International Journal of Data Warehousing and Mining
Volume 18 • Issue 1

Improving Rumor Detection by Image Captioning and Multi-Cell Bi-RNN With Self-Attention in Social Networks

Jenq-Haur Wang, National Taipei University of Technology, Taiwan*
https://orcid.org/0000-0002-6076-7380

Chin-Wei Huang, National Taipei University of Technology, Taiwan

Mehdi Norouzi, University of Cincinnati, USA

ABSTRACT

User-generated contents in social media are not verified before being posted. They could bring many problems if they were misused. Among various types of rumors, the authors focus on the type in which there's mismatch between images and their surrounding texts. They can be detected by multimodal feature fusion in RNNs with attention mechanism, but the relations between images and texts are not well-addressed. In this paper, the authors propose to improve rumor detection by image captioning and RNNs with self-attention. Firstly, they utilize the idea of image captioning to translate images into the corresponding text descriptions. Secondly, these caption words are represented by word embedding models and aggregated with surrounding texts using early fusion. Finally, multi-cell bi-directional RNNs with self-attention are used to learn important features to identify rumors. From the experimental results, the best F-measure of 0.882 can be obtained, which shows the potential of our proposed approach to rumor detection. Further investigation is needed for data in larger scale.

KEYWORDS

Bi-Directional Gated Recurrent Unit, Multimodal Feature Fusion, Rumor Detection, Image Captioning, Self-Attention Mechanism, Sequence-to-Sequence Model

1. INTRODUCTION

With the rapid development of information and communication technology, people can easily share their opinions and get the latest news from social network platforms. Since these user-generated contents are not verified before being posted, people cannot tell if they are real or false. Thus, the unverified false messages are considered as rumors since they are annoying in various aspects of our daily lives. For example, people could receive messages that masquerade as being sent from the government or companies, asking to provide personal information. People can only tell if these messages are real or false by checking if the text contents are relevant or not. To distinguish real

DOI: 10.4018/IJDWM.313189

*Corresponding Author

messages from false ones, rumor detection has been an increasingly important research topic in social media. Nowadays, third-party fact-checking services such as FactCheck.org and Snopes.com have been used for message verification. These fact-checking services usually require manual labeling which needs lots of human efforts. In the face of rapid information dissemination, these services alone cannot be effectively provided in time.

There could be numerous ways to disguise the unverified false messages or rumors as real messages. Since it's very common to post images and texts in the same post in social media, in this paper we define rumors as the message type when there's mismatch between multimedia contents and the surrounding texts. Our research problem for rumor detection is defined as: given a social media post with images and their corresponding surrounding texts, we want to determine if there's a mismatch between semantic information in the images and the surrounding texts. In recent research, deep learning methods are widely used to construct the model and to learn features for rumor detection. For example, as the baseline in our experiments, the method proposed by Jin et. al (2017) utilized a RNN with attention mechanism (att-RNN) to fuse multimodal features, including texts and images. They achieved an accuracy of 68.2% for the Twitter dataset in MediaEval task. The visual neuron attention mechanism gives each neuron different weights for different words. However, the relations between image visual features and text features are not well-addressed. In order to better address the relations between images and the surrounding texts, in this paper, we propose to improve rumor detection by image captioning and multi-cell RNNs with self-attention. Firstly, it's very important if we can aggregate multimodal contents as an effective feature to tell the mismatch between texts and images. Instead of simply adjusting weights of different visual features by the attention mechanism from the RNN results of texts, we first translate images into the most relevant caption words as a more coherent way of feature representation. This helps to closely connect the semantic meanings of images and texts. Secondly, we design a novel type of multi-cell bidirectional RNNs which combine self-attention mechanism for identifying more important features from different sources. Finally, different feature fusion approaches are used to improve the performance of rumor detection.

The main contributions of this paper are as follows: Firstly, we propose a novel multimodal feature fusion approach to rumor detection based on image captioning model that represents image semantics in textual descriptions. The sequence-to-sequence model can extract more meaningful descriptions from images than simple convolutional approaches. To the best of our knowledge, our proposed method is the first to apply image captioning in extracting image semantics for rumor detection. Secondly, instead of one single layer which might not fully capture the relations among words, we design a novel way of stacking bidirectional RNNs called Multi-cell Bi-RNN, which adds more cells in each individual direction of forward and backward passes to learn more deeply in each neuron. A better performance can be obtained than the baseline model. This shows the potential of our proposed approach to rumor detection.

The rest of the paper is structured as follows: In Sec. 2, we provide a review of related previous research works. Then the proposed method is presented and discussed in Sec.3. In Sec.4, we show the experimental results and compare with existing methods. Finally, we conclude the paper in Sec.5.

2. RELATED WORK

With the rapid growth of various kinds of false messages, people become overwhelmed by these misinformation or disinformation. Unverified false messages are called rumors since they might be very similar to real messages. It's hard to tell even for human beings. Thus, we need more information for assessment. There have been many third-party fact-checking services such as FactCheck.org[1] and Snopes.com[2] for verifying the authenticity of messages. But these fact-checking services usually need manual labeling, which are time-consuming and requires lots of human efforts. Recently, machine learning methods are used to train classifiers for rumor detection. For example, Ma et al. (2016) first proposed to detect rumors from microblogs with RNNs. Yu et al. (2017) proposed Convolutional

