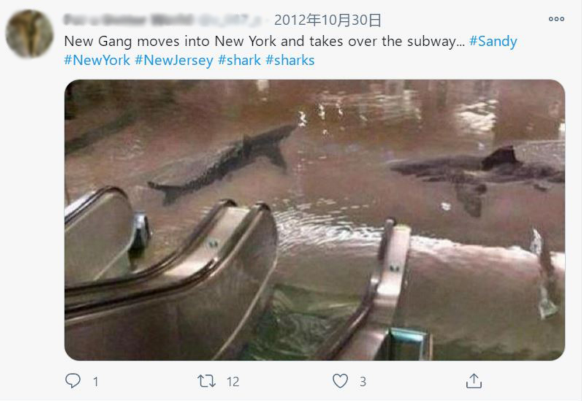
Inconsistency Matters: A Knowledge-guided Dual-inconsistency Network for Multi-modal Rumor Detection

Mengzhu Sun[1] and **Xi Zhang**[1*] and **Jianqiang Ma**[2] and **Yazheng Liu**[1]
[1]Key Laboratory of Trustworthy Distributed Computing and Service (MoE),
Beijing University of Posts and Telecommunications, China
[2] Platform and Content Group, Tencent
{2019110945, zhangx, liuyz}@bupt.edu.cn,
alexanderma@tencent.com

Abstract

Rumor spreaders are increasingly utilizing multimedia content to attract the attention and trust of news consumers. Though a set of rumor detection models have exploited the multi-modal data, they seldom consider the inconsistent relationships among images and texts. Moreover, they also fail to find a powerful way to spot the inconsistency information among the post contents and background knowledge. Motivated by the intuition that rumors are more likely to have inconsistency information in semantics, a novel *Knowledge-guided Dual-inconsistency network* is proposed to detect rumors with multimedia contents. It can capture the inconsistent semantics at the cross-modal level and the content-knowledge level in one unified framework. Extensive experiments on two public real-world datasets demonstrate that our proposal can outperform the state-of-the-art baselines.



[redacted] 2012年10月30日
New Gang moves into New York and takes over the subway... #Sandy #NewYork #NewJersey #shark #sharks

1 12 3

Figure 1. A real-world example of a fake multimedia tweet. It is suspicious to see sharks appear in a subway. Such abnormality should be captured and serve as an essential clue for rumor identification.

1 Introduction

Social media has fostered various false information, including misrepresented or even forged multimedia content, to mislead the readers. The widespread rumors may cause significant adverse effects. For example, some offenders use rumors to guide public opinion, damage the credibility of government, and even interfere with the general election (Allcott and Gentzkow, 2017). Therefore, it is urgent to automatically detect and regulate rumors to promote trust in the social media ecosystem.

Most of existing rumor detection methods focus on textual data to extract distinctive features (Castillo et al., 2011; Chen et al., 2018; Ma et al., 2016; Yu et al., 2017). With multimedia technology development, visual contents have become an important part of rumors to attract and mislead the consumers for more credible storytelling and rapid diffusion (Jin et al., 2016; Qi et al., 2019). Detecting multimedia rumor posts is a double-edged sword. On the one hand, it is more challenging to learn effective feature representations from heterogeneous multi-modal data. On the other hand, it also provides a great opportunity to identify rumors. Xue et al. (2021) shows that, in order to catch eyes of public, rumors tend to use theatrical, comical and attractive images that are irrelevant to the post content. In general, it is often difficult to find pertinent and non-manipulated images to match fictional events, thus posts with mismatched textual and visual information are more likely to be fake (Zhou et al., 2020). Based on these observations, a focus of this paper is to model such gap between the textual and visual information, which we call *cross-modal inconsistency*.

Apart from cross-modal inconsistency, rumor detection can also benefit from knowledge graph (KG), which can provide faithful background knowledge to verify the semantic integrity of post contents. Previous works (Zhang et al., 2019; Wang et al., 2020) used KG to complement the post contents by various data fusion methods. However, they ignore the inconsistent information that may exist between the contents and the background

[*] Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1412–1423
November 7–11, 2021. ©2021 Association for Computational Linguistics

knowledge. For example, in Fig 1, it would be a great help to judge the the truthfulness of the post, given the background knowledge that sharks are very unlikely to occur in a subway. We use *content-knowledge inconsistency* to describe the uncommon co-occurring entities[1] spotting in the multi-modal post contents, such as "shark" and "subway" example in Fig 1.

In this paper, we consider both *cross-modal inconsistency* and *content-knowledge inconsistency*, which are also referred as *dual-inconsistency*. We analyze the data and find that the above dual-inconsistency shows a statistically significant distinction between rumor and non-rumor posts (see details in Sec. 4.2). Such findings in data analysis highlight that the dual-inconsistency can be indicative of the news veracity and should be considered when modeling. However, it is challenging to build models to capture such dual-inconsistency for two reasons. First, text, image and KG data have different structures, which can not be directly integrated. Second, there is no straightforward way to capture their various semantic relationships, especially the inconsistent relationships.

