Focusing on Relevant Responses for Multi-modal Rumor Detection

Jun Li, Yi Bin, Liang Peng, Yang Yang, *Senior Member, IEEE*
Yangyang Li, Hao Jin, and Zi Huang

Abstract—In the absence of an official statement about a rumor, people may expose the truth behind such rumor through their responses on social media. Due to the varying relevance of responses in exposing hidden suspicious points within a rumor claim, it is crucial to prioritize those with higher relevance, rather than considering every responding tweets. As for the multi-modal rumor detection, an effective approach for evaluating relevance is aligning responses with the different modalities of the rumor claim in a fine-grained manner. However, owing to the substantial volume of response tweets, it is both costly and redundant to align all responses with the multi-modal claim. In this paper, we propose a novel two-stage model, termed *Focal Reasoning Model (FoRM)*, to select critical responses for multi-modal rumor detection. More specifically, our FoRM consists of two primary elements: coarse-grained selection and fine-grained reasoning. The coarse-grained selection component employs post-level features of responses to initialize a relevant score for each. Based on these scores, we preserve the responses with higher scores as the candidate ones for subsequent reasoning. Within the fine-grained reasoning component, we develop a relation attention module to investigate fine-grained relationships, specifically token-to-token and token-to-object connections, between the preserved responses and the multi-modal claim, with the goal of discovering valuable clues. Extensive experiments have been conducted on three real-world datasets, and the results demonstrate that our proposed model outperforms all the baselines.

Index Terms—Multi-modal Rumor Detection, Relevant Response, Fine-grained Relation.

1 Introduction

Social media has become a popular and important way for people to gather and share information. A recent survey points out that more than 70% Americans communicate with others and access news content via social media[1]. Such hyper-connected social network not only makes information spread faster but also provides an ideal environment for the spread of misinformation [1]. Among the information in circulation, rumor is the unverified information that could influence public decisions and further lead to social disruption. For instance, rumors would mislead the voters during the 2016 U.S. election [2] and affect the willingness of the public to receive the COVID-19 vaccine [3]. Therefore, it is necessary to detect rumors for providing a better network environment and decreasing the detrimental public effects.

The goal of rumor detection is to verify the truthfulness of a given claim. To achieve this goal, fact-checking websites attempt to invite domain experts to confirm the suspected claim. Such manual method is time-consuming and low-coverage, especially on social media, where produces huge amounts of data every day. In order to automatically verify a claim, some researchers propose to retrieve the relevant evidence from Wikipedia and extract key clues [4], [5], [6]. However, some claims on social media, especially breaking news, report newly emerged events that could not be confirmed from existing databases. These claims always attract public attention and provoke related discussions. Studies [7], [8], about the relevant discussion, discovered that the user responses would provide valuable clues for finding out the truth behind an unverified claim. Inspired

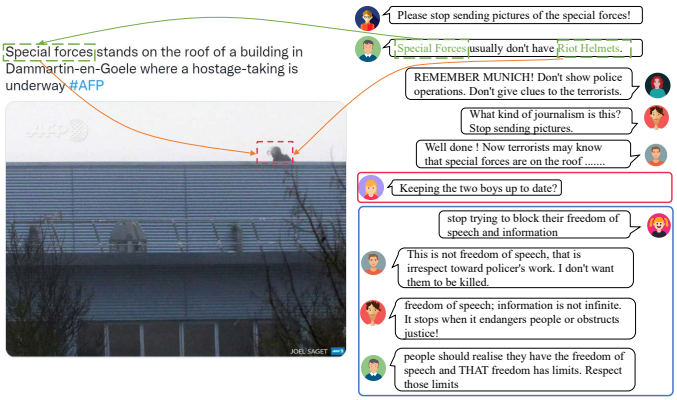

Fig. 1. A conversation constructed with a multi-modal claim and corresponding responses. The tweets in the red box and blue box are irrelevant and unhelpful for verifying the claim respectively. The valuable clues should be captured with the fine-grained reasoning of the responding tweets and the multi-modal claim.

- *Jun Li, Yi Bin, and Liang Peng are with the Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China.*
- *Yang Yang is with the Center for Future Media, and the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China , and also with the Institute of Electronic and Information Engineering, University of Electronic Science and Technology of China, Guangdong 523808, China.*
- *Yangyang Li and Hao Jin are with the National Engineering Research Center for Risk Perception and Prevention, CAEIT, Beijing 100041, China.*
- *Zi Huang is with the School of Information Technology and Electrical Engineering, The University of Queensland, Brisbane, Australia.*
- *Corresponding author: Yang Yang. E-mail: dlyyang@gmail.com.*

Manuscript received April 19, 2005; revised August 26, 2015.

1. https://www.pewresearch.org/internet/fact-sheet/social-media/

by this discovery, many response-based methods have been proposed and achieved good detection accuracy in rumor detection [9], [10], [11], [12], [13], [14], [15], [16].

The majority of response-based methods concentrate on modeling textual information. Early research [9], [10], [11], [17] expended considerable effort on crafting features from social content, user characteristics, and information propagation to learn rumor-indicative clues. However, these hand-crafted features consist solely of statistical information, such as the number of words in a tweet and the fraction of tweets containing a question mark. They lack the semantic depth of textual content and fall short of fully representing the complicated rumor-indicative features. Recently, deep-learning approaches have been proposed to leverage the temporal [12], [13], [18], structural features [19], [20], [21] and content features [22] of user interactions within a conversation thread. Nevertheless, these methods overlook the verification of multi-modal rumor claims, which have already emerged as central elements within social media platforms [23], [24]. To utilize user responses in verifying the multi-modal claim, several multi-modal rumor detection methods [14], [15] are proposed. For instance, Zhang et al. [14] model the multi-modal claim and all responses to construct an event memory with event-invariant features.

Most multi-modal methods mainly fuse different modalities into a post-level representation and aggregate the claim with all the response tweets. Recently, several methods [25], [26] propose to leverage each modality of the multi-modal claim to fuse the responses/evidences with different attention scores. However, they both leverage the global representation to evaluate the relevance of each response, and ignore their fine-grained features. For a multi-modal rumor claim, the focus of a user may be on special words in the text content or an object in the image. Therefore, to identify the relevant responses that critical to verify multi-modal claim, an effective approach is to align responses with the different modalities of the rumor claim in a fine-grained manner. As shown in Fig. 1, based on the words *"special forces"* and the place *"on the roof of a building"* in the claim, we locate that the person in the image is a special force. The response, *"Special Forces usually don't have Riot Helmets"*, connecting to the claim with *"Special Forces"* is inconsistent with the state in the image (the special force wears a riot helmet), which demonstrates the claim is a false rumor. Without the reasoning by the fine-grained interactions, we could not locate the this relevant response. However, due to the substantial volume of response tweets, it is costly and redundant to model all the responses and the multi-modal claim in the fine-grained way, and then select the relevant ones. In Fig. 1, the response in the red box, *"Keeping the two boys up to date?"*, is irrelevant to the multi-modal claim. In addition, tweets in the blue box are talking about the freedom of speech, which may not provide any evidence about the truthfulness of the claim. It may be unnecessary to conduct fine-grained reasoning for these response tweets.

In this paper, we propose a novel two-stage framework, termed **Fo**cal **R**easoning **M**odel (**FoRM**), to verify the multi-modal rumor. More specifically, our FoRM mainly contains two main components: coarse-grained selection and fine-grained reasoning. The coarse-grained selection component employs post-level features of responses to initialize a relevance score for each. Based on these scores, we preserve the responses with higher scores as the candidate ones for subsequent reasoning. Within the fine-grained reasoning component, we develop a relation attention module to investigate fine-grained relationships, *i.e.*, token-to-token and token-to-object connections, between the preserved responses and the multi-modal claim, with the goal of discovering valuable clues. To improve the effectiveness of the coarse-grained selection, we firstly formulate a selection loss to supervise the coarse-grained selection component, aiming to maximize the probability of the presence of relevant responses within the candidate set. Then, we propose to jointly train these two components, which could refine the scores of the coarse-grained selection based on the feedback of the fine-grained reasoning. Extensive experiments have been conducted on three real-world datasets, and the results demonstrate that our proposed model surpasses all baseline models in performance.