Neural Networks (CNNs) for misinformation identification. Sampson et al. (2016) leveraged the implicit linking structure between conversations in social media for emergent rumor detection. Chen et al. (2018) proposed to use deep attention-based RNNs for early rumor detection. Their emphasis is on the emergent stage when there're only a few instances of rumors available. Ma et al. (2018) further proposed a joint framework for rumor detection and stance classification by multitask learning. Their experimental results demonstrate the improvement over individual task with the help of inter-task connections. Jin et al. (2017) investigated the multimodal fusion in Long Short-Term Memory (LSTM) for rumor detection using images and texts. The weights of words in texts from attention output of LSTM are extracted, and an elementwise multiplication to image feature vectors is applied to get the fused information. However, the elementwise correspondence between word vectors and image feature vectors are not clear. Li et al. (2022) proposed a multi-modal model that utilized granularity fusion and image knowledge to help rumor detection. Zhou et al. (2022) proposed a multi-task learning framework which includes domain adaption for extracting visual features from images to improve early rumor detection. Ying et al. (2021) proposed a multilevel multimodal approach to fake news detection. Relations between multi-level semantics of texts extracted by Transformer and image features extracted by ResNet are incorporated in a cross-attention network. Zhao et al. (2018) proposed a sequence-to-sequence based image captioning model that integrates CNN-based image encoder and LSTM-based text decoder. The idea of image captioning was used in Meel and Vishwakarma (2021) for fake news detection. It's an ensemble method which votes by the decisions from different approaches. Specifically, they used cosine to measure the similarity between image caption and news body. But the length of news and captions might be very different.

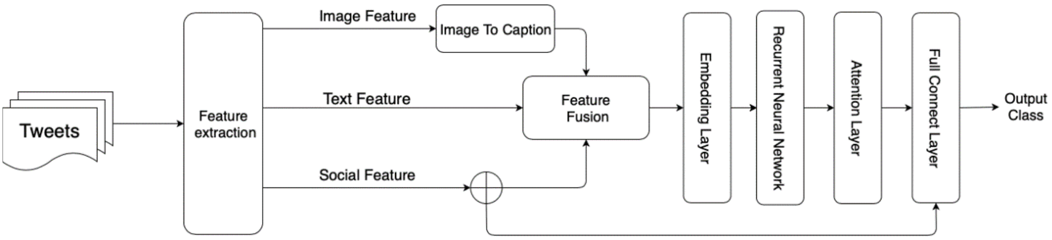
In this paper, we focus on identifying the mismatch between images and surrounding texts. We propose to improve rumor detection by incorporating image captioning for extracting more meaningful visual semantics from images. Then, we focus on feature-level (early) fusion where text features are concatenated with the generated image captions, and represented by word embeddings and further combined with other social features as the representation of the whole social media post. Finally, these fused features are input to multi-cell bi-directional RNNs with self-attention mechanism to identify important features for rumor detection.

3. THE PROPOSED METHOD

We propose a deep attention network model with image captioning and multimodal feature fusion for rumor detection. The architecture is shown in Figure 1.

As shown in Figure 1, images, texts, and social features are first extracted from input tweets by Feature Extraction module. Then, image features are converted to captions by the Image Caption module which consists of CNNs as image encoder and Seq2Seq model to generate descriptive texts. These are further combined with text features and social features by Feature Fusion module. All features will be encoded by Embedding layer, which will then be input to different types of RNNs. Finally, with the Attention layer, we obtain different feature weights for the classification of rumors

Figure 1. System architecture of the proposed method for rumor detection



by a Fully Connected layer. Note that feature fusion of social features could occur in two places: feature-level (early) fusion before the Embedding layer, and decision-level (late) fusion at the Fully Connected layer. Next, we describe each module in more details.

3.1 Feature Extraction

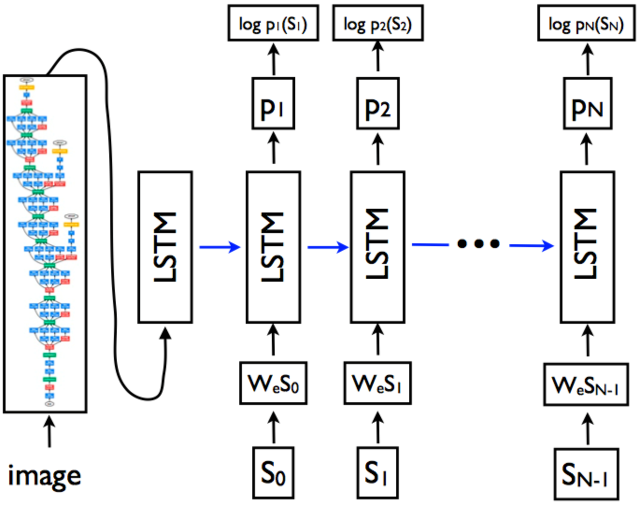
Social media posts such as Tweets consist of three major parts: text contents, images, and social information. First, text contents are extracted as the major feature people used to express the meaning of the post. We use RNNs to extract the contextual information from text contents. Then, in many social media posts, images are usually included to visually show information related to text content. We extract image features and convert to its corresponding short descriptions by Image Caption module. The idea is to extract the semantic meaning inherent in images. Furthermore, we consider various characteristics in social media posts as social features including text sentiments, hashtags, and user characteristics. Users might express their opinions on related topics in tweets. The sentiment polarity of user opinions can be identified by Sentiment Analysis module. It is done by matching words with SentiWordNet[3], whose sentiments are averaged as the overall polarity of the text. In social media, hashtags are provided by users, which highlight the topics of content. Also, people interact with each other by making friends, following others, posting articles, and replying others. To facilitate fair comparison, we consider the same five features as in Jin et al. (2017) including: the number of friends, followers, the ratio of friends in followers, the number of tweets, and if the account has been verified by Twitter or not. These extracted user characteristics are integrated with sentiments and tags as the social features.

