employ cross-modal co-attention mechanism to enhance interaction between different modalities [42, 9, 21, 20, 26]. While these methods effectively improve multi-modal features, the capture of semantic similarity is implicit and insufficient.
- Projection-based methods. Some methods [42, 34] project features of different modalities into a shared feature space through a linear layer with shared weights to capture similarity. However, these methods are limited due to the inherent inconsistency across different modalities [20]. Features extracted from various modalities often exhibit different distributions, and directly using these features may hinder similarity capture since they are not aligned.
- Image caption-based methods. Zhou *et al.* [43] use image caption to convert visual information into textual information to calculate similarity within a unified modality. However, this conversion process weakens the original visual information because its image2sentence model [29] is poor, leading to an inadequate capture of similarity. Zhang *et al.* [41] employ a large language model and reinforcement learning methods to get text-guided image captions, which is effective yet rather complex. Besides, the method focuses on image semantic enhancement, not similarity calculation.
- Image generation-based methods. Chakraborty *et al.* [2] and Li *et al.* [13] generate new images only based on the raw text to enrich data. While these methods have achieved good performance, it is possible to generate noise-like images due to semantically complex text. The absence of raw image misses general outline of the event and magnifies unimportant details such as color or size, resulting in limited similarity information. Moreover, raw text does not integrate with raw image until the diffusion process is over, missing a good fusion opportunity to capture the similarity.

In summary, the existing models face two primary challenges: (1) heavy reliance on external knowledge such as historical information and social context information; (2) insufficient methods for capturing intrinsic semantic similarity.

In this work, we aim to tackle the aforementioned challenges by proposing a novel model termed the **Sem**antic **Sim**ilarity driven multi-modal model (SemSim) for rumor detection. Specifically, we abandon external knowledge and deeply capture the semantic similarity through more comprehensive fusion. We first fuse raw text-image pair into a virtual image via the diffusion process, a new perspective for powerful fusion, and avoid the inherent inconsistency across modalities via virtual and raw images. In computer vision, recent research has found the potential of Stable Diffusion [22] to obtain consistency [8]. If raw image and text are semantically inconsistent like most rumors, the virtual image will differ from raw image thanks to Stable Diffusion, converting semantic dissimilarity between raw image and raw text into dissimilarity between raw and virtual images. Unfortunately, this is ignored by existing rumor detection models. To quantify the semantic text-image similarity, we calculate similarity scores between representations of raw image and virtual image, which serve as intrinsic information driving SemSim.

To further enhance the model, we extract features from the input image and text separately, and employ co-attention to improve representations and further perceive similarity. Based on improved representations, multi-modal feature scores are evaluated. Finally, we balance the semantic similarity score and multi-modal feature score for classification. Though some confusing rumors may have decent similarity scores, low multi-modal feature scores allow for correct detection. As a result, rumors whose images and text are similar such as Figure 1(a) or inconsistent such as Figure 1(b) can both be detected under the balance of the two scores.

The main contributions of our work are as follows:

- We propose a novel approach for rumor detection that emphasizes the capture and utilization of semantic similarity information between visual and textual modalities. Our method prioritizes the intrinsic relationship and alignment between different modalities while extracting multi-modal features in a unified frame.
- We utilize a new method to integrate information between different modalities for capturing similarity more comprehensively.
- We design a new classification method and a new loss function to take advantage of similarity information and achieve alignment.
- Extensive experiments conducted on typical real-world datasets demonstrate that our method is highly effective in identifying rumors and outperforms state-of-the-art rumor detection models.

2 Related Work

Early rumor detection methods [1, 30] only utilize basic textual information including upper and lower case characters, punctuation, and emotional keywords. These methods are too labor-intensive to be widely used. Hence, Ma *et al.* [15] first employ the neural networks, *i.e.*, different types of recurrent neural networks to automatically detect rumors based on textual information. Yu *et al.* [38] utilize convolutional neural network to extract textual features for classification. Ma *et al.* [16] combine the rumor detection task and the stance classification task to better detect rumors. Ma *et al.* [17] introduce an additional adversarial training process to enhance the model robustness. Tian *et al.* [27] utilize a novel signed attention mechanism to improve representations for rumor detection. Nan *et al.* [19] train the model in a multi-stage process to improve the cross-domain generalization capability. Sheng *et al.* [23] propose a model that benefits from perception of the environment. However, these models do not exploit visual features that might be beneficial.

To address this issue, various multi-modal models have been proposed to utilize textual and visual information. Jin *et al.* [9] first utilize cross-modal co-attention for multi-modal rumor detection. In this work, authors extract visual and textual features, respectively. Then the co-attention mechanism is employed to get enhanced features for final detection. Since then, co-attention has been popular. Wang *et al.* [32] design an event discriminator as an auxiliary task to better detect rumors. Some researches utilize BERT-based models to get strong representations [24, 37, 21]. Sun *et al.* [26] combine BERT, ViT, and graph convolution to further improve representations, which is effective but consumes too much memory. Some researches find using graph neural networks to capture social graph information is beneficial [39]. After that, various kinds of external knowledge, such as user preference and historical information are introduced to enhance the models [23, 4, 33]. Unfortunately, these models strongly rely on external knowledge. Sometimes they are not practical because the external knowledge is unavailable or hard to get. Such reliance also limits the real-time capability. As for usage of intrinsic information, as analyzed in Section 1, it is insufficient.