The contributions of this paper are as follows:

- We propose a novel FoRM for multi-modal rumor detection, which could focus on the relevant responses by conducting the fine-grained reasoning of the responses and the multi-modal claim.
- FoRM is designed as the two-stage framework, consisting of coarse-grained selection and fine-grained reasoning, due to the substantial volume of response tweets. We train these two components in the end-to-end manner, which could refine the initial relevance scores of the coarse-grained selection based on the feedback of the fine-grained reasoning and further improve the effectiveness of these relevance scores.
- In coarse-grained selection, to maximize the probability of the presence of relevant responses within the candidate set, we formulate a selection loss to improve the effectiveness of the initial scores.
- Extensive experiments demonstrate that our FoRM could reach better performance and focus on the reasonable responses based on the multi-modal claim for detecting rumors.

2 Related Work

As anyone with an Internet-connected device could share what they may be witnessing or their real-time thoughts on social media [27], the truthfulness of this information is always uncertain. Such unverified information, termed the rumor, has the detrimental effect on society and individuals [28], [29]. Rumor detection aims to identify the veracity of this information and attracts extensive research attention in recent years. Based on the source of clues which are used to verify rumors, the proposed rumor detection methods could be categorized into three types: the claim-based, the fact-checking, and the response-based.

The claim-based methods attempt to find the clues from the claim, such as the inconsistency [23], [30], [31], [32] of different modalities or the event-invariant features [33], [34]. As attention mechanisms are effective in various multi-modal tasks [35], [36], [37], [38], Qian et al. [23] design a contextual attention network to encode the multi-modal context information hierarchically for verifying the truthfulness. Instead of utilizing images as text supplements, Qi et al. [31]

propose a fine-grained model to capture the inconsistency of the entities in text and image. Inspired by adversarial network [39], [40], [41] and domain adaption [42], [43], EANN [33] and MDDA [34] transfer the event-invariant features to newly emerging events.

The fact-checking methods reply on the authoritative sources. Fact-checking websites, such as politifact.com and snopes.com, invite domain experts to confirm the dubious claim. To keep up with the enormous volume of generated online information, Ciampaglia et al. [44] construct a public knowledge graph from Wikipedia for automatically checking claims. Recently, deep learning methods are applied to retrieve the evidence from trustworthy sources and infer the fact [4], [5], [6]. However, some claims, especially breaking news, on social media report newly emerged events which could not be checked from the existing database. For these claims, the fact-checking methods may not work.

The response-based methods actually derive from the crowd wisdom. Studies [7], [8], about the relevant discussion, discovered that the community response would provide valuable clues for finding out the truth behind an unverified claim. Therefore, how to utilize the abundant information of the community response is a key research direction. Early studies [9], [10], [11], [17], [45], [46] focus on extracting manual features from the contents of messages, user profiles, and diffusion patterns. For example, Castillo et al. [9] identify four types of hand-crafted features to characterize each topic and train a supervised classifier to debunk rumors. However, these methods heavily rely on intensive manual efforts and could not fully capture the complicated rumor-indicative features with such statistics. Then deep-learning methods are proposed to infer clues by modelling the timeline and structure of the user interactions. Timeline-based methods [12], [13], [18] concentrate on modelling the temporal pattern of user interactions. Ma et al. [12] leverage Recurrent Neural Networks (RNN) to capture the temporal feature of the sequential reply stream. To extract the high-level interaction of each part of the reply sequence, the CNN-based framework, CAMI, is proposed in [18]. In addition, several works [19], [47], [48] argue the user interactions are tree-structured and propose structure-based methods based on Tree-LSTM [49] and Transformer [50]. Specifically, for parallel processing, BCTree LSTM [47] rebuilds the conversation tree into a binary tree in that each node is always connected with two children. To extract better textual and structural features, Tree-Transformer [48] is proposed and achieves a better performance for conversation trees with different depths. Recent studies [20], [51], [52] attempt to explore both temporal and structural features. By embedding the time delay and structural information into the multi-head attention layer simultaneously, variants of Transformer [20] are proposed. In contrast, conversational-GCN [51] extracts structural and temporal features by GCN and RNN respectively in two steps. Lao et al. [52] design the non-linear structure learning and the linear sequence learning to explore the temporal and the hierarchical characteristic of the user interaction respectively. These approaches have shown promising performance on applying deep learning to rumor detection. However, these methods ignore the multi-modal clues of the thread conversation. Hence, multimodal fusion methods [14], [15], [25] are proposed to fuse the textual and visual features for detecting rumors. Zhang et al. [14] leverage the multi-modal information for a better post representation and build event memory with the event-invariant features of the sequential responding posts.

Instead of considering all the responding tweets in previous works, we propose to exploit valuable tweets from the relevant responses in the fine-grained way.

3 Problem Statement

In this paper, we employ the textual information of responding tweets to verify the corresponding multi-modal claim. The conversation thread of a given claim is defined as:

$$\begin{cases} X = \{s, R\}, \\ s = \{s_T, s_I\}, \\ R = \{r_1, r_2, \cdots, r_N\}, \end{cases} \tag{1}$$

where s is the claim, which consists of the text s_T and the image s_I, R is the group of the responding tweets which replies to the claim s. Hence, the rumor detection task is to calculate the probability $P(y|X, \theta)$, where θ is the parameter of rumor classifier and y is the rumor class label.

4 Method

In this section, we describe the proposed two-stage framework, *Focal Reasoning Model (FoRM)*, in detail. As illustrated in Fig. 2, our model consists of three main components: a feature encoder, coarse-grained selection, and fine-grained reasoning. The feature encoder utilizes pre-trained models to extract token-level and sentence-level features of the text, as well as object-level features of the image. Subsequently, the claim multi-modal representation is fused with the features of different modalities in the claim. The coarse-grained selection component employs the multi-modal representation and sentence-level features of responses to initialize a relevance score for each response. Based on the relevance scores, the top k responses are selected as the candidate ones. The fine-grained reasoning component leverages a relation attention module to model the token-to-token and token-to-object relations between the candidate responses and the claim for evaluating the relevance of preserved responses and finding out the valuable clues.

4.1 Feature Encoder

As the model pretrained in a large dataset typically catches the better semantic information, we leverage the pre-trained models as the textual encoder and the visual encoder, to extract the essential features of the text and the image.

4.1.1 Textual Encoder

Since BERT [53] has achieved great success in natural language processing tasks [54], [55], we apply BERT as the textual encoder. For coarse-grained selection and fine-grained reasoning, the main purpose of our textual encoder is to extract the coarse-grained (sentence-level) features and fine-grained (token-level) features of the text. We feed the claim text s_T and a responding tweet r_i to BERT separately:

$$\begin{aligned} H^s &= \text{BERT}(s_T), \\ H^i &= \text{BERT}(r_i), \end{aligned} \tag{2}$$