3.2 Image Captioning

In this paper, a sequence to sequence approach to image captioning is derived from the model by Vinyals et al. (2015) as shown in Figure 2.

As shown in Figure 2, a CNN image encoder based on Inception Net model (Szegedy et al., 2015) by Google is used for learning image features. With the 42-layer structure, the InceptionNet model includes four different sizes of convolutional kernels, which helps to learn image characteristics in various scales. Then, the output image feature vector will be input to a LSTM-based text decoder,

Figure 2. System architecture of the image captioning module (Vinyals et al., 2015)



followed by the one-hot encoding of text features which was input one at a time. The idea is to predict the next output word given the image feature and previous text contents. At each point of time, the one-hot encoding of a word S_i is input to LSTM where the corresponding probability distribution *log* $P_i(S_i)$ of words is estimated. With the idea of beam search, we only keep the candidate terms with the top *k* probability scores at each point of time. This can better approximate the real probability of the descriptions given the image.

3.3 Feature Fusion

After extracting three different types of features from tweets, we propose a feature fusion approach that better integrates multimodal features to help improve the classification performance.

Given the very diverse nature of different features, we first concatenate the features into one single vector as follows:

Feature tuple = [text encoding, image caption encoding, tag encoding, sentiment polarity encoding, user characteristic encoding]

Text contents are encoded using word embedding models such as Word2Vec to represent words as vectors in 300 dimensions. After being converted to text descriptions by Image Captioning module, images are represented as vectors of the same size as the text encoding. Then, we represent the extracted hashtags and the sentiment polarity of tweets by one-hot encoding. The binary value of 0 and 1 represents the occurrence of each hashtag. For the sentiments, we only consider three classes: positive when the average sentiment score is larger than 1, negative when the score is less than 0, and neutral when the score is between 0 and 1. Finally, we concatenate all the above feature encodings into one single feature vector.

Social features include: the number of friends, followers, the ratio of friends in followers, the number of tweets, and if the account has been verified by Twitter or not, which have very different forms from text and image caption features. In this paper, we consider two different variants of fusion techniques for social features: *feature-level (early) fusion*, and *decision-level (late) fusion*. In feature-level fusion, the raw values of the above-mentioned five features are used as the social features. Then, to make the dimension comparable to that of the word embedding in the Embedding layer, we use an autoencoder to compress the social feature into vectors in 300 dimensions. They are concatenated with text and image caption encodings as the overall feature vector for classification. On the other hand, for decision-level fusion, we only concatenate the raw values of social features to the outputs of RNN and attention layers for classification in the Fully Connected layer.

3.4 Recurrent Neural Networks Structures

In this paper, we design a different way of stacking multiple layers of bi-directional recurrent neural networks for rumor detection. GRUs are used instead of the LSTM networks.

Conventionally, single-layer bi-directional RNN (Bi-RNN) was first proposed by Schuster and Paliwal (1997). It includes two separate one-directional layer: forward pass, and backward pass, which connect to the same output layer, as in Fig. 3.

As shown in Figure 3, input w1 and w3 are separately processed by forward and backward layers. Then, individual outputs w4 and w6 are combined in the output layer, which can be the sum or concatenation of separate outputs.

First, we compare different structures of stacking multiple single-layer Bi-RNNs. It's common to design a Multi-layer Bi-RNN which applies multiple bi-directional RNNs in order. Specifically, two Bi-RNNs can be stacked in Figure 4.

As shown in Figure 4, two Bi-RNNs are stacked in the sense that the second one will only be run after the first has completed. The output of the first Bi-RNN will be input to the second Bi-RNN.

Figure 3. The architecture of the single-layer bi-directional RNN (Schuster and Paliwal, 1997)

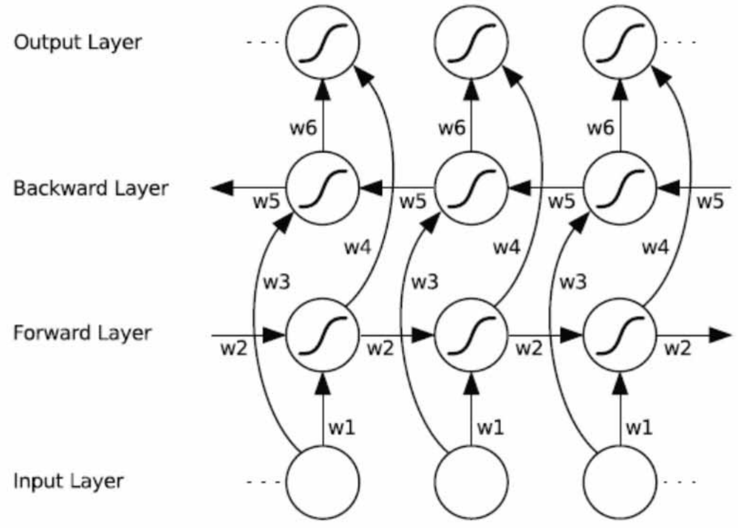
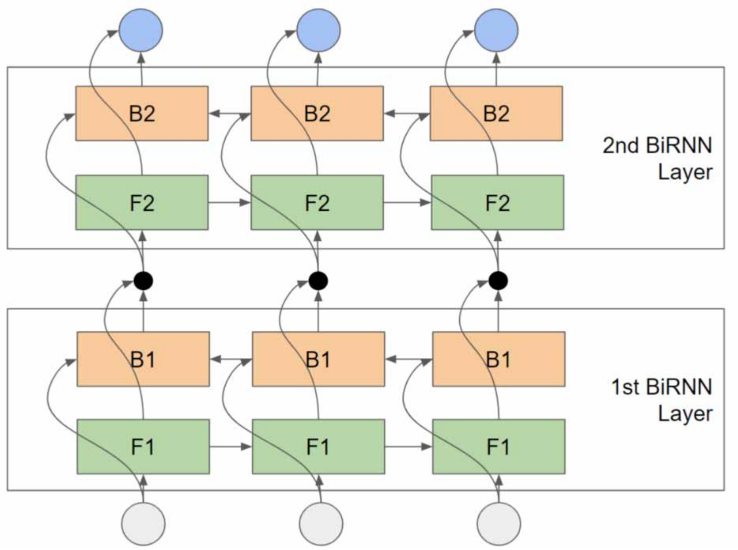