To sum up, existing models suffer from two major issues: (1) the strong reliance on external knowledge rather than intrinsic information; (2) insufficient methods to capture intrinsic semantic similarity. And this work addresses these problems effectively.

3 Methodology

Let *P* = $\{p_1, p_2, ..., p_n\}$ be a set of multimedia posts on social media, where each post $p_i = \{t_i, v_i\}$ includes text t_i and image v_i and has a corresponding ground-truth label $y \in \{0, 1\}$, where $y = 1$ indicates rumor, and 0 indicates non-rumor. The rumor detection task can be formulated as a binary classification task, whose goal is to learn a function $F(t_i, v_i)$ that best approximates the true label y_i for each post p_i.

For each post, we first employ stable diffusion model [22] to generate a virtual image v_i' based on t_i and v_i to have a rich fused visual information. Then the cosine similarity between representations of v_i' and v_i is calculated to measure their semantic similarity for better performance. In this way, we change the goal of rumor detection to learning a function $F(t_i, v_i, v_i')$ that best approximates the true label y_i for each post p_i. In the following, we present our method in detail.

3.1 Model Overview

In this work, we propose a Semantic Similarity driven Multi-modal model (SemSim) for rumor detection, whose architecture is shown in Figure 2. SemSim consists of three modules for exploiting multi-modal features and capturing semantic similarity between different modalities. (1) The multi-modal feature extraction module, which consists of textual and visual feature extractors (marked in green and blue, respectively). (2) The multi-modal information fusion module, which consists of two submodules (marked in orange). The Stable Diffusion submodule generates a virtual image based on the raw image and the prompt obtained by preprocessing the raw text. The Co-attention submodule promotes the interaction between different modalities to enhance the multi-modal features. (3) The multi-modal rumor detection module, which calculates the semantic similarity score and multi-modal feature score, and weights the above two scores for final classification (marked in yellow). Details of these modules are provided in the following subsections.

3.2 Multi-modal Feature Extraction

3.2.1 Textual Representation

We employ Text-CNN to extract the textual features for each post. For each post p_i, assume its text t_i is first padded or truncated to have n tokens, *i.e.* $t_i = \{w_1, w_2, ..., w_n\}$, where w_j denotes the *j*-th word in the text. Next, we feed the pre-trained word vectors to get the initial word embeddings $e^{t_i} = \{e^{w_1}, e^{w_2}, ..., e^{w_n}\} \in \mathbb{R}^{n \times d}$, where d is the embedding dimension. Then a feature map is obtained by a convolution layer, represented as:

$$r_i = \{r_i^1, r_i^2, \ldots, r_i^{n-k+1}\}, \tag{1}$$

where r_i^j represents the result of the convolution operation on $e^{w_{j:j+k-1}} = \{e^{w_j}, e^{w_{j+1}}, ..., e^{w_{j+k-1}}\}$ and k represents the kernel size. Then, the max pooling operation is applied over the obtained feature map r_i for dimension reduction, *i.e.*, $\tilde{r}_i = max\,(r_i)$. Therefore, a filter can output a value $\tilde{r}_i$, and we use a total of d filters with varying convolutional kernel sizes to obtain the semantic features from different views. Finally, we concatenate the results from all the filters to form the overall textual representation $t_r^i \in \mathbb{R}^d$.

3.2.2 Visual Representation

For each post p_i, the pre-trained model ResNet50 [7] with an additional fully connected layer is utilized to capture the visual features of the raw image v_i and the generated virtual image v_i'. The following is the procedure to get representation v_r^i for v_i. First, we extract the hidden features from the second last layer of ResNet50, denoted as v_h^i. Then, the intermediate representation v_h^i is fed into a fully connected layer with an activation function for dimension projection. The output $v_r^i \in \mathbb{R}^d$ is in the same dimension of text feature t_r^i:

$$v_r^i = \sigma(W_v * v_h^i + b_v), \tag{2}$$

where W_v and b_v are trainable matrices in the fully connected layer and $\sigma(\cdot)$ denotes the ReLU activation function. The approach to extract v_v^i for image v_i' is similar to the above procedure. We also extract the hidden features of ResNet50 to pass to the fully connected layer for obtaining the final representation $v_v^i \in \mathbb{R}^d$.

3.3 Multi-modal Information Fusion

3.3.1 Stable Diffusion for Information Fusion

Previous rumor detection models mostly fuse information between different modalities via concatenation or co-attention. We utilize the diffusion process inside Stable Diffusion, a novel view, to integrate information between image v_i and text t_i. The approach unifies information from textual and visual modality into visual modality, overcoming the inherent inconsistency across modalities and deeply mining semantic similarity information. Specifically, a new image v_i' can be obtained by:

$$v_i' = SD_\delta(v_i,\ prompt), \tag{3}$$

where *SD* represents the stable diffusion model, $prompt$ means the guiding description obtained by preprocessing raw content of t_i to modify v_i to v_i', and $\delta \in (0, 1)$ denotes the guiding strength of $prompt$ in generating v_i'. The parameter δ plays an important role in information fusion. A larger value indicates paying less attention to v_i while a small value pays less attention to the guiding $prompt$.