To address these issues, we propose a novel *knowledge-guided dual-inconsistency network* to capture the inconsistency information at the cross-modal level and the content-knowledge level simultaneously. Note that our framework does not require both types of inconsistency to be present to effectively detect rumors. In other words, either type of inconsistent information can be a strong feature to infer a piece of tweet is a rumor. Our framework mainly consists of two sub neural networks: one is to extract cross-modal differences between images and texts, excluding their modal-shared information; the other is to identify the abnormal entity pairs that co-occur in the post contents through measuring their KG representation distances. The two sub neural networks are tightly coupled to achieve the best performance. The contributions of our paper are three-fold:

- We propose a novel knowledge-guided dual-inconsistency network by modeling cross-modal and content-knowledge inconsistencies in one unified framework for multimedia rumor detection.
- To the best of our knowledge, we are the *first* to reveal that rumor posts tend to have larger entity distances on KG than non-rumors, which is a useful signal for rumor detection.
- We empirically show our proposed method can outperform the state-of-the-art baselines on two real world datasets.

2 Related Work

Textual and social contextual rumor detection. Most rumor detection models rely on *textual features*. Recent studies propose deep learning models to capture high-level textual semantics (Ma et al., 2016, 2018; Yu et al., 2017), outperforming traditional machine learning-based models (Zhao et al., 2015; Castillo et al., 2011). *Social context features* represent the user engagements on social media such as retweeting and commenting behaviours (Shu et al., 2019; Tian et al., 2020) and network structures (Wu et al., 2015). However, social context features are usually unavailable at the early stage of the news dissemination.

Multimedia rumor detection. Several recent models begin to explore the role of visual information (Cao et al., 2020). Jin et al. (2017) extracts and fuses multi-modal and social context features with attention mechanism. EANN (Wang et al., 2018) learns post representations by leveraging both the textual and visual information, using an adversarial method to remove event-specific features to benefit newly arrived events. Khattar et al. (2019) proposes a multi-modal variational autoencoder for rumor detection. Zhang et al. (2021) designs a multi-modal multi-task learning framework by introducing the stance task. However, these studies do not consider consistencies between multi-modal information as our work. While both SAFE (Zhou et al., 2020) and MCNN (Xue et al., 2021) consider relevance between textual and visual information, our work differs from theirs in that we distinguish modal-unique from modal-shared information, and also model inconsistencies between content and external knowledge.

Fact-checking with KG. Some studies (Ciampaglia et al., 2015; Fionda and Pirrò, 2018; Pan et al., 2018; Shi and Weninger, 2016) extract structured triples (head, relation, tail) from the post contents, and fact-check them with the faithful triples in KG. A limitation of such approach is that KG is typically incomplete or imprecise to cover the complex relations in the form of triple being extracted from the post.

[1] Note that entity inconsistency are not necessarily cross-modal as in this example.

Consider an extracted triple (Anthony Weiner, cooperate with, FBI) has two entities with the "cooperate with" relation, where both entities are available in KG, but the relation is not (Pan et al., 2018). For such cases, structured triple methods fail to make reliable predictions. By contrast, our method is still applicable, as quantifying entity inconsistencies do not require relations.

Knowledge-enhanced detection. A few studies use the external knowledge to supplement post contents to obtain better representations for rumor detection. A knowledge-guided article embedding is learned for healthcare misinformation detection by incorporating medical knowledge graph and propagating the node embeddings through knowledge paths (Cui et al., 2020). The multimodal knowledge-aware representation and event-invariant features are learned together to form the event representation in Zhang et al. (2019), which is fed into a deep neural network for rumor detection. A knowledge-driven multi-modal graph convolutional network (KMGCN) (Wang et al., 2020) is proposed to model the global structure among texts, images, and knowledge concepts to obtain comprehensive semantic representations. However, these methods don't consider the content-knowledge inconsistency information. Moreover, KMGCN is transductive, requiring the inferred nodes to be present at training time, and time-consuming due to graph construction and learning.

3 Methodology

3.1 Overview

As shown in Fig.2, our framework mainly consists of four components : (1) a *preprocessig component* to obtain entities and their representations; (2) a *cross-modal inconsistency subnetwork* for capturing the inconsistencies between images and texts for each post; (3) a *content-knowledge inconsistency subnetwork* for capturing the inconsistencies between the content and KG through entity distances; (4) a *classification layer* that aggregates various features and produces classification labels.

The data flow as follows. Given a social post with images and texts, we first extract entities and obtain the entity representations. The collection of entity representations are fed into the content-knowledge inconsistency subnetwork to get the knowledge-level inconsistency features. Meanwhile, the image and text data are provided into the cross-modal inconsistency subnetwork to decompose and produce cross-modal inconsistency features and modal-shared features. Then the above features are fused and fed into the classification layer to obtain final classification labels.

3.2 Preprocessing Component

We essentially follow the procedure in Wang et al. (2020) to extract entities from texts and images. For the text content, we use the entity linking solution TAGME[2] to link the ambiguous entity mentions in the text to the corresponding entities in KG. For the visual content, we utilize the off-the-shelf pre-trained YOLOv3 (Redmon and Farhadi, 2018) to extract semantic objects as visual words. The labels of detected objects, such as person and dog, are treated as entity mentions. These mentions are linked to entities in KG. In this paper, we take Freebase[3] as the reference KG. We then obtain the pre-trained entity representations publicly available from OpenKE[4] , which are trained with TransE (Bordes et al., 2013) on Freebase. An entity representation $e_l \in \mathbb{R}^{d_e}$. Thus, our model accepts quadruple inputs : Text, Image, Entity set, Pre-trained KG.