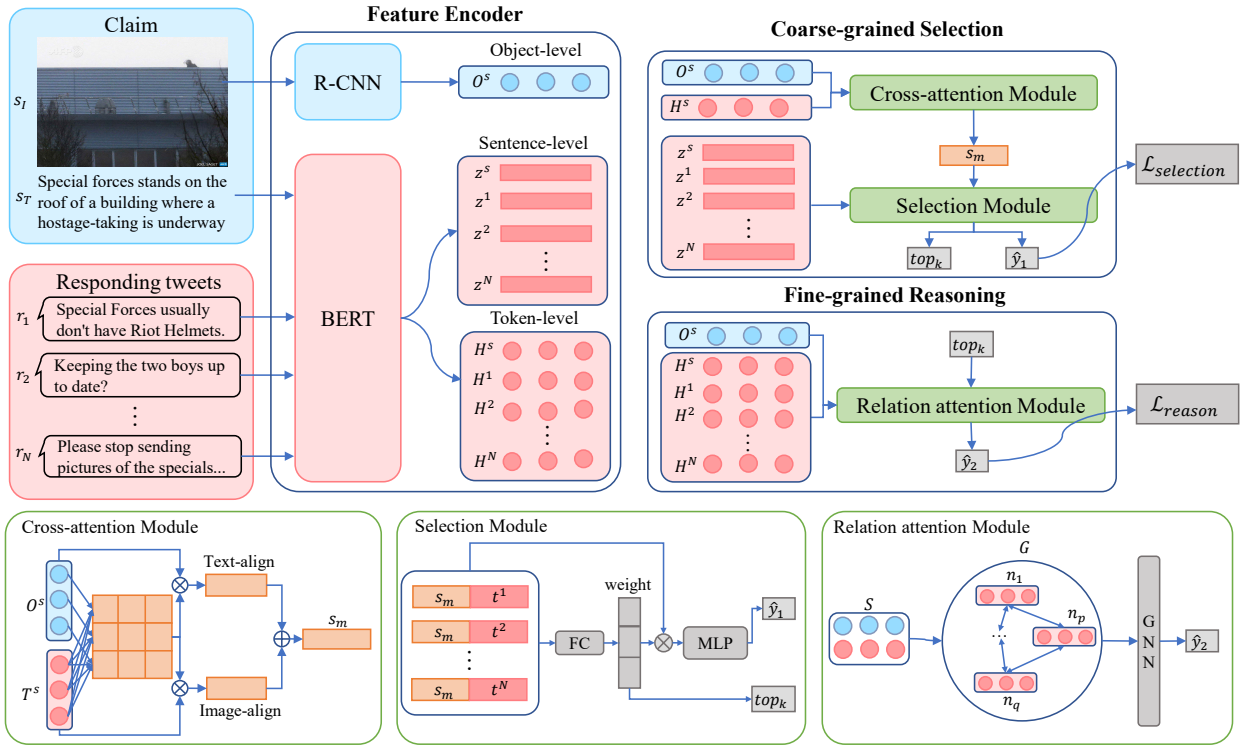

Fig. 2. Our proposed FoRM consists of three main components: feature encoder, coarse-grained selection and fine-grained reasoning. Specifically, the feature encoder extracts the fine-grained (token-level and object-level) features and the coarse-grained (sentence-level) features. The coarse-grained selection applies the cross-attention module to fuse the multi-modal feature of the claim and selects the candidate tweets with their sentence-level features. The fine-grained reasoning leverages the relation attention module to construct a full-connected graph of the selected tweets and the multi-modal claim for capturing their fine-grained relations.

where $H^s \in \mathbb{R}^{d_t \times M}$ and $H^i \in \mathbb{R}^{d_t \times M}$ are the token hidden states of s_T and r_i respectively, M denotes the number of tokens in s_T or r_i. We use the hidden state of the first token [CLS] to obtain the representation of the sentence:

$$z^s = \sigma(W^t H_0^s), \\ z^i = \sigma(W^t H_0^i), \tag{3}$$

where $W^t \in \mathbb{R}^{d_t \times d_t}$ is the shared parameter to obtain the sentence-level representations, $\sigma(\cdot)$ is the hyperbolic tangent activation function ($Tanh(\cdot)$), $z^s \in \mathbb{R}^{d_t \times 1}$ and $z^i \in \mathbb{R}^{d_t \times 1}$ denote sentence-level representations of the s_T and r_i;

4.1.2 Visual Encoder

VGGNet [56] and ResNet [57] are usually used as the visual encoder in previous work [14], [15], [23]. However, they fragment the image into a grid of regions, which barely conform to the semantics of the image [58]. To better fuse the semantic information of the text and the image, we apply Faster R-CNN [59] with bottom-up and top-down attention mechanisms as our visual encoder to detect K objects and extract the object-level features of s_I:

$$O^s = \text{Fast-RCNN}(s_I), \tag{4}$$

where $O^s = [o_1, o_2, \cdots, o_K] \in \mathbb{R}^{d_i \times K}$ are the features of the K objects in s_I.

4.2 Coarse-grained Selection

Due to the openness of social media, where anyone can respond to a claim, the responses have different relevance for verifying a rumor claim. As for a multi-modal claim, the focus of a user may be on specific words specific textual in the text or an object within the image. Therefore, it is effective to evaluate relevance by aligning responses with the different modalities of the rumor claim in a fine-grained manner. Yet, given the considerable volume of response tweets, executing such fine-grained alignment for each response is both costly and redundant. Consequently, we design the coarse-grained selection module to select candidate responses with the post-level representation. These chosen responses are then utilized for conducting fine-grained reasoning with the multi-modal claim.

4.2.1 Cross-attention Module

Instead of simply concatenating features of different modalities within the rumor claim, we leverage the cross-attention module to align the semantics between them. Specifically, we first project the token-level feature H^s and the object-level feature O^s into the same dimensional space:

$$V^s = \sigma(W^o O^s), \\ T^s = \sigma(W^h H^s), \tag{5}$$

where $W^o \in \mathbb{R}^{d \times d_i}$ and $W^h \in \mathbb{R}^{d \times d_t}$ are the training parameters, $V^s = [v_1, v_2, \cdots, v_K] \in \mathbb{R}^{d \times K}$ and $T^s = [w_1, w_2, \cdots, w_M] \in \mathbb{R}^{d \times M}$ are projected object features and token features respectively. Then, to obtain the relevant objects based on the tokens, we regard each token $w_i \in T^s$

as the query to calculate the cosine similarity of each object $v_j \in V^s$:

$$c_{ij}^{t\to o} = \cos(w_i, v_j). \tag{6}$$

The token-aligned feature is generated by computing the weighted sum of object-level features:

$$w_i^{t\to o} = \sum_{j=1}^{K} c_{ij}^{t\to o} v_j. \tag{7}$$

Finally, we combine all the token-aligned features to the text-aligned representation $s^{t\to o} \in \mathbb{R}^{d\times 1}$:

$$s^{t\to o} = \frac{1}{M}\sum_{i=1}^{M} w_i^{t\to o}. \tag{8}$$

To obtain the image-aligned representation, we regard the object v_j as the query and all the tokens in the claim text as the key and value:

$$\begin{aligned} c_{ji}^{o\to t} &= \cos(v_j, w_i), \\ v_j^{o\to t} &= \sum_{i=1}^{M} c_{ji}^{o\to t} w_j, \\ s^{o\to t} &= \frac{1}{K}\sum_{j=1}^{K} v_j^{o\to t}, \end{aligned} \tag{9}$$

where $c_{ji}^{o\to t}$ is the cosine similarity of the object v_j and the token w_i, $v_j^{o\to t}$ denotes the object-aligned features of v_j, $s^{o\to t}$ is the image-aligned representation of the claim.

The multi-modal representation of the claim is the weight summation of the text-aligned representation and the image-aligned representation:

$$s_m = \sigma(W^{o\to t} s^{o\to t}) + \sigma(W^{t\to o} s^{t\to o}), \tag{10}$$

where $W^{o\to t}$ and $W^{t\to o}$ are the training parameters, $s_m \in \mathbb{R}^{d\times 1}$ is the multi-modal representation of the claim.

4.2.2 Selection Module

In our work, the relevant response represents a tweet that is crucial for the verification of a rumor. To narrow the scope of relevant responses, a selection module utilizes post-level representations to sort all responses based on their significance in predicting rumor labels.

Since the multi-modal claim serves as the source of community responses, it is imperative to provide the context for each response by combining it with the claim:

$$A = s_m \mathbb{1}^T \circ \sigma(W^z Z), \tag{11}$$

where $A \in \mathbb{R}^{2d\times N}$ is the tweet matrix with claim context, $\mathbb{1} \in \mathbb{R}^{N\times 1}$ is set to replicate the claim to each responding tweet, $\circ$ is the concatenation operator, $W^z \in \mathbb{R}^{d_t\times d}$ is the parameter, $Z = \{z^1, z^2, \cdots, z^N\} \in \mathbb{R}^{d_t\times N}$. Then tweet matrix A is fed to generate the significance scores of tweets $\alpha \in \mathbb{R}^{1\times N}$ by a **softmax** function:

$$\alpha = \textbf{softmax}(W^a A), \tag{12}$$

where $W^a \in \mathbb{R}^{1\times 2d}$ is the training parameter. To improve the effectiveness of such significance in identifying rumors, we leverage it to predict the rumor label and maximize the accuracy of the prediction:

$$\hat{y_1} = \textbf{MLP}(\sum_{i=0}^{N} \alpha_i z^i), \tag{13}$$

where α_i denotes the significance score of tweet z^i, $\hat{y_1}$ is the coarse-grained prediction, the activation function and the hidden dimension of **MLP** are ReLU and d_h respectively. Based on such significance scores of responding tweets, the top k tweets $R^* = \{r_1^*, r_2^*, \cdots, r_k^*\}$ are selected for further fine-grained reasoning.