The information fusion process inside stable diffusion models can be divided into the forward process which adds noise and the backward process for denoising, sampling and generating virtual images. The forward process can be represented as:

$$noise\ =\ SD_1\,(v_i\ ,\ z)\,, \tag{4}$$

where SD_1 denotes the iterative process to add noise, z indicates the gradually added noise in Gaussian distribution, and $noise$ is nearly an isotropic pure noise. During the iterations of gradually transforming from v_i to $noise$, a series of noise is obtained serving as labels. The potential way of adding noise is learned to facilitate sampling new images during denoising. The reverse denoising process can be formalized as:

$$v_i' = SD_2\,(noise\ ,\ prompt)\,. \tag{5}$$

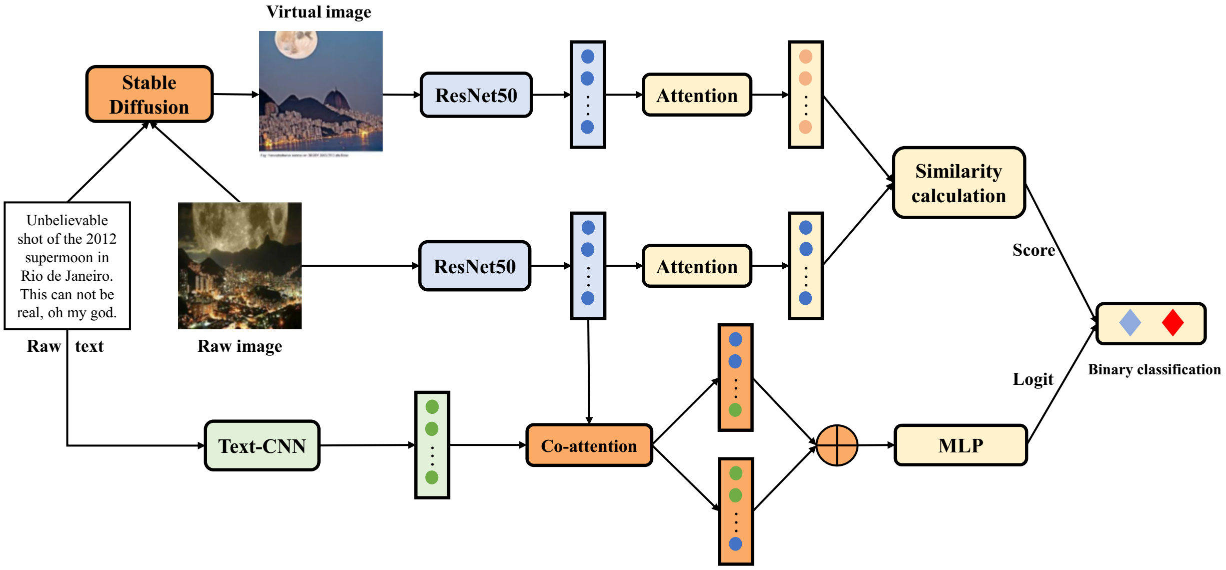


Figure 2: The overall architecture of our SemSim model. For each post, we first generate a virtual image based on the raw image and the prompt obtained from the raw text. Then we obtain representations of the virtual image, raw image, and raw text through feature extractors. We use multi-head self-attention to enhance representations of the two images, and calculate their similarity score. Meanwhile, the cross-modal co-attention mechanism is also used to fuse information and gain stronger representations between the textual and visual modalities of the input, which are then passed to a Multi-Layer Perceptron (MLP) to get a multi-modal feature score at the logit layer. In the end, the logit score and semantic similarity score are weighted for the final classification.

This process starts with noises and uses labels containing visual modality information obtained from the iterative SD_1. Gradually, SD_2 incorporates the textual information included in *prompt* and the visual information included in a series of noise to ultimately generate new virtual images.

The stable diffusion model's operation of integrating textual and visual information can be intuitively explained as: the generated virtual image v_i' fuses the semantic consistency information of the raw image v_i and text t_i. If v_i and t_i are semantically inconsistent, v_i' will differ from v_i to a relatively large extent. The dissimilarity between v_i and v_i' can be suspicious, which provides a useful clue to SemSim to detect rumors.

3.3.2 Co-attention Mechanism for Fusion

Based on the multi-head self-attention mechanism, Jin *et al.* [9] proposed the Co-Attention mechanism, a variant of attention to adapt to multi-modal situations:

$$Q_v^i = v_r^i W^Q, K_t^i = t_r^i W^K, V_t^i = t_r^i W^V, \tag{6}$$

$$Z_{vt}^i = MA(Q_v^i, K_t^i, V_t^i), \tag{7}$$

$$MA(Q, K, V) = (\overset{H}{\underset{h=1}{||}} \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V)W^o, \tag{8}$$

where $W^Q, W^K, W^V \in \mathbb{R}^{d \times \frac{d}{H}}$ are trainable projection matrices, $||$ denotes the concatenation operation, H indicates the number of heads, $d_k = d/H$ is the last dimension of the key matrix K, $W^o \in \mathbb{R}^{d \times d}$ is the output linear transformation, and Z_{vt}^i is the enhanced textual representation with visual information. Similarly, we can exchange v_r^i and t_r^i to get Z_{tv}^i, the enhanced visual representation with textual information. In this way, we fuse textual and visual representations obtained by feature extractors for better performance. Finally, we concatenate these two representations to further fuse the information:

$$Z^i = Z_{vt}^i \oplus Z_{tv}^i, \tag{9}$$

where $\oplus$ denotes the concatenation and Z^i is the final multi-modal representation for post p_i.