3.3 Cross-modal Inconsistency Subnetwork

This subnetwork consists of two encoders for texts and images, respectively, a decomposition layer to obtain the corresponding modal-unique features and modal-shared features, and a fusion layer to produce cross-modal inconsistency features.

Text and image encoding. We map texts and images into feature representations. For each text, all textual words are firstly mapped into embedding vectors $w_j \in \mathbb{R}^{d_w}$. Then, we utilize the bidirectional long short term memory (Bi-LSTM) network to encode the textual sequence into a representation vector. It maps the word embedding w_j into its hidden state $h_j \in \mathbb{R}^{d_0}$, where w_j denotes the embedding of the j-th word from a word sequence with length M. We concatenate $\overleftarrow{h_0}$ and $\overrightarrow{h_M}$ to obtain the hidden state of the textual content $h \in \mathbb{R}^{2d_0}$. After that, we encode the textual representation into a d-dimensional vector,

$$H_T = \mathbf{ReLU}(\boldsymbol{w_T} * h + \boldsymbol{b_T}), \qquad (1)$$

[2] TAGME is available at https://tagme.d4science.org/tagme/

[3] Freebase data dumps is available at https://developers.google.com/freebase/

[4] OpenKE is available at http://openke.thunlp.org

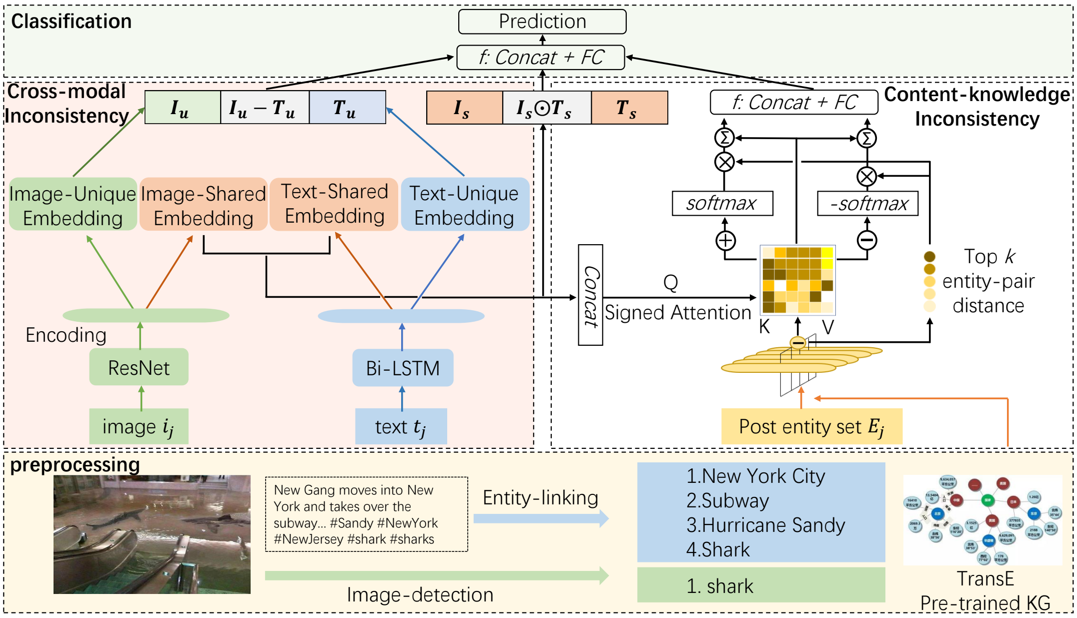


Figure 2. The framework of the proposed knowledge-guided dual-inconsistency network. It consists of four components: (1) bottom: the data preproceessing component to extract and represent entities from multimedia contents; (2) middle left: the cross-modal inconsistency subnetwork; (3) middle right: the content-knowledge inconsistency subnetwork; and (4) top: the rumor classification layer. *Concat* denotes the concatenation operating, and *FC* represents the fully-connected layer.

where $\boldsymbol{w_T}$ and $\boldsymbol{b_T}$ are learnable weights and bias parameters. Similarly, we encode an image into a d-dimensional vector with a pre-trained convolutional neural network (CNN),

$$H_I = \textbf{ReLU}(\boldsymbol{w_I} * \textbf{CNN}(Image) + \boldsymbol{b_I}), \quad (2)$$

where $\boldsymbol{w_I}$ and $\boldsymbol{b_I}$ are learnable parameters.