4.3 Fine-grained Reasoning

Compared with all the responses, the number of preserved responses is smaller. Hence, we design a relation attention network to model the fine-grained (token-to-token and token-to-object) relations between the preserved responses and the multi-modal claim. Inspired by [6], the relation attention module constructs a full-connected graph of the candidate tweets and multi-modal claim. The prediction process is split into two parts: rumor prediction based on a special node n_q after information propagation $P(y|n_q, G, S)$ and relevance prediction of such special node $P(n_q|G, S)$:

$$P(y|G, S) = \sum_{q\in G} P(y|n_q, G, S) P(n_q|G, S), \tag{14}$$

where the node n_q in graph G denotes a selected tweet $r_q \in R^*$ and is initialized with the sentence-level features z^q and token-level features H^q, S represents the claim, which consists of the fine-grained feature $S^m = [T^s \circ V^s]$ and the coarse-grained feature s_m.

4.3.1 Information Propagation

The information propagation aims to aggregate the information of neighbors and generate a new representation. Hence, we first extract the fine-grained features of neighbor node n_p based on the node n_q and the claim S. The token-level features H^p of the node n_p is regarded as the query to calculate the cosine similarity of the key (H^q of the node n_q and S^m of the claim):

$$\begin{aligned} T^p, T^q &= \sigma(W^p H^p), \sigma(W^q H^q), \\ C^{p\leftarrow\{q,s\}} &= \cos(T^p, [S^m \circ T^q]), \end{aligned} \tag{15}$$

where $W^p \in \mathbb{R}^{d\times d_t}$ and $W^q \in \mathbb{R}^{d\times d_t}$ are the parameters, T^p and T^q denote the token-level features of n_p and n_q projected into the dimensional space d, $C^{p\leftarrow\{q,s\}} \in \mathbb{R}^{M\times(2M+K)}$ is the cosine similarity matrix. Then, we obtain fine-grained features $T^{p\leftarrow\{q,s\}}$ of the neighbor n_p:

$$\begin{aligned} \alpha^{p\leftarrow\{q,s\}} &= \textbf{softmax}(C^{p\leftarrow\{q,s\}}), \\ T^{p\leftarrow\{q,s\}} &= \alpha^{p\leftarrow\{q,S\}}[S \circ T^q] + T^p, \end{aligned} \tag{16}$$

where $\alpha^{p\leftarrow\{q,s\}} \in \mathbb{R}^{M\times(2M+K)}$ denotes the significance of the tokens in n_p and the tokens and objects in S. To the end, we fuse all the token features in $T^{p\leftarrow\{q,s\}}$ to obtain the representation of n_p with the attention mechanism:

$$\begin{aligned} \beta &= \textbf{softmax}(W^{p\leftarrow\{q,s\}} T^{p\leftarrow\{q,s\}})), \\ z^{p\leftarrow\{q,s\}} &= \sum_{i=1}^{M} \beta_i T_i^{p\leftarrow\{q,s\}}, \end{aligned} \tag{17}$$

where $\beta \in \mathbb{R}^{1\times M}$ denotes the weight of all the token in n_p, $W^{p\leftarrow\{q,s\}}\mathbb{R}^{1\times d}$ is the training parameter, $z^{p\leftarrow\{q,s\}} \in \mathbb{R}^{1\times d}$ is the representation of n_p,.

To generate the propagated representation of n_q, we aggregate the representations of all its neighbors. We fuse the neighbor's representation $z^{p\leftarrow\{q,s\}}$, the sentence-level features z^q of n_q and the multi-modal feature s_m of the claim to calculate the importance of λ_p the neighbor node n_p:

$$\lambda_p = \textbf{softmax}_p(\textbf{MLP}(z^{p\leftarrow\{q,s\}} \circ s_m \circ \sigma(W^z z^q)), \quad (18)$$

where $\textbf{softmax}_p$ selects the value of n_p from the output of **softmax** function. The propagated representation v^q of n_q is represented as:

$$v^q = (\sum_{p\in G} \lambda_p z^{p\leftarrow\{q,s\}}) \circ \sigma(W^z z^q). \quad (19)$$

Finally, we predict the probability of the rumor label based on the special node n_q:

$$P(y|n_q, G, S) = \textbf{softmax}(W^y(v^q \circ s_m)), \quad (20)$$

where $W^y \in \mathbb{R}^{4\times 3d}$ is the parameter.

4.3.2 Relevance Prediction

Each node represents a responding tweet which may provide clues to verify the claim, so we use the claim to evaluate the relevance of a node.

Given a node n_q, we calculate the similarity between the fine-grained claim features S^m and tokens in T^q:

$$C^{s\leftarrow q} = \cos(S^m, T^q), \quad (21)$$

where $C^{s\leftarrow q} \in \mathbb{R}^{(M+K)\times M}$ is the translation matrix between the multi-modal claim and the tokens in T^q. Then, we obtain the claim fine-grained features $S^{s\leftarrow q} \in \mathbb{R}^{d\times(M+K)}$ by the weighted summation of all the tokens' features in T^q:

$$\begin{aligned} \alpha^{s\leftarrow q} &= \textbf{softmax}(C^{s\leftarrow q}), \\ S^{s\leftarrow q} &= \alpha^{s\leftarrow q} T^q + S^m. \end{aligned} \quad (22)$$

The relevance score of n_q is predicted as:

$$\begin{aligned} s^{s\leftarrow q} &= \sum_{i=1}^{M+K} S_i^{s\leftarrow q}, \\ P(n_q|G, S) &= \textbf{softmax}_q(W^{s\leftarrow q} s^{s\leftarrow q}), \end{aligned} \quad (23)$$

where $s^{s\leftarrow q} \in \mathbb{R}^{d\times 1}$ is the representation vector of the claim, $W^{s\leftarrow q} \in \mathbb{R}^{1\times d}$ is the parameter, $\textbf{softmax}_q$ selects the value of n_q from the output of **softmax** function. Finally, the rumor prediction by the whole graph is aggregated following Eq. 14.

4.4 Model Training

In the proposed two-stage framework, the coarse-grained selection module and the fine-grained reasoning module are jointly trained. Hence, during training, the scores of the coarse-grained selection could be refined by the feedback of the fine-grained reasoning module. Besides, to maximize the probability of the presence of relevant responses within the candidate set, we formulate a selection loss for the coarse-grained selection module to improve the effectiveness of the initial relevance scores:

$$\mathcal{L}_{selection} = \textbf{CrossEntropy}(y, \hat{y_1}), \quad (24)$$

TABLE 1
The statistics of the datasets.

Statistic	TWitter15	Twitter16	Weibo
Total conversation	1490	818	4664
Total tweets	331,612	204,820	3,805,656
Unverified	374	203	0
True	372	205	0
False	370	205	2,313
Non-rumor	374	205	2,351
images	737	400	3,675

where y denotes the rumor class label, $\hat{y_1}$ is the label prediction based on the initial relevance scores of all responses.

As for the whole framework, we leverage the cross entropy loss to minimizes the difference between the ground truth y and the predicted distribution $P(y|G, S)$:

$$\mathcal{L}_{reason} = \textbf{CrossEntropy}(y, P(y|G, S)). \quad (25)$$

Towards the end, the total loss function is defined as the linear combination of the above two loss functions:

$$\mathcal{L} = \mathcal{L}_{reason} + \beta\mathcal{L}_{selection}, \quad (26)$$

where β is a trade-off parameter.

5 EXPERIMENTS

5.1 Experimental Setting

5.1.1 Datasets

We train and evaluate our FoRM on three public datasets: Twitter15 [60], Twitter16 [60] and Weibo [12]. Twitter15 and Twitter16 both consist of Twitter threads, which contain a claim and a set of responding tweets. To expand a given claim into a multi-modal one, we crawl the corresponding image based on its ID and textual information. These two datasets contains four rumor labels: Unverified, Ture, False and Non-rumor. As for Weibo, there are two labels: False and Non-rumor, and we also crawl the corresponding image for each claim. We follow [61] to process and split the datasets. The statistics of the processed datasets are shown in Table 1 and the processed data is publicly accessible[2].

5.1.2 Implementation Details

In our experiments, the textual encoder inherits huggingface's implementation[3] and the visual encoder is implemented by bottom-up-attention[4]. Limited by the memory of GPUs, we set the number of responding tweets N to 100, the max number of tokens M to 35, and the max number of objects K to 36. We use the special token [PAD] in BERT embedding to fill the tweets less than 100 and the tokens less than 35. For the objects less than 36, we leverage the one-padding to fill them. The dimension of the textual (token-level, sentence-level) feature d_t and object-level feature d_i are 768 and 2048 respectively. To fuse features of different modalities, we project them into the same dimensional space $d = 768$. The hidden dimension d_h of **MLP** is set to 128. Finally, the model is optimized by Adam optimizer. We set

2. https://github.com/chunyuanY/RumorDetection
3. https://github.com/huggingface/transformers
4. https://github.com/peteanderson80/bottom-up-attention

TABLE 2
Rumor verification results on Twitter15 and Twitter16 (FR: False rumor; TR: True rumor; UR: Unverified rumor; NR: Non-rumor).