3.4 Multi-modal Rumor Detection

3.4.1 Obtaining Similarity Scores

Here we present a method to calculate the semantic similarity score between features of the raw image and virtual image, *i.e.*, v_r^i and v_v^i. Each pair of v_r^i and v_v^i is passed to the fully connected layer, followed by multi-head self-attention for enhancing the features.

$$\tilde{v}_r^i = W_{fc_1} * v_r^i, \ \tilde{v}_v^i = W_{fc_2} * v_v^i, \tag{10}$$

$$Q^i = \tilde{v}_r^i W^Q, K^i = \tilde{v}_r^i W^K, V^i = \tilde{v}_r^i W^V, \tag{11}$$

$$Q_v^i = \tilde{v}_v^i W^Q, K_v^i = \tilde{v}_v^i W^K, V_v^i = \tilde{v}_v^i W^V, \tag{12}$$

$$R^i = MA(Q^i, K^i, V^i), R_v^i = MA(Q_v^i, K_v^i, V_v^i), \tag{13}$$

where W_{fc_1} and W_{fc_2} represent weight matrices, and Q^i, K^i and V^i are query, key and value matrix for the raw image while Q_v^i, K_v^i and V_v^i are those for the generated virtual image. MA is the multi-head self-attention mechanism shown in Eq. 8. R^i and R_v^i are enhanced features for raw and virtual image, respectively.

Then, we utilize a modified cosine function [43] to measure the semantic consistency:

$$Sim(R^i, R_v^i) = \frac{R^i \cdot R_v^i}{2\|R^i\| \cdot \|R_v^i\|} + \frac{1}{2}, \tag{14}$$

where $Sim(R^i, R_v^i) \in [0, 1]$ is the similarity score revealing the relevance. We regard $1 - Sim(R^i, R_v^i)$ as the probability that p_i is a rumor solely from the perspective of semantic similarity. Specifically, if raw image and raw text do not match at all semantically, the generated image will be much different from the raw image, leading to a very low similarity score. On the contrary, if raw image and raw text are consistent, the similarity score will exceed the borderline of 0.5. This result is consistent with human's empirical knowledge that rumors are more likely to have unmatched image and text. Compared to existing researches that ignore the semantic consistency or simply project features from different modalities into a high-dimensional space to measure similarity, our approach of obtaining the similarity scores is more reasonable and effective.

3.4.2 Detecting Rumors with Similarity Scores

Almost all the previous works on rumor detection directly perform binary classification task based on the fused features. In contrast, we propose a new classification method called Classification With Similarity (CWS), which takes both multi-modal representations and the semantic similarity into consideration. We first evaluate the original classification score of multi-modal features Z^i, denoted as:

$$logit_i = MLP(Z^i), \tag{15}$$

where MLP represents a Multilayer Perception consisting of two fully connected layers with a ReLU function. The output $logit_i$ is in two dimensions, with $logit_i[1]$ the initial score as rumor and $logit_i[0]$ the initial score as non-rumor. Then, we mix up the above multi-modal feature score and similarity score as follows:

$$logit_i'[0] = logit_i[0] + \varphi * Sim(R^i, R_v^i), \tag{16}$$

$$logit_i'[1] = logit_i[1] + \varphi * (1 - Sim(R^i, R_v^i)), \tag{17}$$

where φ is used to balance the two factors. $1 - Sim(R^i, R_v^i) \in [0, 1]$ denotes the semantic inconsistency score, and $logit_i'$ is the revised result with similarity score. Both $logit_i[1]$ and $1 - Sim(R^i, R_v^i)$ are clues to categorize p_i as rumor. Then we calculate the predicted probability $\tilde{y}_i$ of p_i being a rumor and use the cross-entropy loss function as:

$$\tilde{y}_i = \text{softmax}(logit_i')[1], \tag{18}$$

$$L_{classify} = \sum_{i=1}^{n} -(y_i \log(\tilde{y}_i) + (1 - y_i)\log(1 - \tilde{y}_i)), \tag{19}$$

where y_i is the ground truth for p_i. Besides, we define a new loss function based on the MSE loss, which is designed for alignment:

$$L_{similarity} = \frac{1}{n}\sum_{i=1}^{n}(1 - Sim(R^i, R_v^i) - y_i)^2. \tag{20}$$

Here we provide an example to better understand the defined loss function. Assuming there is a rumor whose label is 1, we pass the rumor to our model to calculate the similarity score. If $Sim(R^i, R_v^i)$ is very low, it means our model strongly believes that the post is exactly a rumor. According to Eq. 20, the defined loss for this post is also low, which rewards our model for clear classification. However, if $Sim(R^i, R_v^i)$ is very high, it will become a punishment to our model, and vice versa. The final loss can be written as:

$$L = \lambda_c L_{classify} + \lambda_s L_{similarity}, \tag{21}$$

where λ_c and λ_s are used to balance the two losses.

Table 1: Statistics of the datasets.

Dataset	Non-rumors	False Rumors	Images	Comments
PHEME	1428	590	2018	7388
Weibo	1460	1127	2587	4534

3.4.3 Incorporating Adversarial Training

Previous works [42] have shown that, the PGD (Projected Gradient Descent) method [18], a widely used adversarial training technique, helps enhance the model's robustness at the text embedding level during training. Hence, we also incorporate the PGD Adversarial Training to enhance the representations. Specifically, at each training iteration, we calculate the gradient for the textual features and utilize it to compute the adversarial perturbation that is applied to the text embedding. Then the gradient is recalculated based on the updated textual features. Such process is repeated for m times, and we confine the extent of perturbation within a spherical space. Finally, we accumulate the adversarial gradients with the original gradient, which is then used for the parameter updates.