Multi-modal decomposition. Enlightened by the idea of projecting the multi-modal representations into different spaces (Xu et al., 2020), we break down the raw visual and textual representations into modal-unique spaces and modal-shared space. While a shared layer is proposed to extract modal-invariant shared features f^*_{shared}, an image or text layer is used to extract the corresponding modal-unique features f^*_{unique}, that is

$$\begin{aligned} I_s &= W_{shared} H_I \in \mathbb{R}^{d_s} \\ I_u &= P_I H_I \in \mathbb{R}^{d_u} \\ T_s &= W_{shared} H_T \in \mathbb{R}^{d_s} \\ T_u &= P_T H_T \in \mathbb{R}^{d_u} \end{aligned} \quad (3)$$

where H_I and H_T are the features of visual and textual modality. $W_{shared} \in R^{d_s \times d}$ and $\{P_I, P_T\} \in \mathbb{R}^{d_u \times d}$ are projection matrices for modal-shared space and modal-unique space, respectively.

To ensure that the decomposed modal-shared space is unrelated with the modal-unique spaces, the orthogonal constrain is introduced as:

$$\begin{aligned} W_{shared}(P_I)^T &= 0 \\ W_{shared}(P_T)^T &= 0 \end{aligned} \quad (4)$$

which can be converted into the orthogonal loss as

$$\mathcal{L}_o = ||W_{shared}(P_I)^T||_F^2 + ||W_{shared}(P_T)^T||_F^2, \quad (5)$$

where $|| \cdot ||_F^2$ denotes the Forbenius norm.

After obtaining two modal-unique features and two modal-shared features in Eqn.(3), we combine them as the cross-modal inconsistency representation f_{unique} and the overall modal-shared representation f_{share}, that is

$$\begin{aligned} f_{unique} &= [T_u; T_u - I_u; I_u] \\ f_{share} &= [T_s; T_s \odot I_s; I_s], \end{aligned} \quad (6)$$

where $\odot$ denotes the element-wise multiplication operation.

3.4 Content-knowledge Inconsistency Subnetwork

Here we introduce how to capture the content-knowledge inconsistency features.

Entity pair sorting. We measure their Manhattan distance for each pair of entity representations within a post and retain the top-k ($k = 5$) entity pairs and their corresponding distance values. Note that for a few posts where the number of entities is less than 4, we make a supplement with some pseudo entities whose representations are vectors with random values. We concatenate the pairwise entity representations to get the entity pair representation $\boldsymbol{EP}_i \in \mathbb{R}^{2d_e}$ $(i \in [1, k])$. Also we get the entity pair distance $dis^i \in \mathbb{R}$ $(i \in [1, k])$

Content-knowledge fusion with signed attention. To incorporate KG with post contents, we propose to fuse the top-k largest-distance entity pairs with the modal-shared contents with the attention mechanism. We use the modal-shared content as a query Q, and the entity pair representation EP as the value and key. Since the entity pairs may have multi-aspect correlations with the contents, we adopt the signed attention mechanism (Tian et al., 2020) to capture both positive and negative correlations simultaneously. In the traditional attention mechanism, if the correlations between query and keys are negative (i.e., their compatibility (e.g., dot product) value is negative), we would treat it as insignificant. However, such a negative correlation may represent the opposing semantics that can be beneficial to the rumor detection task. The signed attention mechanism, on the contrary, adds a “-softmax" operation using the opposite compatibility values between queries and keys as input to the softmax function to amplify the negative correlations. Thus the compatibility values would go through two channels, i.e., both traditional softmax and "-softmax" functions, to capture both positive and negative relationships between the modal-shared contents and the top-k largest distance entity pairs. We thus obtain two attention weights corresponding to the two channels, that is,

$$\begin{aligned} \boldsymbol{Q} &= \mathbf{Concat}(I_s, T_s) \\ \alpha^j_{pos} &= softmax(\frac{\boldsymbol{Q(EP_j)}}{\sqrt{2d_e}}) \\ \alpha^j_{neg} &= -softmax(-\frac{\boldsymbol{Q(EP_j)}}{\sqrt{2d_e}}) \end{aligned} \tag{7}$$

where the modal-shared feature Q is the concatenation of modal-shared features for images and texts. Both α^j_{pos} and α^j_{neg} denote the attention weights of the j-th entity pair but reflecting the positive and negative correlations, respectively. A larger α^j_{pos} (resp. α^j_{neg}) means that the entity pair is more positively (resp. negatively) semantically related to the content.

Meanwhile, an entity pair with a larger entity distance should influence the learning object more significantly. Following this intuition, we devise the final attention weight for each of the entity pairs by taking both of the factors into consideration and employ the weights to calculate the weighted sum of the entity pair representations, that is,

$$\begin{aligned} \beta^i_* &= \frac{dis^i \alpha^i_*}{\sum^k_{j=1} dis^j * \alpha^j_*} \\ f^*_{kg} &= \sum^k_{i=1} \beta^i_* (\boldsymbol{EP}_i) \\ f_{kg} &= Concat(f^{pos}_{kg}, f^{neg}_{kg}), \end{aligned} \tag{8}$$

where dis^i $(i \in [1, k])$ denotes the entity distance for the i-th entity pair. β^i_* $(* \in \{pos, neg\})$ is the distance-aware signed attention weights. f^*_{kg} $(* \in \{pos, neg\})$ is the positive/negative entity-pair embedding based on the signed attention weights. f_{kg} denotes the final semantic vector that represents the content-knowledge inconsistency features.