	Method	Twitter15					Twitter16				
		Acc.	FR-F1	TR-F1	UR-F1	NR-F1	Acc.	FR-F1	TR-F1	UR-F1	NR-F1
Textual	DTC	0.454	0.355	0.317	0.415	0.733	0.465	0.393	0.419	0.403	0.643
	GRU2	0.646	0.574	0.608	0.592	0.792	0.633	0.489	0.686	0.593	0.772
	RvNN	0.723	0.758	0.821	0.654	0.682	0.737	0.743	0.835	0.708	0.662
	PPC	0.842	0.875	0.818	0.790	0.811	0.863	**0.898**	0.843	0.837	0.820
	dEFEND	0.839	0.872	0.849	0.813	0.820	0.859	0.818	0.936	0.860	0.820
	GLAN	0.890	0.880	0.908	0.841	**0.929**	0.897	0.848	0.938	0.876	**0.923**
	PLAN	0.845	0.858	0.895	0.802	0.823	0.874	0.839	0.917	0.888	0.853
Multi-modal	att-RNN	0.774	0.778	0.859	0.780	0.662	0.767	0.781	0.898	0.723	0.634
	EANN	0.804	0.837	0.870	0.795	0.701	0.815	0.831	0.895	0.804	0.729
	LIIMR	0.819	0.826	0.754	0.889	0.807	0.837	0.841	0.787	0.889	0.826
	CAFE	0.821	0.811	0.760	0.889	0.826	0.832	0.819	0.809	0.889	0.804
	MKEMN	0.836	0.819	0.905	0.793	0.829	0.848	0.842	0.903	0.860	0.787
	Retrieval-based	0.848	0.877	0.881	0.794	0.832	0.870	0.886	0.920	0.851	0.817
	MFAN	0.851	0.838	0.900	0.864	0.809	0.875	0.864	0.957	0.867	0.821
	FoRM	**0.902**	**0.886**	**0.946**	**0.885**	0.890	**0.913**	0.894	**0.958**	**0.905**	0.891

the learning rate and batch size to 5e-5 and 4 respectively. All the models is implemented based on PyTorch with a 24GB NVIDIA GeForce RTX 3090. The params of our FoRM is 90M. For Weibo dataset, the training time is about 2 hours and the inference time for the whole test set is about 2 mins.

5.2 Baselines

We compare the proposed FoRM with the following two categories of baselines:

5.2.1 Single Modality Methods

- **DTC [9]** predicts the credibility of a Twitter event by a decision tree classifier and four types of hand-crafted features, including message-based, user-based, topic-based, and propagation-based features.
- **GRU2 [12]** constructs all responding tweets as variable-length time series, and uses a multilayer GRU network to model the temporal pattern of the rumor diffusion.
- **RvNN [19]** organizes a conversation thread as a tree-structure propagation based on the reply relationship, and recursively generates the representation of such propagation from different angles (bottom-up or top-down) to identify rumors.
- **PPC [62]** uses both recurrent and convolutional networks to model the propagation path of each claim, with the aim of the early detection.
- **dEFEND [22]** is a text-based method, which uses bi-GRU to encode text content and co-attention to select the key comments for explainable detection.
- **GLAN [61]** regards all claim tweets, responses and users as a heterogeneous graph, and models the local semantic relation (a claim and corresponding responses) and the global structure (all claim tweets and all users) of rumors with the attention network.
- **PLAN [20]** is a Transformer-based method that models the full-connected structure of all responses considering the uncertain response relations.

5.2.2 Multi-modal Methods

- **EANN [33]** applies the adversarial neural networks to capture event-invariant features of the multi-modal claim for debunking fake news in newly emerged events.
- **att-RNN [63]** identifies rumors by leveraging the attention mechanism to extract multi-modal features from the claim, including textual, visual and social context features.
- **LIIMR [64]** explores the importance of different modalities in identifying fake news.
- **CAFE [24]** regards the cross-modal ambiguity as a gate to adaptively aggregate unimodal features or capture cross-modal correlations.
- **MKEMN [14]** captures latent temporal features of the current event with its all corresponding responses and then builds the external memory to restore the event-invariant features for transferring from seen data to newly emerged events.
- **Retrieval-based [26]** proposes a retrieval baseline which leverages the claim text or the claim image to do evidence retrieval in a coarse-grained manner.
- **MFAN [25]** makes the first attempt to integrate the heterogeneous multi-modal data (text, image and social graph) in one framework, and verifies rumors by modeling complex data relationships with the graph network.

5.3 Performance Comparison

Table 2 and Table 3 demonstrates the rumor verification results of the compared methods and our model on three real-word datasets. Following [19], [61], we leverage the accuracy as the evaluation metric over the four classes and F_1 score for each one. As for Weibo, the accuracy is the main evaluation metric and precision, recall and F1 score are adopted to evaluate each class.

The single modality methods model community response from different perspectives. Without the deep learning, DTC performs worse than other methods, because the hand-crafted features could not enumerate all the features hidden in the community response. Although the deep learning network is applied in the other models, the result of GRU2 is poorer due to the failure to capture semantic

TABLE 3
Rumor verification results of Weibo.

	Method	Acc.	FR			TR		
			Precision	Recall	F1	Precision	Recall	F1
Textual	DTC	0.831	0.847	0.815	0.831	0.815	0.847	0.830
	GRU2	0.910	0.876	0.956	0.914	0.952	0.864	0.906
	PPC	0.921	0.896	0.962	0.923	0.949	0.889	0.918
	dEFEND	0.908	0.905	0.908	0.907	0.911	0.905	0.908
	GLAN	0.946	0.943	0.948	0.945	0.949	0.943	0.946
	PLAN	0.920	0.918	0.921	0.920	0.922	0.919	0.920
Multi-modal	att-RNN	0.899	0.914	0.879	0.896	0.885	0.919	0.902
	EANN	0.909	0.949	0.862	0.903	0.875	0.955	0.913
	LIIMR	0.914	0.913	0.917	0.915	0.915	0.912	0.913
	CAFE	0.921	0.934	0.909	0.922	0.908	0.933	0.921
	MKEMN	0.923	0.935	0.908	0.921	0.912	0.938	0.925
	Retrieval-based	0.931	0.950	0.910	0.929	0.915	0.953	0.933
	MFAN	0.935	0.954	0.914	0.933	0.918	**0.957**	0.937
	FoRM	**0.962**	**0.956**	**0.967**	**0.961**	**0.967**	**0.957**	**0.962**

information of special response. GLAN and PLAN regard each response as a unit and model the their interaction with post-level embedding, so they reach the better performance, which indicates the importance of the semantic information in user responses.

Among all the multi-modal baselines, the response-based methods (MKEMN, Retrieval-Based and MFAN) outperform the claim-based methods (att-RNN, EANN, LIIMR and CAFE). Thus, discovering clues from user responses for rumor detection is a practical way. Benefited from BERT, the powerful text encoder, the performance of LIIMR and CAFE are both better than att-RNN and EANN. However, even introducing the multi-modal claim, the result of MFAN is also poorer than GLAN, which indicates the effective method of modelling the multi-modal claim and user responses has not been well exploited.

The multi-modal coarse-grained selection baseline (Retrieval-Based) has a better performance than the textual one (dEFEND). This indicates that the multi-modal information in the claim is crucial to identify rumors. Retrieval-Based leverages the each modality of the claim as query to search the evidences independently, and just concatenate the their result for rumor verification. Thus, Retrieval-Based ignores the cross-modality relation of the claim. Besides, the fine-grained features of the responses and the multi-modal claim are also ignored. These two reasons may result in the poorer performance than our FoRM.

To the end, our FoRM leverages the multi-modal information of the claim to select the relevant responses from the whole responses, which could avoid the irrelevant tweets distract the model from locating the key clues. Considering the focus of a user may be on special words in the text content or an object in the image, we propose to evaluate the relevance by aligning the response and the multi-modal claim in the fine-grained way. This is different from dEFEND, PLAN and Retrieval-Based that leverages the post-level attention to focus on the crucial responses. Hence, FoRM achieves the best performance among all the baselines, not only single modality methods but also the multi-modal methods.