4 Experiments

In this Section, we conduct extensive experiments to demonstrate the effectiveness and practicality of our motivation to deeply capture semantic similarity information for better detecting rumors. We introduce datasets, baselines and implementation details in Section 4.1to Section 4.3. In Section 4.4 and Section 4.5, we compare our SemSim with baselines and recently popular large language models. We conduct ablation experiments in Section 4.6 to analyze the importance of different components in SemSim. In Section 4.7 and Section 4.8, we respectively conduct sensitivity analysis and individual case analysis.

4.1 Datasets

We evaluate the proposed SemSim on two typical real-world datasets for rumor detection: Pheme [44] and Weibo [25]. The Pheme dataset consists of tweets from the Twitter platform based on five breaking news. The Weibo dataset is collected from Weibo platform. In this work, we focus on text, images, and the semantic similarity information between them, and some baselines need comments for tweets. Following [42], we remove tweets that do not have text or image. And we remove tweets with only retweet information. The detailed statistics of the two datasets after removing are listed in Table 1.

4.2 Baselines

We consider the following baselines, which have covered the whole four types of methods mentioned in Section 1. EANN [32] is a GAN-based model to mine event-invariant features. MVAE [10] learns shared representations of texts and images using a variational autoencoder. QSAN [27] uses signed attention mechanism to enhance text encoding. SAFE [43] jointly exploits multi-modal features and similarity of text and images to learn representations. SpotFake [24] uses BERT and VGG-19 to extract textual and visual features respectively, then concatenates them for classification. HMCAN [21] uses cross-modal co-attention to fuse features of BERT every four layers with image features. M^{3}APS [13] generates images only based on raw text and generates text via image caption to enrich data. MFAN [42] combines textual, visual, and social graph features in a multi-modal feature-enhanced attention network to effectively detect rumors. However, none of them fuse information via the diffusion process and make full use of semantic similarity as we do.

Table 2: Comparison results of SemSim and the baselines.

	PHEME				Weibo			
Method	Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score
EANN	77.13±0.96	71.39±1.07	70.07±2.19	70.44±1.69	80.96±2.26	80.19±2.37	79.68±2.46	79.87±2.40
MVAE	77.62±0.64	73.49±0.81	72.25±0.90	72.77±0.81	71.67±0.89	70.52±0.95	70.21±1.01	70.34±0.98
QSAN	75.13±1.19	69.97±2.03	65.80±1.72	66.87±1.70	71.01±1.81	71.02±0.95	67.54±3.27	67.58±3.59
SAFE	81.49±0.84	79.88±1.22	79.50±0.81	79.68±0.70	84.95±0.85	84.98±0.82	84.95±0.91	84.96±0.86
SpotFake	84.42±0.95	80.71±0.99	83.72±1.87	82.19±1.46	85.13±1.88	85.18±1.97	84.98±2.14	85.08±1.95
HMCAN	87.44±0.72	85.02±0.71	84.08±0.93	84.49±0.81	89.04±0.97	89.07±1.02	88.42±0.83	88.67±0.95
M^{3}APS	88.01±0.86	85.99±0.87	84.55±1.01	85.27±0.94	88.12±0.77	87.85±0.91	87.63±0.78	87.65±0.87
MFAN	88.62±0.63	87.01±0.58	85.72±0.97	86.28±0.65	88.75±1.10	88.74±1.49	88.22±1.47	88.24±1.48
SemSim	**89.94±0.89**	**88.77±0.98**	**86.56±1.17**	**87.61±0.93**	**90.97±0.98**	**90.92±0.88**	**89.83±1.36**	**90.31±0.98**

4.3 Implementation Details

Following [42], we split datasets for training, validating, and testing with a ratio of 7:1:2. We use accuracy, precision, recall and F1 as evaluation metrics. The Adam optimizer [11] with a learning rate of 2e-3 is employed to optimize trainable parameters. The batch size is set to 64. We use word vectors in [39] as initial word embeddings. The number of heads H is set to 8, the scaling factor $\varphi = 3$, and λ_c and λ_s are set to 1 and 0.4, respectively. We perform five runs throughout all experiments and report the average results with standard deviation. The version of stable diffusion model is stable-diffusion-2. The guiding strength δ is set to 0.8. The guiding $prompt$ is the core information automatically extracted from raw text by GPT-3.5, to satisfy constraints of stable diffusion model on token length, improve the information density of prompts and reduce random noise from calls for retweeting and other irrelevant tags [13]. The input to GPT begins with "I hope you are a core information extractor. Please extract the core information of following sentences".

4.4 Comparisons with the Baselines

We compare the performance of SemSim with up to eight competitors, covering the whole four types of methods mentioned in Section 1. In Table 2, we report the comparison results. Our model outperforms all the baselines in terms of all metrics on both datasets, demonstrating its effectiveness and superior performance.