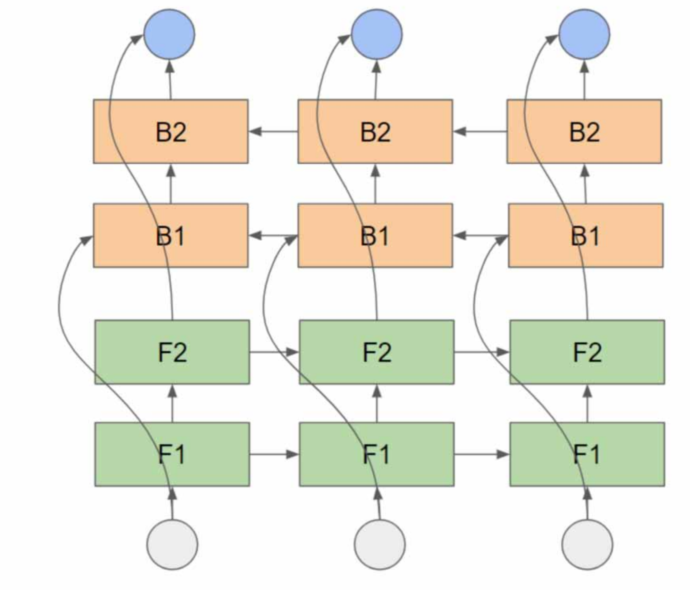
Figure 4. The architecture of the Multi-layer bi-directional RNN



In this paper, we propose a slightly different architecture called Multi-cell Bi-RNN where the forward and backward layers are separately stacked multiple times. An example case of combining two Bi-RNNs as Multi-cell RNNs is shown in Figure 5.

As shown in Figure 5, this architecture increases the number of cells in each of the forward and backward layers. Each input will be processed independently in separate direction through multiple cells. The combination of outputs from two directions only occurs at the output layer of the whole network. For example, the inputs are fed separately into the forward layer through F1 and F2, and the backward layer through B1 and B2. Then, the outputs from F1 and F2 will only be combined with the outputs from B1 and B2 at the final output layer.

Figure 5. The architecture of the Multi-cell bi-directional RNN



3.5 Attention Layer

Attention mechanism has been applied in RNNs to obtain the relative importance of words within text contents. In this paper, we apply the idea of self-attention in our proposed Multi-cell Bi-RNNs. The major difference of self-attention is that no external information is needed. The core idea of self-attention is the scaled dot-product attention, which is a variant of dot-product attention (Luong et al., 2015). It has been shown to be much faster than additive attention (Bahdanau et al., 2015) in single-layer feed forward neural networks by Vaswani et al. (2017) and Tan et al. (2018). Given the query vectors Q, keys K, and the values V, the attention score for scaled dot-product attention can be calculated by the following equation:

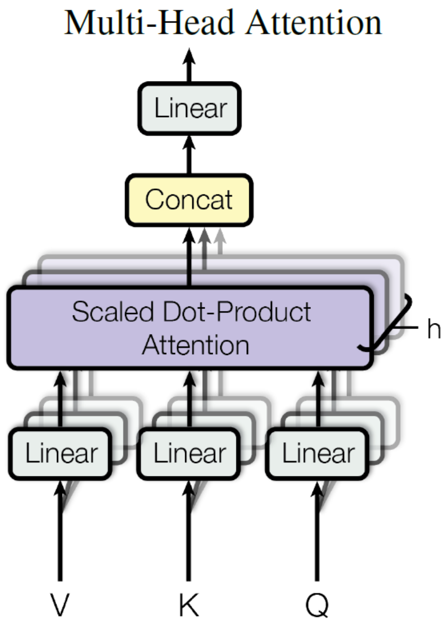
$$Attention\left(Q,K,V\right)=softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where d_k is the dimension of keys. When the dimension becomes larger, the dot products of Q and K will grow larger. To counter such effect, the scaling factor of (1/sqrt(d_k)) is used. After normalized by softmax function, the weight is multiplied by V to update the values. To speed up the computation, a multi-head attention is applied as shown in Figure 6.

As shown in Figure 6, instead of calculating a single attention function, we linearly project Q, K, and V *h* times with different projections, and then the scaled dot-product attention function is performed in parallel. Then, these are concatenated and projected to get the final values. The equation for multi-head attention is as follows:

$$MultiHead\left(Q,K,V\right)=Concat\left(head_1,\ldots,head_h\right)W^O$$

Figure 6. The idea of multi-head attention



where:

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \quad (2)$$

With self-attention mechanism, dependencies between each word will be calculated. This can avoid the gradient vanishing problem in RNNs when the path is too long. Thus, in this paper, we will use the results of self-attention for final classification at the Fully Connected layer.