3.5 Rumor Classification Layer

At last, we concatenate the cross-modal inconsistency features, content-knowledge inconsistency features and the modal-shared features, and feed it into a fully-connected layer with Sigmoid activation function to obtain the predicted probability for instance i, that is,

$$\hat{y_i} = \sigma(\boldsymbol{w_f}[f_{unique} \oplus f_{share} \oplus f_{kg}] + \boldsymbol{b_f}) \tag{9}$$

where $\boldsymbol{w_f}$ and $\boldsymbol{b_f}$ are the weight and bias parameters. We then use cross entropy loss as the rumor classification loss:

$$\mathcal{L}_c = -\sum_i y_i log \hat{y}_i \tag{10}$$

where y_i is the ground truth label of the i-th instance. In addition, we also incorporate the orthogonal loss for multi-modal decomposition in Eqn.(5). Thus, the final total loss is

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_o \tag{11}$$

where λ is the weight of the orthogonal loss.

4 Experiments

In this section, we present the experiments to evaluate the effectiveness of our proposal.

	#Posts	#False	#True	#Images	# Entities/Post
Twitter	15557	10184	5373	410	5.536
Pheme	2374	686	1688	2374	5.363

Table 1: The statistics of the two datasets.

	Entity Distance		Image-text Similarity	
	Twitter	Pheme	Twitter	Pheme
Rumors	97.13	89.13	-0.058	-0.043
Non-rumors	90.20	82.89	0.041	0.091

Table 2: The average sum of the five largest entity distances and the average image-text similarity on two datasets.

4.1 Dataset

We conduct experiments on two real-world datasets, i.e., Twitter (Boididou et al., 2015) and Pheme (Zubiaga et al., 2017), both collected from Twitter. As one primary objective of our proposal is to incorporate the text and image information, we remove the data instances without any text or image. Moreover, we also remove the data instances from which no entities can be extracted, as at least one entity is required in our model. The statistics of the resulting datasets are shown in Table 1. Note that if there are multiple images attached to one post, we randomly retain one image and discard the others. For Twitter dataset, one image can be shared by various posts.

4.2 Preliminary Analysis of Dual Inconsistency

We conduct data analysis to validate that two inconsistencies have a statistically significant distinction between rumors and non-rumors.

Entity Distance Analysis We conduct entity distance analysis to show that the largest entity distance of a post is statistically different towards rumors and non-rumors. Specifically, we measure their Manhattan distance for each pair of entity representations within a post and retain the top-k ($k = 5$) largest distance values (as described in Sec. 3.4). The average sum of the five largest distances for all rumor and non-rumor posts are shown in Table 2. We can observe that, on average, the sum of entity distances for rumors is larger than that for non-rumors.

To statistically verify the observation, we make it as a hypothesis and conduct hypothesis testing. For each dataset, two equal-sized collections of rumor and non-rumor tweets are sampled. And two-sample one-tail t-test is conducted on the 100 data instances to validate whether there is sufficient statistical correlation to support the hypothesis. Let μ_f be the mean of five largest entity distances of the rumor collection and μ_r represent that of non-rumors. The null hypothesis is H_0, and the alternative hypothesis is H_1. The hypothesis of interest is:

$$\begin{aligned} H_0 &: \mu_f - \mu_r \leq 0 \\ H_1 &: \mu_f - \mu_r > 0 \end{aligned} \tag{12}$$

The results show that there are statistical evidences on both of the datasets. On Pheme, the result, t = 4.090, df = 90, p-value = 0.000047 (significance alpha= 5%), rejects the H_0 hypothesis. And the confidence interval CI is $[0.212, 42.112]$, the effect size is 0.826. The conclusion is similar on Twitter dataset.

Image-text Similarity Analysis We also conduct the image-text similarity analysis towards rumors and non-rumors. In particular, we first decompose the raw textual and visual representations to obtain image-unique and text-unique embeddings excluding their shared information (refer to Eqn. (3) in Sec. 3.3 for details), and measure their cosine similarity to get the image-text similarity. The average similarity results are shown in Table 2. We can observe that the similarity for rumors is smaller than that for non-rumors on both datasets, in line with our expectations. Moreover, we also perform the hypothesis testing and confirm there is statistical evidence on both datasets. Please see Appendix B for details. Our analysis shows that on each dataset, the rumors own distinct content-knowledge inconsistency and cross-modal inconsistency from non-rumors, which can be helpful for distinguishing rumors and non-rumors.

In the above data analysis as well as the methodology section, we consider top-k ($k = 5$) largest distances between entities, rather than averaging distances between all entity pairs, as the latter would weaken the contrast between rumors and non-rumors, as the gap between the average distances of non-rumors and rumors decreases significantly by the increase of k in preliminary analysis, where when $k > 5$, the gap is almost closed. This is because that even for rumors, there are almost always some consistent entities. For the example in Fig. 1, a shark appears in water is reasonable, and a subway station usually has elevators. We empirically show in the later Table 4 that considering top-5 achieves best results.