Utilizing similarity explicitly improves the performance of SAFE significantly, compared with MVAE and EANN. However, poor multi-modal representations make SAFE obviously inferior to SemSim. Benefiting from fully utilizing multi-modal co-attention every four hidden layers, HMCAN enhances multi-modal features and shows good performance. However, consistency information can be too implicit, making it inferior to SemSim. Although MFAN and M^{3}APS are recent advanced models for rumor detection, our proposed SemSim still beats them. M^{3}APS enriches information via image caption and image generation based on only text. Besides, both M^{3}APS and MFAN rely on the social context information such as retweet or response, which we abandon for real-time capability. Even in this condition, the fine-grained fusion via the diffusion process and the full use of semantic similarity between raw text-image pairs empower our model to win. To sum up, our SemSim shows the best detecting performance, because it benefits from capturing and utilizing semantic consistency more adequately, and obtaining excellent multi-modal features.

4.5 Further Comparison with LLMs

To further check the effectiveness of SemSim, we conduct a preliminary comparison with recently popular Large Language Models (LLMs). We compare SemSim with two open source large models, *i.e.*, chatglm-6b [40] and llama-2-13b [28]. We regret not to use raw images because the above two large models only focus on text. If we directly feed the posts into LLMs and ask them to detect rumors, they will refuse and reply "As an AI language model, I cannot perform real-time fact-checking or verify the specific post for its truthfulness or rumor status". To handle this problem, we obtain robust textual embeddings via the large model and add an MLP as a classifier to detect rumors. We only train parameters in the classifier while freezing parameters inside LLMs due to resource limitation. Finally, we evaluate the performance on the test dataset.

Table 3: Preliminary comparison results of SemSim with LLMs.

Method		Accuracy	Precision	Recall	F1 score
PHEME	chatglm-6b	78.62	74.21	72.83	73.58
	llama-2-13b	79.44	75.62	73.11	74.19
	SemSim	**89.94**	**88.77**	**86.56**	**87.61**
Weibo	chatglm-6b	83.73	83.17	84.22	83.41
	llama-2-13b	81.10	81.17	80.97	80.12
	SemSim	**90.97**	**90.92**	**89.83**	**90.31**

As shown in Table 3, our model has achieved significantly better detection results on both datasets. These LLMs can be fine-tuned to gain better performance. Due to the limitations of experimental resources, we only do a preliminary comparison.

4.6 Ablation Study

To further explore the importance of different components in SemSim, we compare SemSim with its sub-models "-w/o SD", "-w/o CWS", and "-w/o simloss". They respectively denote the variants without stable diffusion, the proposed new classification method CWS, and the new loss function. Stable diffusion is a new strategy to fuse information for better capturing the semantic similarity while CWS and simloss are both strategies to utilize the semantic similarity. The comparison results are shown in Table 4.

We can observe all ablation variants perform significantly worse than complete SemSim on both datasets. Without stable diffusion, the captured semantic similarity is inadequate, leading to the significantly decreased performance. These results offer following insights: (i) capturing semantic similarity between different modalities is in great need; (ii) classification with similarity scores is effective; (iii) the new loss based on similarity helps with alignment.

In order to further verify the effectiveness of semantic similarity between images and text, we detect rumors only based on similarity scores. Specifically, this variant only uses components in the upper two lines in Figure 2, which we define as SemSim- for convenience. According to Table 5, although we only explore the semantic similarity scores, the performance of SemSim- is still competitive, further demonstrating that our motivation to deeply mine semantic similarity

Table 4: Comparison results of the variants of SemSim.

Method		Accuracy	Precision	Recall	F1 score
PHEME	-w/o SD	86.51	85.16	81.27	83.17
	-w/o CWS	86.75	85.07	82.09	83.65
	-w/o simloss	88.03	86.12	84.47	85.29
	SemSim	**89.94**	**88.77**	**86.56**	**87.61**
Weibo	-w/o SD	87.19	86.67	86.95	86.83
	-w/o CWS	87.66	86.72	87.99	87.62
	-w/o simloss	88.39	88.01	87.85	87.97
	SemSim	**90.97**	**90.92**	**89.83**	**90.31**

Table 5: Comparison results of SemSim- and the baselines.

Method		Accuracy	Precision	Recall	F1 score
PHEME	EANN	77.13	71.39	70.07	70.44
	MVAE	77.62	73.49	72.25	72.77
	QSAN	75.13	69.97	65.80	66.87
	SemSim-	**78.86**	**75.04**	**73.16**	**73.52**
Weibo	EANN	80.96	80.19	79.68	79.87
	MVAE	71.67	70.52	70.21	70.34
	QSAN	71.01	71.02	67.54	67.58
	SemSim-	**82.59**	**83.72**	**81.39**	**82.29**

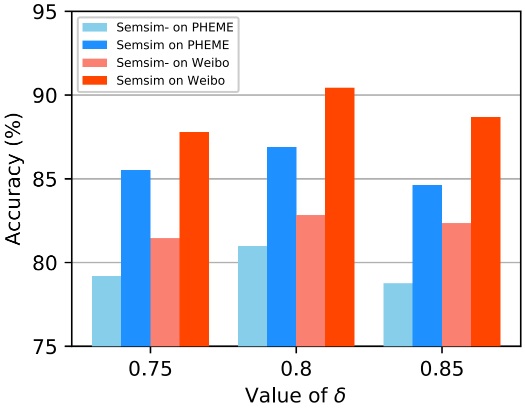
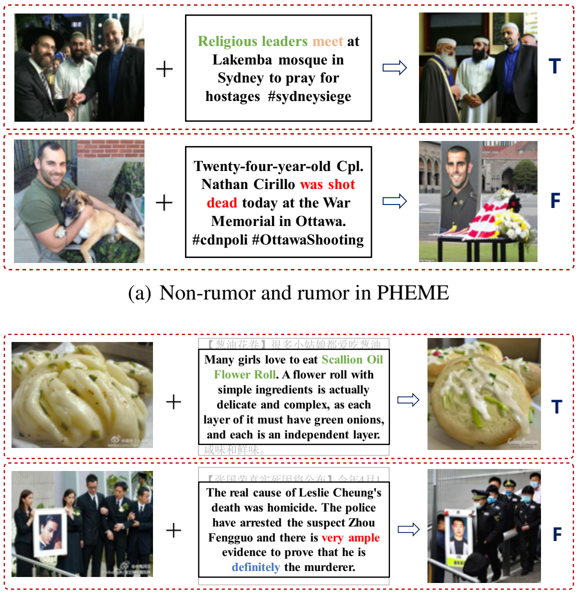