4. EXPERIMENTS

To evaluate the performance of our proposed method, we need two types of datasets: one for image captioning, and the other for rumor detection. Firstly, for rumor detection, we used the Twitter dataset for Verifying Multimedia Use task in MediaEval 2015, 2016 (Boididou et al., 2016) since it has been verified by Twitter. Secondly, since Microsoft COCO 2014 dataset (Tan et al., 2018) was widely used in image tasks, such as image recognition and image feature detection, they can meet our needs for image captioning for the following reasons. First of all, image captioning is usually used for identifying objects as the major components comprising an image. On Twitter, users usually share images together with the related opinions in surrounding texts. To tell if the texts are consistent with the major components in images, image captioning can be used. In MS COCO 2014 dataset, although image captions might not directly address the contents in tweets, the image objects represent the common ideas in daily life, where tweets might also discuss about. They are useful to mimic human image understanding. Thus, in this paper, the MS COCO dataset was used to train semantic relations between images and texts. The statistics of the two datasets are shown in Table 1.

For performance evaluation of image captioning, we use the Bilingual Evaluation Understudy (BLEU) metric. Bilingual Evaluation Understudy (BLEU) was first proposed by Papineni et al. (2002)

Table 1. Statistics in the two datasets for our experiments

	MS COCO 2014 dataset		Twitter Rumor Detection dataset	
	Images	Captions	Real Events	Fake Events
Training	82,783	413,915	189	157
Test	36,454	182,270	21	24
Total	119,237	596,185	210	181

for evaluating the performance of machine translation. It's defined as the geometric mean of the modified *n*-gram precision P_n, using *n*-grams up to length *N* and weights summing to one as follows:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} W_n \log p_n\right)$$

$$BP = \begin{cases} 1 & if\ \ c > r \\ e^{1-r/c} & if\ \ c \le r \end{cases} \tag{3}$$

where *c* denotes the length of the candidate, and *r* denotes the effective reference corpus length. The modified *n*-gram precision P_n denotes the sum of clipped *n*-gram counts divided by the number of candidate *n*-grams as follows:

$$P_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}\left(n - gram\right)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count_{clip}\left(n - gram'\right)} \tag{4}$$

where the clipped counts are calculated as follows:

$$Count_{clip} = \min\left(Count, Max\_Ref\_Count\right) \tag{5}$$

For rumor detection, since we focus on binary classification, we use performance evaluation metrics including accuracy, precision, recall and F-measure. In order to validate the robustness of the evaluation results, we further use student's T-test to verify the statistical significance. The accuracy measure indicates the ratio of correctly predicted instances. It measures the proportion of correctly classified documents to the total number of documents available for classification. F1 Score is also called F-Measure, which can be calculated as follows:

$$F1\ Score = \frac{2 * \left(Precision * Recall\right)}{Precision + Recall} \tag{6}$$

where Precision refers to the percentage of observations that are relevant and Recall (also known as sensitivity) refers to the percentage of total relevant observations that are correctly classified. The F1 score conveys the weighted average between precision and recall. It is usually more useful than accuracy especially in highly imbalanced datasets. Student's T-test (Pearson et al., 1990) was used as a statistical significance test as follows:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\tilde{s}_1^{\,2}}{n_1} + \frac{\tilde{s}_2^{\,2}}{n_2}}} \tag{7}$$

where S is the standard deviation. We can calculate the corresponding P value to evaluate the effectiveness of the model.

4.1 Effect of Image Captioning

First, we evaluated the effects of image captioning on the two datasets with BLEU score. In MSCOCO 2014 dataset, there are five short descriptions for each image. In this experiment, we randomly selected three descriptions for training, and two remaining for testing.

As shown in Figure 7, for MSCOCO 2014 dataset, the BLEU scores for unigrams and bigrams are 0.695 and 0.51, respectively. This is similar to the results in Xu et al. (2015). Also, the BLEU scores for MSCOCO 2014 dataset are better than those of MediaEval 2015, 2016, when *n*-grams (*n*=1 to 3) are used. After inspecting the Media 2015, 2016 dataset in details, we found surrounding texts that might not necessarily describe the image. Thus, the images are not suitable to be used in our experiment. Since the goal of our image captioning model is to learn the semantic relations between texts and images, in this paper, we use the model trained from MSCOCO 2014 dataset for Image Captioning.

4.2 Effect of Word Embedding

To compare the effect of word embedding on rumor detection, we carried out experiments with two different embedding models: random initialization, and pretrained Word2Vec model. For random initialization, we randomly generated a numeric vector for each term consisting of values between -1 and 1. Then, weights in the vector will be updated in the neural network (denoted as Random_Update). For pretrained Word2Vec model, we use the model parameters pretrained with Google News as the initial weights. Then, we compare the performance when the model is updated (denoted as Word2Vec_Update) or not (denoted as Word2Vec) as follows in Table 2.