Method	Modality			Twitter				Pheme			
	Text	Image	KG	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
BERT	✓			0.835	0.821	0.81	0.815	0.819	0.809	0.726	0.765
Transformer	✓			0.791	0.772	0.791	0.781	0.774	0.755	0.648	0.697
TextGCN	✓			0.712	0.721	0.744	0.732	0.810	0.775	0.744	0.759
EANN	✓	✓		0.697	0.695	0.698	0.697	0.766	0.701	0.687	0.693
KMGCN	✓	✓	✓	0.825	0.813	0.788	0.800	0.812	0.775	0.753	0.764
Our Model	✓	✓	✓	**0.920**	**0.905**	**0.930**	**0.918**	**0.846**	**0.815**	**0.804**	**0.809**

Table 3: Results of comparison among different models on Twitter and Pheme Datasets.

4.3 Experimental Setup

The details of the two datasets and the preprocessing steps have been introduced in Sec. 4.1. We split the Pheme dataset into training, validation, and testing set with a split ratio of 6:2:2 without overlapping. For the Twitter dataset, we keep the same splitting scheme (approximately 13:1) in the raw data. In terms of parameter setting, the learning rate is {0.005, 0.0005}, batch size is {32, 128}. Our algorithms are implemented on Pytorch framework (Paszke et al., 2017) and trained with Adam (Kingma and Ba, 2015). The weight of the orthogonal loss is $\lambda = 1.5$. We use the pre-trained BERT (Wolf et al., 2020) as initial word embeddings for text encoding in our model. For other models that don't adopt BERT, we use GloVe [5] instead. We employ accuracy, precision, recall, and F1 as evaluation metrics. We adopt an early stop strategy and dynamic learning rate reducing for model training on both of the datasets.

4.4 Baselines

The baselines are listed as follows:

BERT (Devlin et al., 2019) is a pre-trained language model based on deep bidirectional transformers, and we use it to get the representation of the post text for classification.

Transformer (Vaswani et al., 2017) uses the self-attention mechanism and position encoding to extract textual features for sequence to sequence learning. We only use its encoder here.

TextGCN (Yao et al., 2019) uses a graph convolution network to classify documents. The whole corpus is modeled as a heterogeneous graph. It learns word and document embedding.

EANN (Wang et al., 2018) uses an event adversarial neural network to extract event-invariant features from images and texts for rumor detection.

KMGCN (Wang et al., 2020) is a state-of-the-art rumor detection model that uses a graph convolution network to incorporate visual information and KG to enhance the semantic representation.

4.5 Results and Discussion

Table 3 demonstrates the performance of all the compared models on two datasets. We can observe that our model significantly outperforms all the baselines in all the metrics, which confirms that considering the dual-inconsistency information would benefit the rumor detection task.

Among the three state-of-the-art textual representation models, BERT outperforms Transformer and TextGCN on both datasets, demonstrating its superior capability in capturing the textual semantics for rumor detection. Although EANN considers both visual and textual information, it performs not as well as BERT and TextGCN. The possible reason is that EANN uses CNN to extract the textual feature, which is not as powerful as Transformer and GCN. It indicates that textual representations play a crucial role in rumor detection. KMGCN achieves comparable and better performance compared to TextGCN. Since both of them adopt graph convolution networks as the backbone, it indicates that the image and knowledge features can provide complementary information and improve performance. We can attribute our proposal's superiority to two critical properties: (1) we model two types of inconsistent information, which are more suitable to rumor identification: (2) we adopt BERT as the initial text representation to capture textual semantics. We conduct ablation tests in the following subsection for validation.

4.6 Performance of the Variations

We investigate the effects of our proposed components by defining the following variations:

w/o Visual: the variant that removes visual information.

[5] GloVe: Global Vectors for Word Representation:https://nlp.stanford.edu/projects/glove/

Method	Twitter		Pheme	
	ACC	F1	ACC	F1
Our Model	**0.920**	**0.918**	**0.846**	**0.809**
-w/o Visual	0.836	0.813	0.806	0.751
-w/o BERT	0.905	0.893	0.830	0.787
-concat TV	0.905	0.897	0.808	0.763
-w/o KE	0.864	0.853	0.805	0.764
-mean KE	0.865	0.854	0.813	0.787
-rm 1 KE	0.915	0.909	0.822	0.780
-rm 2 KEs	0.874	0.866	0.821	0.774
-rm 3 KEs	0.871	0.868	0.814	0.773
-rm 1 KE pair	0.910	0.907	0.811	0.766
-rm 2 KE pairs	0.880	0.877	0.794	0.747
-rm 3 KE pairs	0.877	0.869	0.785	0.715

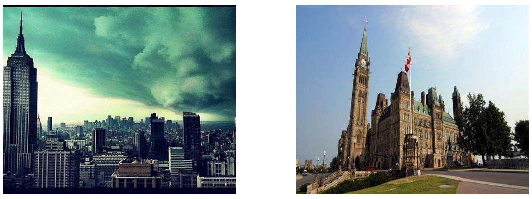
Table 4: Results of comparison among different variants on Twitter and Pheme datasets.

w/o KE: the variant that removes the content-knowledge inconsistency subnetwork.
mean KE: the variant that utilizes the mean pooling of the entity representations instead of the content-knowledge inconsistency features.
concat TV: the variant that concatenates the textual and visual representations instead of their cross-modal inconsistency and modal-shared features.
rm n **KE:** the variant that randomly removes n ($n \in \{1, 2, 3\}$) entities from the post entity set.
rm n **KE pair:** the variant that randomly removes top-n ($n \in \{1, 2, 3\}$) largest distance entity pairs from the post entity set.