TABLE 4
The effectiveness of different components in FoRM on Weibo.

Models	Acc.	FR F1	NR F1
FoRM	0.962	0.961	0.962
w/o Visual	0.935	0.935	0.935
w/o Fine-grained	0.945	0.945	0.944
w/o Selection loss	0.928	0.927	0.928
w/o Jointly train	0.934	0.934	0.934

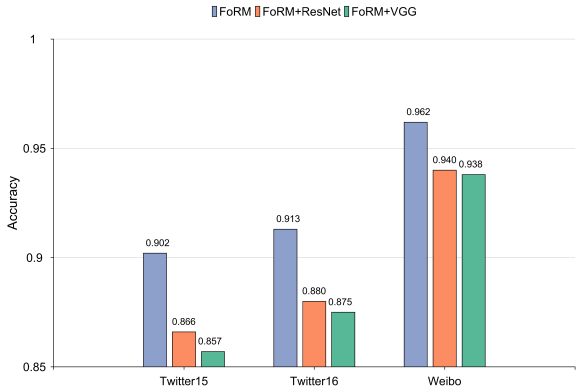

Fig. 3. Performance comparison with different visual encoders.

5.4 Ablation Study

To validate the effectiveness of the components in the proposed FoRM, we design several variants based on our model for ablation experiments.

In Table 4, to show the effectiveness of each components of FoRM, we remove the key components and design four variants: **FoRM w/o Visual**, **FoRM w/o Fine-grained**, **FoRM w/o Selection loss**, and **FoRM w/o Jointly train**. Without the selection loss, the performance of FoRM w/o selection loss drops significantly, which demonstrates the designed selection loss could improve the effectiveness of the candidate responses for subsequent fine-grained reasoning. Besides, we first train the coarse-grained selection module

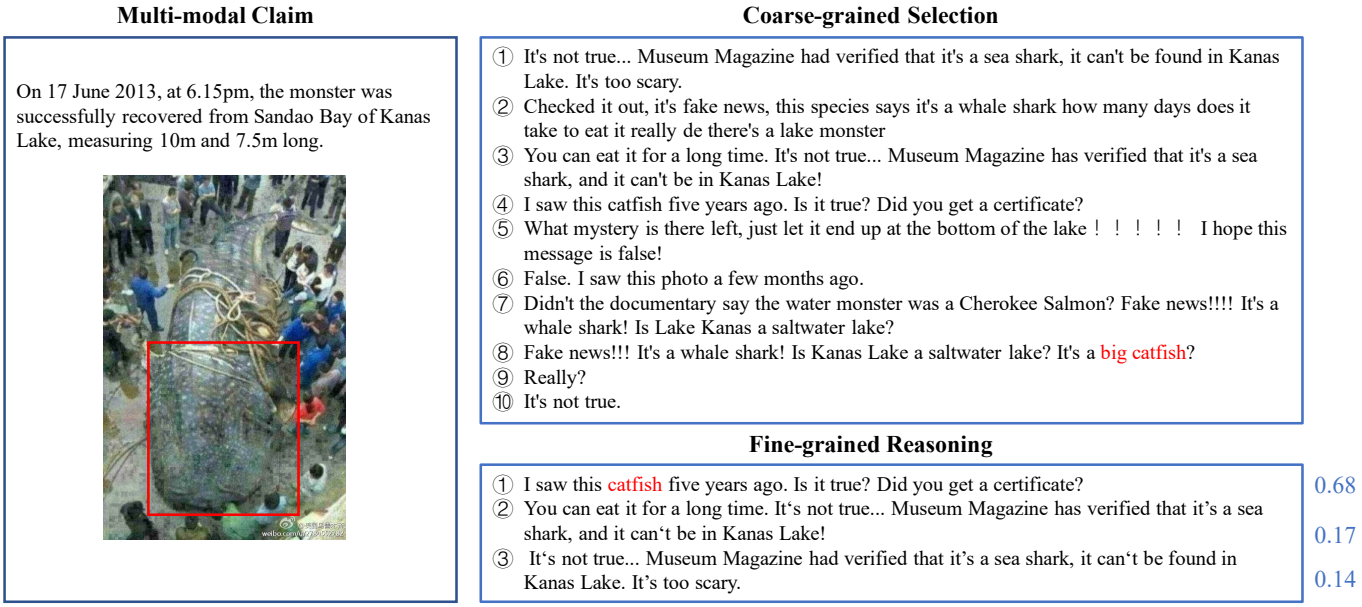

Fig. 4. A sample of false rumor in Weibo dataset and we translate the Chinese sentences for reference. In the right part, coarse-grained selection module selects 10 responses as the candidates. Subsequently, fine-grained reasoning module resorts these candidate responses and assigns higher relevance scores (blue numbers) for three. The red words and red image region represent the focus of FoRM to debunk this false rumor.

TABLE 5
Results of different fine-grained module on Weibo.

Models	Acc.	FR F1	NR F1
FoRM	0.962	0.961	0.962
+GAT with sentence-level features	0.943	0.941	0.944
w/o information propagation	0.937	0.938	0.937
w/o relevance prediction	0.952	0.954	0.951

TABLE 6
Results of different selection module on Weibo.

Models	Acc.	FR F1	NR F1
FoRM	0.962	0.961	0.962
+attention	0.941	0.939	0.942
+cosine similarity	0.930	0.928	0.931
+non-linear layer	0.940	0.940	0.940
w/o cross-attention module	0.953	0.953	0.954

with selection loss separately. Then, leveraging its outputs of relevance scores to train the fine-grained reasoning module. We observe that the performance of FoRM w/o Jointly train even poorer than FoRM w/o Fine-grained. Because of the lack of the fine-grained interaction, the relevance scores initialized by the coarse-grained selection module are not reasonable enough. Based on these scores, the selected candidate responses may confuse the fine-grained reasoning module and result in the drop of accuracy. Thus, training this two modules in the end-to-end manner would refine the scores of the coarse-grained selection based on the feedback of the fine-grained reasoning, which could also improve the effectiveness of the selection.

In Fig 3, we compare the performance of FoRM with different visual encoders. As the pre-trained VGG and ResNet are widely used for extracting the visual feature, we leverage them to replace our visual encoder respectively. Compared with FoRM, the accuracy of FoRM+ResNet and FoRM+VGG are both poorer. These results indicate that the feature extraction of Faster R-CNN conforms to the semantics of the image, which performs better in our model.

To further demonstrate the effectiveness of the fine-grained component, we design three variants to fuse the candidate responses from the selection module. In the information propagation, the semantic representation of responses are inferred by the fine-grained aggregation with the multi-modal claim and other candidate responses. Without the information propagation, the accuracy of **FoRM w/o information propagation** drops significantly, which indicate that response representation by fine-grained aggregation is more important for rumor detection. Besides, although the fine-grained fusion is used in each response and the multi-modal claim, FoRM w/o information propagation still gets the poorer performance than FoRM + GAT. This demonstrates that only replying on a single response would not identify rumors enough, and other responses may provide additional evidences to refine current response. The high accuracy of **w/o relevance prediction**, which directly leverage the attention aggregation of the response representations, also reveals that the importance of more complicated response representation by the information propagation.

Within the coarse-grained selection module, we concatenate the multi-modal claim and responses with post-level representation. Then a linear layer with softmax function is used to initialize the relevance score of each response. In Table 6, we leverage attention and cosine similarity to initialize the relevance scores. Specifically, **FoRM+attention** and **FoRM+cosine similarity** regard the claim as a query to calculate the attention scores and cosine similarity scores

to initialize the relevance scores. We use the non-linear layer, **FoRM+non-linear layer** (a linear layer with ReLU and softmax), to initialize the relevance scores, rather than the linear layer. The accuracy of these three variants all drop significantly, which demonstrates the concatenation of the multi-modal claim and the responses is more suitable for our FoRM. Additionally, the cross-attention module is crucial to extract the multi-modal representation of the claim. When replacing the cross-attention module with element-wise addition, the accuracy drops almost 1.0%.