Figure 7. The evaluation result of image captioning

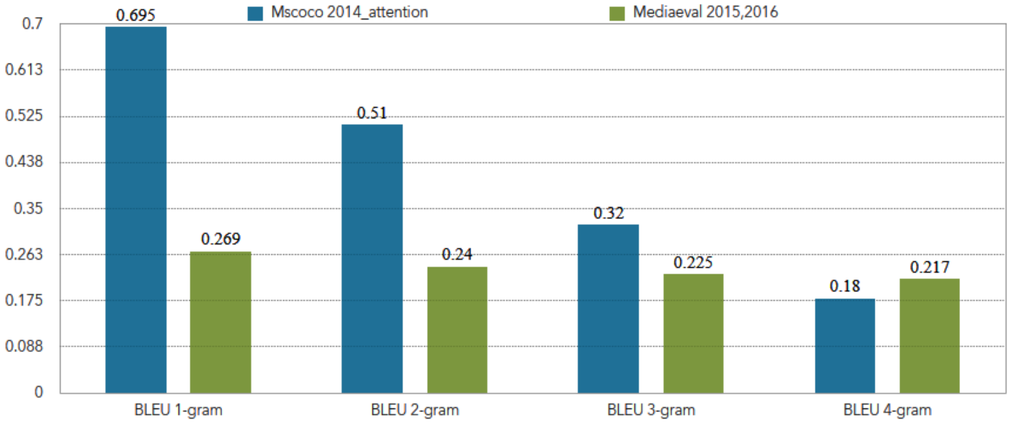


Table 2. The effect of word embedding on rumor detection

Word embedding	Accuracy	Precision	Recall	F-measure
Word2Vec	0.72	**0.811**	0.796	0.803
Word2Vec_Update	0.574	0.738	0.57	0.643
Random_Update	**0.749**	0.786	**0.861**	**0.822**

As shown in Table 2, we can observe better performance for random initialization, with the best F1 score of 0.822. It's slightly better than using the pretrained Word2Vec model with the F1 score of 0.803. After careful inspection, we observed that: Firstly, since the pretrained Word2Vec embedding was trained from Google News articles, they learned some contexts of words in news articles. Although this might be different from the word distributions for tweets, it's still possible to learn some inter-word relations which applies to tweets. Secondly, for unknown terms in the pretrained Word2Vec model, they will be ignored which lead to inferior performance. Also, we observed that the pretrained Word2Vec embedding without updates performed much better than with updates. This verified that there are some relations between word vectors in the pretrained model. Updating these vectors using data in different domains will lose the original meanings trained from the news articles. It validates the effects of word embedding models for the classification performance.

4.3 Effect of RNN Structures

To evaluate the effects of RNN structures on rumor detection, we compared the performance of our proposed multi-cell RNNs with different variants when using only text features as in Table 3.

As shown in Table 3, when considering text features only, the best performance can be obtained for both attention-based Multi-layer Bi-RNN and Multi-cell Bi-RNN, with a F1 score of 0.816. Combining attention mechanism with multiple layers for Bi-RNN outperform single-layer one. This shows that when stacking more Bi-RNN layers, it's possible to learn more relations between words. Attention mechanism further emphasizes the relative importance among words. Also, all three types of the attention-based Bi-RNNs greatly outperform single-layer LSTM-based networks by Jin et al. (2017). This shows the advantage of stacking multiple layers of bidirectional RNNs with GRU units when combined with self-attention mechanism.

4.4 Effect of Feature Fusion

In addition to the various designs of RNN structures, we further investigate the effect of multimodal feature fusion with different combinations of features. Firstly, we evaluated the performance when combining all features as in Table 4.

As shown in Table 4, after combining all features, Multi-cell Bi-RNNs achieve the best performance with F-measure of 0.882, followed by single layer Bi-RNN and Multi-layer Bi-RNNs. From our analysis, we can see the advantage of Multi-cell Bi-RNNs since after being processed by

Table 3. The effect of RNN structures on rumor detection (with only text features)

Model	Accuracy	Precision	Recall	F-measure
Jin et al., 2017 (only Text)	0.532	0.598	0.541	0.568
Single-layer BRNN (only Text)	0.738	0.774	0.826	0.799
Multi-layer BRNN (only Text)	**0.739**	**0.776**	0.861	**0.816**
Multi-cell BRNN (only Text)	0.724	0.74	**0.91**	**0.816**

Table 4. The effect of multimodal feature fusion on rumor detection (with all features)

Model	Accuracy	Precision	Recall	F-measure
Jin et al., 2017 (All)	0.682	0.78	0.615	0.689
Single-layer BRNN (Early fusion)	0.8	0.81	0.91	0.859
Multi-layer BRNN (Early fusion)	0.77	0.776	0.924	0.843
Multi-cell BRNN (Early fusion)	**0.827**	**0.816**	**0.959**	**0.882**

deeper hidden layers with more hidden neurons in each separate direction, the relation among features in a sequence of words can be better learned. On the other hand, for Multi-layer Bi-RNN, when we stacked two separate Bi-RNNs, the first one already learned the relations among words in texts. Then, when it comes to the second Bi-RNN, since the input is different from the raw text, there will be not much difference from the outputs of the first one. That also verifies why Multi-layer Bi-RNN has similar performance to single Bi-RNN. To further validate these results, we further checked with the statistical test by Student's T-test between Multi-layer and Multi-cell Bi-RNNs, the p-value is 0.039 which shows the statistical significance of their difference.

To compare the effects of two different fusion strategies, the performance evaluation for different types of RNNs using early and late fusion are shown in Table 5.

As shown in Table 5, the best performance can be obtained for feature-level (early) fusion for Multi-cell Bi-RNNs with a F1 score of 0.882. Even for single layer Bi-RNNs, early fusion helps to improve the performance from 0.828 to 0.859. This shows the advantage of early fusion since when social features are combined in early fusion stage before entering Bi-RNNs, the relation between social features can be learned together with other features, with their weights being adjusted by the attention mechanism.

4.5 Effect of Attention Mechanism

To verify the effects of attention mechanism with Bi-RNNs on the performance of rumor detection, we implemented the Hierarchical Attention Network (HAN) architecture proposed by Yang et al.