The ablation study in Table 4 demonstrates that the proposed components: cross-modal inconsistency features and content-knowledge inconsistency features, are indispensable for achieving the best performance. Visual features and BERT representations can also improve the performance. To make a more fair comparison, we use the same input but alternate aggregating mechanisms instead of the inconsistency mechanisms. The results of *mean KE* and *concat TV*, lower than the proposed model, show that the inconsistency features are more effective than the aggregated features for rumor detection.

To verify the effectiveness of the knowledge information, we conduct the sensitivity analysis with varying number of entities and entity pairs. As shown in Table 4, when one or more entities are removed from the entity set of a post, the performance degrades. Similar trends can be observed when removing one or more entity pairs in the content-knowledge inconsistency subnetwork. It shows that considering the top-5 entity pairs achieves the best performance.

4.7 Qualitative Evaluation



(a) Zombie apocalypse approaches RT @thinkprogress: Sandy approaches NYC Sandy hurricane.

(b) NHL postpones Maple Leafs-Senators game after tragic shootings in Ottawa.

Figure 3. Two rumor cases detected by our model.

We analyze two rumor cases that our model can recognize accurately. They are from Twitter and Pheme, respectively. In Fig 3 (a), the extracted entity set is *{Zombie, Tropical cyclone, New York City, RT (TV network), ThinkProgress}*. The average sum of the five largest entity distances is 119.73, larger than the average sum of the rumors in Twitter (i.e., 97.13 shown in Table 2), implying the existence of content-knowledge inconsistency. Its image-text similarity value is 0.277, much larger than the average value for rumors (-0.058 in Table 2), indicating the image and text are well matched. In Fig 3 (b), it is obvious that the image and text are not well-matched, verified by its low image-text similarity value (only -0.133). The two cases help to confirm that our model can effectively capture the two types of inconsistent information for rumor identification.

5 Conclusion

We have revealed the necessity of capturing the inconsistent semantics for detecting rumors. We thus propose the knowledge-guided dual-inconsistency network, which involves the cross-modal inconsistency and content-knowledge inconsistency information in one unified framework. We have demonstrated our proposal's effectiveness in capturing and fusing both types of inconsistency features to achieve the best performance. Note that the inconsistency features can be easily plugged into other rumor detection frameworks to further improve the performance. In future work, we plan to explore more effective inconsistency features and devise a more explainable model.

Acknowledgements

This work was supported by the Natural Science Foundation of China (No.61976026, No.U1836215) and 111 Project (B18008).

References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–36.

Christina Boididou, Katerina Andreadou, Symeon Papadopoulos, Duc-Tien Dang-Nguyen, Giulia Boato, Michael Riegler, Yiannis Kompatsiaris, et al. 2015. Verifying multimedia use at mediaeval 2015. *MediaEval*, 3(3):7.

Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.

Juan Cao, Peng Qi, Qiang Sheng, Tianyun Yang, Junbo Guo, and Jintao Li. 2020. Exploring the role of visual content in fake news detection. *Disinformation, Misinformation, and Fake News in Social Media*, pages 141–161.

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 675–684. ACM.

Tong Chen, Xue Li, Hongzhi Yin, and Jun Zhang. 2018. Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 40–52. Springer.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis M Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PloS one*, 10(6):e0128193.

Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. DETERRENT: knowledge guided graph attention network for detecting healthcare misinformation. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 492–502. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Valeria Fionda and Giuseppe Pirrò. 2018. Fact checking via evidence patterns. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 3755–3761. ijcai.org.

Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 795–816.

Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia*, 19(3):598–608.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2915–2921. ACM.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pages 3818–3824. IJCAI/AAAI Press.

Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on Twitter with tree-structured recursive neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1980–1989, Melbourne, Australia. Association for Computational Linguistics.

Jeff Z Pan, Siyana Pavlova, Chenxi Li, Ningxi Li, Yangmei Li, and Jinshuo Liu. 2018. Content based fake news detection using knowledge graphs. In *International semantic web conference*, pages 669–683. Springer.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch.

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

Joseph Redmon and Ali Farhadi. 2018. Yolov3: An incremental improvement. *ArXiv preprint*, abs/1804.02767.

Baoxu Shi and Tim Weninger. 2016. Discriminative predicate path mining for fact checking in knowledge graphs. *Knowledge-based systems*, 104:123–133.

Kai Shu, Limeng Cui, Suhang Wang, Dongwon Lee, and Huan Liu. 2019. defend: Explainable fake news detection. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 395–405. ACM.

Tian Tian, Yudong Liu, Xiaoyu Yang, Yuefei Lyu, Xi Zhang, and Binxing Fang. 2020. QSAN: A quantum-probability based signed attention network for explainable false information detection. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1445–1454. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, pages 849–857. ACM.