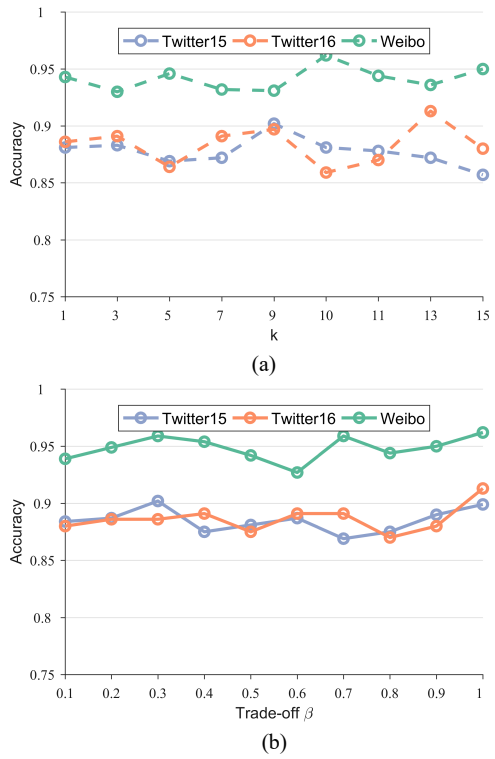

Fig. 5. The influence of the number k of candidate responses and the trade-off β on the performance of our model.

Finally, Fig. 5 shows the influence of the number of candidate responses (top_k) and the trade-off β of the loss function on performance. As it is costly and redundant to align all responses and the multi-modal claim in the fine-grained way, we propose to select top_k responses as the candidate ones based on the initial relevance scores from the coarse-grained selection module. In Fig. 5(a), our FoRM achieves best performance on Twitter15, Twitter16 and Weibo, when k is 9, 13, 10 respectively. These results suggest that the more responding tweets used, the worse performances may be obtained, since more tweets may include noise tweets and propagate the noise to other tweets through the information propagation. Besides, another important hyper-parameter is the trade-off β of combining the selection loss. In Fig. 5(b), the results show that the best performance is reached by setting β to 0.3, 1.0, and 1.0 for Twitter15, Twitter16 and Weibo respectively.

5.5 Case Study

Fig. 4 shows the example of multi-modal false rumor in Weibo dataset. As a two-stage framework, FoRM firstly use the coarse-grained selection module to select 10 responses by their post-level representations. Benefited from the selection loss, these candidate responses include the query, *"Really?"*, and denying, *"It's not true."*, which are both helpful to debunk rumors. Subsequently, the fine-grained reasoning module resorts the candidate ones with the fine-grained relation attention. Rather than only focusing on the stance, denying the claim, our model pays more attention on finding clues, thus the three responses with higher relevance scores both contain the stance of querying or denying, *"It's not true"*, *"Is it true?"* and the reasons, such as, *"I saw this catfish five years ago."*. The highest relevance score response, *"I saw this catfish five years ago. Is it true? Did you get a certificate?"*, shows the effectiveness of our relation attention module. The key word, "catfish", connects this response and the image region in the claim, resulting in the highest relevance score. This also indicates the main clues for debunking this multi-modal claim are in the image and explains why not pay more attention on the response including the same word, "Kanas Lake", in the claim text. Within the information propagation, with the same word, "catfish", the response, "I saw this catfish five years ago. Is it true? Did you get a certificate?", connects the eighth response, "Fake news!!! It's a whale shark! Is Kanas Lake a saltwater lake? It's a big catfish?". Finally, the clues are incorporated as: (1) this image was posted five years ago; (2) the monster in this image is a whale shark; (3) Kanas Lake is not a saltwater lake, so it is not a suitable living area for whale sharks. Hence, this multi-modal claim is debunked as false rumor.

6 CONCLUSION

In this paper, we proposed a two-stage framework, termed FoRM, to focus on the relevant response for multi-modal rumor detection model. To evaluate the relevance of responses for verifying the multi-modal claim, an effective approach is to align responses with the different modalities of the rumor claim in a fine-grained manner. However, due to the substantial volume of response tweets, it is both costly and redundant to align all responses with the multi-modal claim. Therefore, our FoRM consists of the coarse-grained selection module and the fine-grained reasoning module. Specifically, the coarse-grained selection module firstly leveraged the post-level features to select top k responses as the candidates. Subsequently, the fine-grained reasoning module captured the fine-grained relations, *i.e.*, token-to-token and token-to-object relations, of the multi-modal claim and the candidate tweets to identify the relevant responses for debunking rumors. The experiments on three public datasets showed that our model outperforms all the baselines and could select more reasonable tweets based on the multi-modal claim for further rumor detection.

ACKNOWLEDGMENTS

This work was partially supported by the National Natural Science Foundation of China under grant 62220106008, U20B2063, and 62102070. This work was also partially supported by Sichuan Science and Technology Program under grant 2023NSFSC1392. We also acknowledge the thoughtful discussions and paper revisions from Zeyu Ma.

REFERENCES

[1] H. Webb, P. Burnap, R. Procter, O. Rana, B. C. Stahl, M. Williams, W. Housley, A. Edwards, and M. Jirotka, "Digital wildfires: Propagation, verification, regulation, and responsible innovation," *ACM TOIS*, vol. 34, no. 3, pp. 1–23, 2016.

[2] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[3] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, "Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa," *Nature human behaviour*, vol. 5, no. 3, pp. 337–348, 2021.

[4] J. Thorne, A. Vlachos, O. Cocarascu, C. Christodoulopoulos, and A. Mittal, "The fact extraction and verification (fever) shared task," in *Workshop of EMNLP*, 2018, pp. 1–9.

[5] J. Zhou, X. Han, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Gear: Graph-based evidence aggregating and reasoning for fact verification," in *ACL*, 2019, pp. 892–901.

[6] Z. Liu, C. Xiong, M. Sun, and Z. Liu, "Fine-grained fact verification with kernel graph attention network," in *ACL*, 2020, pp. 7342–7351.

[7] R. Procter, J. Crump, S. Karstedt, A. Voss, and M. Cantijoch, "Reading the riots: What were the police doing on twitter?" *Policing and society*, vol. 23, no. 4, pp. 413–436, 2013.

[8] H. Li and Y. Sakamoto, "Computing the veracity of information through crowds: A method for reducing the spread of false messages on social media," in *2015 48th Hawaii international conference on system sciences*. IEEE, 2015, pp. 2003–2012.

[9] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *WWW*, 2011, pp. 675–684.

[10] F. Yang, Y. Liu, X. Yu, and M. Yang, "Automatic detection of rumor on sina weibo," in *KDD*, 2012, pp. 1–7.

[11] X. Liu, A. Nourbakhsh, Q. Li, R. Fang, and S. Shah, "Real-time rumor debunking on twitter," in *CIKM*, 2015, pp. 1867–1870.

[12] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, and M. Cha, "Detecting rumors from microblogs with recurrent neural networks," in *IJCAI*. AAAI Press, 2016.

[13] T. Chen, X. Li, H. Yin, and J. Zhang, "Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection," in *Pacific-Asia conference on knowledge discovery and data mining*. Springer, 2018, pp. 40–52.

[14] H. Zhang, Q. Fang, S. Qian, and C. Xu, "Multi-modal knowledge-aware event memory network for social media rumor detection," in *ACM Multimedia*, 2019, pp. 1942–1951.

[15] H. Zhang, S. Qian, Q. Fang, and C. Xu, "Multi-modal meta multi-task learning for social media rumor detection," *IEEE TMM*, vol. 24, pp. 1449–1459, 2021.

[16] J. Ma, J. Li, W. Gao, Y. Yang, and K.-F. Wong, "Improving rumor detection by promoting information campaigns with transformer-based generative adversarial learning," *IEEE TKDE*, 2021.

[17] V. Qazvinian, E. Rosengren, D. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *EMNLP*, 2011, pp. 1589–1599.

[18] F. Yu, Q. Liu, S. Wu, L. Wang, T. Tan *et al.*, "A convolutional approach for misinformation identification." in *IJCAI*. AAAI Press, 2017, pp. 3901–3907.

[19] J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on twitter with tree-structured recursive neural networks," in *ACL*, 2018, pp. 1980–1989.

[20] L. M. S. Khoo, H. L. Chieu, Z. Qian, and J. Jiang, "Interpretable rumor detection in microblogs by attending to user interactions," in *AAAI*, vol. 34, no. 05, 2020, pp. 8783–8790.