Figure 3: The influence of the hyper-parameter δ on our model SemSim and the variant SemSim-.


+
T
+
F
+
T
【葱油花卷】很多小姑娘都爱吃葱油
咸味和鲜味。
+
F
【张国荣真实死因将公布】今年4月1

Religious leaders meet at Lakemba mosque in Sydney to pray for hostages #sydneysiege

Twenty-four-year-old Cpl. Nathan Cirillo was shot dead today at the War Memorial in Ottawa. #cdnpoli #OttawaShooting

(a) Non-rumor and rumor in PHEME

Many girls love to eat Scallion Oil Flower Roll. A flower roll with simple ingredients is actually delicate and complex, as each layer of it must have green onions, and each is an independent layer.

The real cause of Leslie Cheung's death was homicide. The police have arrested the suspect Zhou Fengguo and there is very ample evidence to prove that he is definitely the murderer.

(b) Non-rumor and rumor in Weibo

Figure 4: Cases in PHEME and Weibo. Chinese text in Weibo is translated into English. 'T' denotes non-rumors while 'F' for rumors.

of raw text-image pairs is practical and effective.

4.7 *Sensitivity Analysis*

We conduct the sensitivity analysis to show the influence of stable diffusion parameter δ. The fusion effect of stable diffusion and the quality of generated images highly rely on δ, the guiding strength of the prompt, which is usually 0.8 in image generation [22]. To figure out the sensitivity of δ in our model, we set it to be 0.75, 0.8, and 0.85, and report the results of SemSim and its variant SemSim- on both datasets in Figure 3. As δ can only affect the generated images to contribute to the similarity scores without affecting the multi-modal feature scores, the results of SemSim- reveal the influence of δ on the pure similarity while results of SemSim reveal influence on final performance. Although each value of δ only differs by 0.05, there is already a considerable gap on both datasets, demonstrating the importance of the fusion process and the quality of generated images. We set $\delta = 0.8$ in our model, because both SemSim and SemSim- achieve the best accuracy on two datasets in this setting.

4.8 *Case Study*

As shown in Figure 4, some rumors and non-rumors are correctly classified under the balance of multi-modal feature score and similarity score. We preprocess raw text to obtain guiding prompts to reduce random noise. Given the prompt and raw image as input, a virtual image is generated. Specifically, in Figure 4(a), raw image (left) and text (middle) of the upper post are about "religious leaders". SemSim captures semantic similarity to generate a similar virtual image (right) and correctly classifies it as a non-rumor. For the lower rumor, the forged text contains "was shot dead" while the image does not contain this information. So the scene of memorial ceremony appears in virtual image. Such inconsistency and forgery are perceived by SemSim to detect this rumor. In Figure 4(b), image and text are both about "Scallion Oil Flower Roll" for the upper non-rumor and there is no suspicious clue. SemSim classifies it as a non-rumor correctly. For the rumor one, similarity score slightly tends to classify it as a non-rumor because both image and text seem to be about the death, while the exaggerated certainty in the text is suspicious, which is captured by multi-modal feature score. Under the balance of two scores, SemSim detects the rumor as expected. Though it is wrongly classified only from the perspective of similarity, the balance with feature score allows for detecting the rumor finally.

5 Conclusion

This paper proposed a novel model called SemSim (semantic similarity driven multi-modal model) for rumor detection. SemSim begins by generating a virtual image based on raw image and raw text, fusing information into visual modality via the diffusion process. Then the similarity score between raw and virtual images is calculated as intrinsic information to drive SemSim. Additionally, SemSim evaluates a multi-modal feature score based on raw text-image pair. Finally, SemSim combines the semantic similarity score and multi-modal feature score to make final classification. Evaluations and comparisons on two typical datasets for rumor detection demonstrate that our model outperforms the state-of-the-art baselines.

Our SemSim brings a new view to fuse information between different modalities for rumor detection. In the future, how to further explore this fusion method, *i.e.*, the diffusion process to enhance interaction for better detecting rumors could be a potential direction.

Acknowledgements

This work is supported by National Natural Science Foundation (62076105,U22B2017).

References

[1] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *Proceedings of the International Conference on World Wide Web*, 2011.

[2] M. Chakraborty, K. Pahwa, A. Rani, S. Chatterjee, and et al. Factify3m: A benchmark for multimodal fact verification with explainability through 5w question-answering, 2023.