Table 5. The effect of feature fusion strategies on rumor detection

Model	Accuracy	Precision	Recall	F-measure
Single-layer BRNN (Early fusion)	0.8	0.81	0.91	0.859
Multi-layer BRNN (Early fusion)	0.77	0.776	0.924	0.843
Multi-cell BRNN (Early fusion)	**0.827**	0.816	**0.959**	**0.882**
Single-layer BRNN (Late fusion)	0.757	0.789	0.872	0.828
Multi-layer BRNN (Late fusion)	0.772	0.772	0.94	0.848
Multi-cell BRNN (Late fusion)	0.801	**0.834**	0.879	0.856

(2016) and compared its performance on rumor detection with our proposed method. The architecture of that paper is shown in Figure 8.

As shown in Figure 8, there are two layers in hierarchical attention network: word attention, and sentence attention. To make this architecture work on our rumor detection dataset, we assume that all tweets discussing the same event are considered as the texts describing that event. The experimental results using only text features are shown in Table 6.

As shown in Table 6, our proposed attention-based Bi-RNN achieves the best performance with a F-measure of 0.799, which is 10% higher than that of HAN. Although HAN is better than Jin et al. (2017), our proposed method still outperforms that of HAN. Next, we further compared feature fusion of all features as in Table 7.

As shown in Table 7, we can observe the best performance for early fusion in Bi-RNNs with a F-measure of 0.827, which is 23.3% improvement over HAN. We observed that since the tweets are more subjective, they might not contain complete descriptions of the related events. It might cause hierarchical attention networks (HAN) not being able to learn the whole meaning with syntactic structures such as words and sentences. This further validates the superior performance of our proposed attention-based Bi-RNN than hierarchical attention networks.

Figure 8. The architecture of hierarchical attention network (HAN) (Yang et al., 2016)

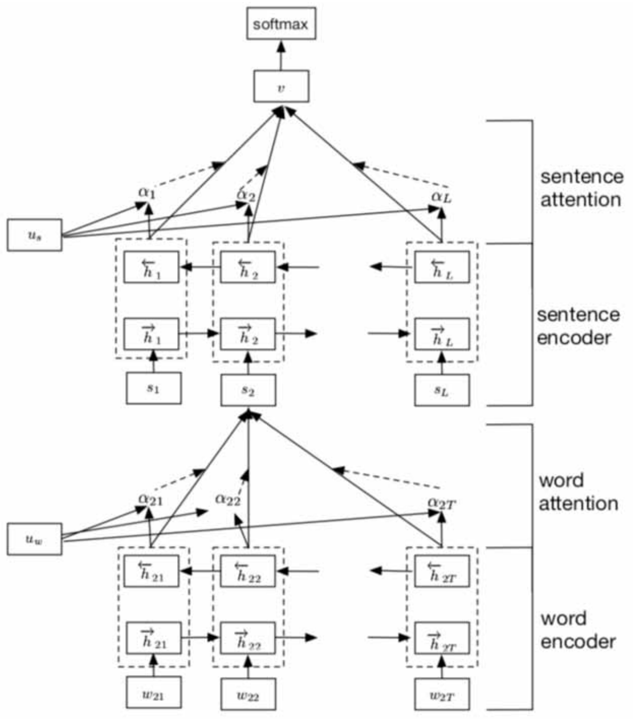


Table 6. Comparison with hierarchical attention network (with only text feature)

Model	Accuracy	Precision	Recall	F-measure
Jin et al., 2017 (only Text)	0.532	0.598	0.541	0.568
HAN (only Text)	0.723	0.576	**0.882**	0.697
Attention-RNN (only Text)	**0.738**	**0.774**	0.826	**0.799**

Table 7. Comparison with hierarchical attention network (with all features)

Model	Accuracy	Precision	Recall	F-measure
Jin et al., 2017 (All)	0.682	**0.78**	0.615	0.689
HAN (All)	0.68	0.55	0.647	0.594
Attention-RNN (Late fusion)	0.646	0.712	0.795	0.751
Attention-RNN (Early fusion)	**0.748**	0.768	**0.897**	**0.827**

4.6 Effect of Social Feature Selection

To further verify the effects of various social features on rumor detection, we evaluated the performance using different combinations of social features as in Table 8.

As shown in Table 8, for early fusion of different social features, the best performance can be observed when we combine all features except for user feature. The best performance with a F-measure of 0.882 can be obtained, which is better than using all features with a F-measure of 0.827. Also, we can observe that among single social features, tags play the most important roles. Similarly, we compared different social features for late fusion as in Table 9.

As shown in Table 9, we can again validate the superior performance for using all social features except for user feature in late fusion, which gives a F1 score of 0.856. This shows that user features such as the number of friends, followers, or tweets cannot be reliably used to distinguish between rumors and normal messages.