[21] L. Fang, K. Feng, K. Zhao, A. Hu, and T. Li, "Unsupervised rumor detection based on propagation tree vae," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

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

[23] S. Qian, J. Wang, J. Hu, Q. Fang, and C. Xu, "Hierarchical multi-modal contextual attention network for fake news detection," in *SIGIR*, 2021, pp. 153–162.

[24] Y. Chen, D. Li, P. Zhang, J. Sui, Q. Lv, L. Tun, and L. Shang, "Cross-modal ambiguity learning for multimodal fake news detection," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 2897–2905.

[25] J. Zheng, X. Zhang, S. Guo, Q. Wang, W. Zang, and Y. Zhang, "Mfan: Multi-modal feature-enhanced attention networks for rumor detection." IJCAI, 2022.

[26] X. Hu, Z. Guo, J. Chen, L. Wen, and P. S. Yu, "Mr2: A benchmark for multimodal retrieval-augmented rumor detection in social media," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 2901–2912.

[27] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Computing Surveys*, vol. 51, no. 2, pp. 1–36, 2018.

[28] A. Zubiaga, M. Liakata, R. Procter, G. Wong Sak Hoi, and P. Tolmie, "Analysing how people orient to and spread rumours in social media by looking at conversational threads," *PloS one*, vol. 11, no. 3, p. e0150989, 2016.

[29] J. Cao, J. Guo, X. Li, Z. Jin, H. Guo, and J. Li, "Automatic rumor detection on microblogs: A survey," *arXiv preprint arXiv:1807.03505*, 2018.

[30] D. Khattar, J. S. Goud, M. Gupta, and V. Varma, "Mvae: Multimodal variational autoencoder for fake news detection," in *WWW*, 2019, pp. 2915–2921.

[31] 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 *ACM Multimedia*, 2021, pp. 1212–1220.

[32] M. Sun, X. Zhang, J. Ma, S. Xie, Y. Liu, and S. Y. Philip, "Inconsistent matters: A knowledge-guided dual-consistency network for multi-modal rumor detection," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[33] 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 *KDD*, 2018, pp. 849–857.

[34] H. Zhang, S. Qian, Q. Fang, and C. Xu, "Multimodal disentangled domain adaption for social media event rumor detection," *IEEE TMM*, vol. 23, pp. 4441–4454, 2020.

[35] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, "Describing video with attention-based bidirectional lstm," *IEEE transactions on cybernetics*, vol. 49, no. 7, pp. 2631–2641, 2018.

[36] X. Huang, S. Qian, Q. Fang, J. Sang, and C. Xu, "Csan: Contextual self-attention network for user sequential recommendation," in *ACM Multimedia*, 2018, pp. 447–455.

[37] C. Liu, Z. Mao, T. Zhang, A. Liu, B. Wang, and Y. Zhang, "Focus your attention: A focal attention for multimodal learning," *IEEE TMM*, 2020.

[38] L. Yu, J. Zhang, and Q. Wu, "Dual attention on pyramid feature maps for image captioning," *IEEE TMM*, vol. 24, pp. 1775–1786, 2021.

[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[40] M. Zhao, S. Yue, D. Katabi, T. S. Jaakkola, and M. T. Bianchi, "Learning sleep stages from radio signals: A conditional adversarial architecture," in *ICML*. PMLR, 2017, pp. 4100–4109.

[41] Y. Yang, J. Zhou, J. Ai, Y. Bin, A. Hanjalic, H. T. Shen, and Y. Ji, "Video captioning by adversarial lstm," *IEEE TIP*, vol. 27, no. 11, pp. 5600–5611, 2018.

[42] X. Ma, T. Zhang, and C. Xu, "Deep multi-modality adversarial networks for unsupervised domain adaptation," *IEEE TMM*, vol. 21, no. 9, pp. 2419–2431, 2019.

[43] S. Li, C. H. Liu, B. Xie, L. Su, Z. Ding, and G. Huang, "Joint adversarial domain adaptation," in *ACM Multimedia*, 2019, pp. 729–737.

[44] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational fact checking from knowledge networks," *PloS one*, vol. 10, no. 6, p. e0128193, 2015.

[45] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on sina weibo by propagation structures," in *IEEE ICDE*. IEEE, 2015, pp. 651–662.

[46] J. Ma, W. Gao, Z. Wei, Y. Lu, and K.-F. Wong, "Detect rumors using time series of social context information on microblogging websites," in *CIKM*, 2015, pp. 1751–1754.

[47] S. Kumar and K. M. Carley, "Tree lstms with convolution units to predict stance and rumor veracity in social media conversations," in *ACL*, 2019, pp. 5047–5058.

[48] J. Ma and W. Gao, "Debunking rumors on twitter with tree transformer," in *COLING*, 2020, pp. 5455–5466.

[49] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *ACL*, 2015.
[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NIPS*, 2017.
[51] P. Wei, N. Xu, and W. Mao, "Modeling conversation structure and temporal dynamics for jointly predicting rumor stance and veracity," in *EMNLP*, 2019, pp. 4787–4798.
[52] A. Lao, C. Shi, and Y. Yang, "Rumor detection with field of linear and non-linear propagation," in *WWW*, 2021, pp. 3178–3187.
[53] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019, pp. 4171–4186.
[54] R. Guan, H. Zhang, Y. Liang, F. Giunchiglia, L. Huang, and X. Feng, "Deep feature-based text clustering and its explanation," *IEEE TKDE*, 2020.
[55] J. Li, A. Sun, J. Han, and C. Li, "A survey on deep learning for named entity recognition," *IEEE TKDE*, vol. 34, no. 1, pp. 50–70, 2020.
[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[57] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[58] L. Peng, Y. Yang, X. Zhang, Y. Ji, H. Lu, and H. T. Shen, "Answer again: Improving vqa with cascaded-answering model," *IEEE TKDE*, 2020.
[59] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *CVPR*, 2018, pp. 6077–6086.
[60] J. Ma, W. Gao, and K.-F. Wong, "Detect rumors in microblog posts using propagation structure via kernel learning," in *ACL*, 2017.
[61] 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 *2019 IEEE international conference on data mining (ICDM)*. IEEE, 2019, pp. 796–805.
[62] Y. Liu and Y.-F. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *AAAI*, vol. 32, no. 1, 2018.
[63] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal fusion with recurrent neural networks for rumor detection on microblogs," in *ACM Multimedia*, 2017, pp. 795–816.
[64] S. Singhal, T. Pandey, S. Mrig, R. R. Shah, and P. Kumaraguru, "Leveraging intra and inter modality relationship for multimodal fake news detection," in *Companion Proceedings of the Web Conference 2022*, 2022, pp. 726–734.

Jun Li received the bachelor's degree and master's degree from Chongqing University in 2013 and 2016. He is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China. His current research interests include Natural Language Processing and Social Media Analytics.

Yi Bin is currently with the University of Electronic Science and Technology of China, Chengdu, China. He received the Ph.D. degree from UESTC in 2020. His research interests include multimedia analysis, vision understanding and deep learning.

Liang Peng received his Ph.D. degree in School of Computer Science and Engineering, University of Electronic Science and Technology of China. His research interests include vision understanding, natural language processing and deep learning.

Yang Yang (Senior Member, IEEE) received the bachelor's degree from Jilin University, Changchun, China, in 2006, the master's degree from Peking University, Beijing, China, in 2009, and the Ph.D. degree from The University of Queensland, Brisbane, QLD, Australia, in 2012, all in computer science. He is currently with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include multimedia content analysis, computer vision, and social media analytics.

Yangyang Li received the B.S. degree from the Nanjing University of Information and Technology, in 2009, and the Ph.D. degree in computer science from the Beijing University of Posts and Telecommunications, in 2015. He is currently a Senior Engineer at the National Engineering Research Center for Risk Perception and Prevention. His research interests include the cyberspace security, social networks, and data science.

Hao Jin received the B.S. degree in computer science from the Hefei University of Technology in 2013, and the Ph.D. degree in Communication and Information System from the University of Chinese Academy of Sciences, in 2018. She is currently a Senior Engineer at the National Engineering Research Center for Risk Perception and Prevention. Her research interests include the cyberspace security and social networks.

Zi Huang received her BSc degree from the Department of Computer Science, Tsinghua University, China, and the PhD degree in computer science from The University of Queensland. She is a professor and ARC Future Fellow in The School of Information Technology and Electrical Engineering, The University of Queensland. Her research interests mainly include multimedia indexing and search, social data analysis, and knowledge discovery.