Youze Wang, Shengsheng Qian, Jun Hu, Quan Fang, and Changsheng Xu. 2020. Fake news detection via knowledge-driven multimodal graph convolutional networks. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 540–547.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*, pages 651–662. IEEE.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3777–3786, Online. Association for Computational Linguistics.

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

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7370–7377.

Feng Yu, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. 2017. A convolutional approach for misinformation identification. AAAI Press.

Huaiwen Zhang, Quan Fang, Shengsheng Qian, and Changsheng Xu. 2019. Multi-modal knowledge-aware event memory network for social media rumor detection. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 1942–1951.

Huaiwen Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2021. Multi-modal meta multi-task learning for social media rumor detection. *IEEE Transactions on Multimedia*, pages 1–1.

Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 1395–1405. ACM.

Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. **SAFE**: Similarity-aware multi-modal fake news detection. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 354–367. Springer.

Arkaitz Zubiaga, Maria Liakata, and Rob Procter. 2017. Exploiting context for rumour detection in social media. In *International Conference on Social Informatics*, pages 109–123. Springer.

A On Reproducibility

In this section, we provide more details of the experimental setting and configuration to enable our work's reproducibility.

A.1 Baseline Implementation

We compared the proposed framework with five baseline methods discussed in Section 4.4, including BERT, Transformer, TextGCN, EANN, KMGCN. Baselines were obtained as follows:

- BERT: We use BERT with fine-tuning to detect rumors, which is available at https://github.com/huggingface/transformers.
- Transformer: we use the publicly available implementation at https://github.com/jayparks/transformer.
- TextGCN: we use the publicly available implementation at https://github.com/chengsen/PyTorch_TextGCN.
- EANN: we used the authors' implementation, which is available at https://github.com/yaqingwang/EANN-KDD18.
- KMGCN: we implemented the codes by ourselves. We followed the implementation details described in KMGCN except for choosing a different KG. Instead of using Probase and Yago in the original KMGCN, we used Freebase as the reference knowledge graph and acquired isA relation of the entities. The Freebase isA relation data dump is available at https://freebase-easy.cs.uni-freiburg.de/dump/

  The reason why we chose Freebase as the knowledge source is three-fold: (1) Freebase has a much larger-scale set of entities than Probase and Yago, which would facilitate the rumor detection task. (2) There are off-the-shelf pre-trained entity embeddings that can be used directly by our model; (3) for KMGCN, we need to use the same KG source as our model to make a fair comparison.

A.2 Implementation details for Our Model

The Twitter dataset is available at https://github.com/MKLab-ITI/image-verification-corpus, and the Pheme dataset is at https://figshare.com/articles/PHEME_dataset_of_rumours_and_non-rumours/4010619.

In preprocessing, we use three pre-trained models as follows:

- Entity linking: we use the existing entity linking solution TAGME to link the ambiguous entity mentions in the text to the corresponding entities in Freebase. TAGME is available at: https://tagme.d4science.org/tagme/.
- Image detection: we employ the YOLOv3 detector to search objects in each image. For the Pheme dataset, we use a pre-trained model provided in https://pjreddie.com/darknet/yolo/#demo. Due to the low image quality of the Twitter dataset, we employ the pre-trained YOLOv3 model and YOLOv3 detector pre-trained on the dataset that we have collected from the web and Open Image Dataset. We labeled about 50 different kinds of objects on the images we collected.
- Pre-trained entity representations: we use the entity representations publicly available at http://openke.thunlp.org. The scale of pre-trained embeddings is 86054151, and the embedding dimension is 50.

We conducted all the experiments on a server with three Intel(R) Xeon(R) Silver 4210 CPU @ 2.20GHz, 125 GB memory, 7.16 TB HDD and four GeForce RTX 2080Ti GPU cards.

In the training procedure, our proposed model's average run time is about 14260s for the Twitter dataset and about 3362s for the Pheme Dataset. The total number of trainable parameters is 25179753. The method of choosing hyperparameter values is manual tuning, and we use F1 as the criterion to select hyperparameters values.

B Preliminary Data Analysis: Image-text Similarity

Due to the main manuscript's space limits, we present the hypothesis testing of image-text similarity here to supplement Sec. 4.2.

The rumor and non-rumor collections are set the same as Section 4.2. Let θ_f be the mean of cosine-similarity of the rumor collection and θ_r represent that of non-rumors. The null hypothesis is H_0^s, and the alternative hypothesis is H_1^s. The hypothesis of interest is:

$$\begin{aligned} H_0^s &: \mu_f - \mu_r \geq 0 \\ H_1^s &: \mu_f - \mu_r < 0 \end{aligned} \tag{13}$$

The results show that there are statistical evidences on both of the datasets. On Twitter dataset, the result, t = -3.7925, df = 97, p-value = 0.000129 (significance alpha= 5%), rejects the H_0 hypothesis. And the confidence interval CI is $[-0.425888, -0.002151]$, the effect size is -0.7662. We also found statistical evidences on Pheme dataset. On Pheme dataset, the result, t = -7.9051, df = 94, p-value = 2.4769×10^{-12} (significance alpha= 5%), rejects the H_0 hypothesis. And the confidence interval CI is $[-0.317446, -0.001603]$, the effect size is -1.5970.