Table 8. Effects of social feature selection (for early fusion)

Features	Accuracy	Precision	Recall	F-measure
Early fusion (only Tag)	0.754	0.815	0.819	0.817
Early fusion (only Sentiment)	0.6	0.704	0.691	0.698
Early fusion (only User)	0.588	0.712	0.653	0.681
Early fusion (All features)	0.748	0.768	0.897	0.827
Early fusion (All except User)	**0.827**	**0.816**	**0.959**	**0.882**

Table 9. Effects of social feature selection (for late fusion)

Dataset	Accuracy	Precision	Recall	F-measure
Late fusion (only Tag)	0.74	0.777	0.861	0.817
Late fusion (only Sentiment)	0.63	0.726	0.72	0.723
Late fusion (only User)	0.571	0.741	0.55	0.635
Late fusion (All features)	0.646	0.712	0.795	0.751
Late fusion (All except User)	**0.801**	**0.834**	**0.879**	**0.856**

5. CONCLUSION

In this paper, we proposed to improve rumor detection by image captioning for extracting semantic meanings from images, and multi-cell Bi-RNNs with self-attention for learning the relative weights among words. Firstly, we adopted various types of features including texts, images, and social features for rumor detection. We applied a Seq2Seq model for image captioning to convert image features into texts, and combine with text and social features. Then, we proposed Multi-cell Bi-RNN, which stacks multiple single-layer RNNs in each direction before the outputs are combined. We also included self-attention mechanism to further modify the weights based on their relative importance. The best performance with a F-measure of 0.882 can be obtained when we used GRU-based Multi-cell Bi-RNNs with early fusion of all features except for user feature. It's superior to our baseline model. In future, we plan to further evaluate the proposed method with datasets in larger scales.

ACKNOWLEDGMENT

This paper is partly supported by the research grants from Ministry of Science and Technology, Taiwan, under the grant number MOST111-2221-E-027-121, and partly supported by the research grants from National Taipei University of Technology International Joint Research Project, under the grant number NTUT-IJRP-109-02. Jenq-Haur Wang and Mehdi Norouzi have equal contribution to this paper.

REFERENCES

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *Proceedings of ICLR 2015*.

Boididou, C., Papadopoulos, S., Dang-Nguyen, D.-T., Boato, G., Riegler, M., Middleton, S. E., Petlund, A., & Kompatsiaris, Y. (2016). Verifying multimedia use at MediaEval 2016. *Proceedings of MediaEval 2016 Workshop*.

Chen, T., Li, X., Yin, H., & Zhang, J. (2018). Call attention to rumors: Deep attention based recurrent neural networks for early rumor detection. In *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 40-52). doi:10.1007/978-3-030-04503-6_4

Jin, Z., Cao, J., Guo, H., Zhang, Y., & Luo, J. (2017). Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia* (pp. 795-816). doi:10.1145/3123266.3123454

Li, B., Qian, Z., Li, P., & Zhu, Q. (2022). Multi-modal Fusion Network for Rumor Detection with Texts and Images. In *Proceedings of MMM 2022* (pp. 15–27). doi:10.1007/978-3-030-98358-1_2

Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)* (pp.1412-1421). doi:10.18653/v1/D15-1166

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.-F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI 2016)* (pp. 3818-3824). Academic Press.

Ma, J., Gao, W., & Wong, K.-F. (2018). Detect rumor and stance jointly by neural multi-task learning. In *Proceedings of the International World Wide Web Conference 2018* (pp. 585-593). doi:10.1145/3184558.3188729

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

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL 2002)* (pp. 311-318). Academic Press.

Pearson, E. S., Gosset, W. S., Plackett, R., & Barnard, G. A. (1990). *Student: A statistical biography of William Sealy Gosset*. Oxford University Press.

Sampson, J., Morstatter, F., Wu, L., & Liu, H. (2016). Leveraging the implicit structure within social media for emergent rumor detection. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM 2016)* (pp. 2377-2382). doi:10.1145/2983323.2983697

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681. doi:10.1109/78.650093

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2015* (pp. 1-9). IEEE.

Tan, Z., Wang, M., Xie, J., Chen, Y., & Shi, X. (2018). Deep semantic role labeling with self-attention. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence 2018* (pp. 4929–4936). AAAI.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In Proceedings of neural information processing systems 2017 (pp. 5998-6008). Academic Press.

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition 2015* (pp. 3156-3164). doi:10.1109/CVPR.2015.7298935

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of international conference on machine learning 2015* (pp. 2048-2057). Academic Press.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1480-1489). Academic Press.

Ying, L., Yu, H., Wang, J., Ji, Y., & Qian, S. (2021). Multi-Level Multi-Modal Cross-Attention Network for Fake News Detection. *IEEE Access: Practical Innovations, Open Solutions*, *9*, 132363–132373. doi:10.1109/ACCESS.2021.3114093

Yu, F., Liu, Q., Wu, S., Wang, L., & Tan, T. (2017). A convolutional approach for misinformation identification. In *Proceedings of IJCAI 2017* (pp. 3901-3907). doi:10.24963/ijcai.2017/545

Zhao, W., Wang, B., Ye, J., Yang, M., Zhao, Z., Luo, R., & Qiao, Y. (2018). A multi-task learning approach for image captioning. In *Proceedings of IJCAI 2018* (pp. 1205-1211). doi:10.24963/ijcai.2018/168

Zhou, H., Ma, T., Rong, H., Qian, Y., Tian, Y., & Al-Nabhan, N. (2022). MDMN: Multi-task and Domain Adaptation based Multi-modal Network for early rumor detection. *Expert Systems with Applications*, *195*, 116517. doi:10.1016/j.eswa.2022.116517

ENDNOTES

[1] https://www.factcheck.org/
[2] https://www.snopes.com/
[3] http://swn.isti.cnr.it/

Jenq-Haur Wang received the Ph.D. degree from National Taiwan University, in 2002. He currently works as a Professor in the Department of Computer Science and Information Engineering, National Taipei University of Technology, Taiwan. His research interests include social media mining, Web information retrieval, machine learning, and big data analytics. He has published in major journals including Information Processing & Management, International Journal of Intelligent Systems, Multimedia Tools and Applications, and IEEE Access.