[3] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[4] Y. Ding, B. Guo, Y. Liu, H. Wang, H. Shen, and Z. Yu. Piercingeye: Identifying both faint and distinct clues for explainable fake news detection with progressive dynamic graph mining. In *ECAI 2023 - 26th European Conference on Artificial Intelligence*, pages 549–556, 2023.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[6] Y. Dou, K. Shu, C. Xia, P. S. Yu, and L. Sun. User preference-aware fake news detection. In *The International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021.

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[8] S. Jia, M. Huang, Z. Zhou, Y. Ju, J. Cai, and S. Lyu. Autosplice: A text-prompt manipulated image dataset for media forensics. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 893–903, 2023.

[9] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the ACM Multimedia Conference*, pages 795–816, 2017.

[10] D. Khattar, J. S. Goud, M. Gupta, and V. Varma. MVAE: multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*, pages 2915–2921, 2019.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[12] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[13] B. Li, Z. Qian, P. Li, and Q. Zhu. Multi-modal rumor detection on modality alignment and multi-perspective structures. In *Advanced Intelligent Computing Technology and Applications*, pages 472–483, 2023.

[14] D. Li, J. Zhu, M. Wang, J. Liu, X. Fu, and Z.-J. Zha. Edge-aware regional message passing controller for image forgery localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8222–8232, 2023.

[15] J. Ma, W. Gao, P. Mitra, S. Kwon, and et al. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2016.

[16] J. Ma, W. Gao, and K. Wong. Detect rumor and stance jointly by neural multi-task learning. In *Companion of the The Web Conference*, 2018.

[17] J. Ma, W. Gao, and K. Wong. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The World Wide Web Conference*, pages 3049–3055, 2019.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[19] Q. Nan, D. Wang, Y. Zhu, Q. Sheng, Y. Shi, J. Cao, and J. Li. Improving fake news detection of influential domain via domain- and instance-level transfer. In *Proceedings of the International Conference on Computational Linguistics*, pages 2834–2848, 2022.

[20] P. Qi, J. Cao, X. Li, H. Liu, Q. Sheng, X. Mi, Q. He, Y. Lv, C. Guo, and Y. Yu. Improving fake news detection by using an entity-enhanced framework to fuse diverse multimodal clues. In *Proceedings of the ACM International Conference on Multimedia*, pages 1212–1220, 2021.

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

[22] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10674–10685, 2022.

[23] Q. Sheng, J. Cao, X. Zhang, R. Li, D. Wang, and Y. Zhu. Zoom out and observe: News environment perception for fake news detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4543–4556, 2022.

[24] S. Singhal, R. R. Shah, T. Chakraborty, P. Kumaraguru, and S. Satoh. Spotfake: A multi-modal framework for fake news detection. In *IEEE International Conference on Multimedia Big Data*, pages 39–47, 2019.

[25] C. Song, C. Yang, H. Chen, C. Tu, Z. Liu, and M. Sun. Ced: credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3035–3047, 2019.

[26] T. Sun, Z. Qian, P. Li, and Q. Zhu. Graph interactive network with adaptive gradient for multi-modal rumor detection. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 2023.

[27] T. Tian, Y. Liu, X. Yang, Y. Lyu, X. Zhang, and B. Fang. QSAN: A quantum-probability based signed attention network for explainable false information detection. In *The ACM International Conference on Information and Knowledge Management*, pages 1445–1454, 2020.

[28] H. Touvron, L. Martin, K. Stone, P. Albert, and et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[29] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. pages 652–663, 2017.

[30] A. Vlachos and S. Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the Workshop on Language Technologies and Computational Social Science*, pages 18–22, 2014.

[31] M. Wang, X. Fu, J. Liu, and Z.-J. Zha. Jpeg compression-aware image forgery localization. In *Proceedings of the ACM International Conference on Multimedia*, pages 5871–5879, 2022.

[32] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao. EANN: event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 849–857, 2018.

[33] T. Xiao, S. Guo, J. Huang, R. Spolaor, and X. Cheng. Hipo: Detecting fake news via historical and multi-modal analyses of social media posts. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 2805–2815, 2023.

[34] J. Xue, Y. Wang, Y. Tian, Y. Li, L. Shi, and L. Wei. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 2021.

[35] J. Xue, Y. Wang, Y. Tian, Y. Li, and L. Wei. Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 2021.

[36] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 2019.

[37] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Annual Conference on Neural Information Processing Systems 2019*, pages 5754–5764, 2019.

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

[39] C. Yuan, Q. Ma, W. Zhou, J. Han, and S. Hu. Jointly embedding the local and global relations of heterogeneous graph for rumor detection. In *IEEE International Conference on Data Mining*, pages 796–805, 2019.

[40] A. Zeng, X. Liu, Z. Du, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[41] Q. Zhang, J. Liu, F. Zhang, J. Xie, and Z.-J. Zha. Hierarchical semantic enhancement network for multimodal fake news detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 3424–3433, 2023.

[42] J. Zheng, X. Zhang, S. Guo, Q. Wang, W. Zang, and Y. Zhang. MFAN: multi-modal feature-enhanced attention networks for rumor detection. In *Proceedings of the International Joint Conference on Artificial Intelligence,*, pages 2413–2419, 2022.

[43] X. Zhou, J. Wu, and R. Zafarani. SAFE: similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on knowledge discovery and data mining*, pages 354–367, 2020.

[44] A. Zubiaga, M. Liakata, and R. Procter. Exploiting context for rumour detection in social media. In *Social Informatics: International Conference*, pages 109–123, 2017.