Accepted Manuscript

Attention-based Convolutional Approach for Misinformation Identification from Massive and Noisy Microblog Posts

Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan

PII: S0167-4048(18)30399-7
DOI: https://doi.org/10.1016/j.cose.2019.02.003
Reference: COSE 1467

To appear in: *Computers & Security*

Received date: 20 April 2018
Revised date: 10 November 2018
Accepted date: 4 February 2019

Please cite this article as: Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan, Attention-based Convolutional Approach for Misinformation Identification from Massive and Noisy Microblog Posts, *Computers & Security* (2019), doi: https://doi.org/10.1016/j.cose.2019.02.003

Attention-based Convolutional Approach for Misinformation Identification from Massive and Noisy Microblog Posts

Feng Yu, Qiang Liu, Shu Wu, Liang Wang, Tieniu Tan*

Center for Research on Intelligent Perception and Computing (CRIPAC)
National Laboratory of Pattern Recognition (NLPR)
Institute of Automation, Chinese Academy of Sciences (CASIA)
University of Chinese Academy of Sciences (UCAS), Beijing 100190, China
{feng.yu, qiang.liu, shu.wu, wangliang, tnt}@nlpr.ia.ac.cn

Abstract

The fast development of social media fuels massive spreading of misinformation, which harm information security at an increasingly severe degree. It is urgent to achieve misinformation identification and early detection in social media. However, two main difficulties hinder the identification of misinformation. First, an event about a piece of suspicious news usually comprises massive microblog posts (hereinafter referred to as post), and it is hard to directly model the event with *massive-volume* posts. Second, information in social media is of *high noise*, i.e., most posts about an event have little contribution to misinformation identification. To resolve the difficulty of massive volume, we propose an Event2vec module to learn distributed representations of events in social media. To overcome the difficulty of high noise, we mine significant posts via content and temporal co-attention, which learn importance weights for content and temporal information of events. In this paper, we propose an Attention-based Convolutional Approach for Misinformation Identification (ACAMI) model. The Event2vec module and the co-attention contribute to learning a good representation of an event. Then the Convolutional Neural Network (CNN) can flexibly extract key features scattered among an input sequence and shape high-level interactions

*Corresponding author

among significant features, which help effectively identify misinformation and achieve practical early detection. Experimental results on two typical datasets validate the effectiveness of the ACAMI model on misinformation identification and early detection tasks.

Keywords: information security; social network; misinformation identification; early detection; convolutional neural network; co-attention

1. Introduction

Nowadays, social media, such as Facebook and Twitter, enable increasingly easy access and extensive applications for users. On the one hand, users can enjoy convenient lives and easy access to information anytime anywhere with the help of social media. On the other hand, social media provides fertile breeding ground for misinformation dissemination. According to statistics of Facebook (the most popular social network worldwide), there are more than 2 billion monthly active users and 23% of users say to have shared misinformation either knowingly or not[1]. Social media will amplify harm of misinformation via wide propagation, which will likely harm information security, mislead public opinion, impact political election[2] and further pose huge threat to public security and social stability. Moreover, a feasible solution to preventing the spread of misinformation is to detect misinformation at an early stage and launch directed and effective counter campaigns [1]. Therefore, it is more and more urgent to identify misinformation from a mass of social media information and detect misinformation as early as possible.

The tasks in this work are misinformation identification and early detection, both of which identify an event in social media as misinformation or true information. Here an event is about a piece of news propagating in social media,

[1] http://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/

[2] http://www.npr.org/2016/11/08/500686320/did-social-media-ruin-election-2016

such as "Ballistic missile threat inbound to hawaii"[3]. Moreover, an event usually comprises many posts including postings, repostings and comments. To be specific, the task of misinformation identification is to detect whether an event is misinformation or not by analyzing a sequence of posts of the event, and the task of early detection is to identify misinformation or true information only using partial posts of the early stage of an event.

To identify misinformation, some conventional models have been proposed based on handcrafted features, which are extracted from user credibility and post content at a post level [2] [3] [4], at an event level [5] [6] [7] or aggregating from the post level to the event level [8]. Some other works adopt more effective handcrafted features, such as conflict viewpoints [9], temporal properties [5] [6], users' replies [10] [11] and signals tweets containing skepticism [7]. However, handcrafted features may not cover potentially informative features in dynamic and complicated social media scenarios. What's worse, a rough mergence of different handcrafted features cannot shape high-level interactions among significant features. Lastly, these feature engineering methods are also labor-intensive for so many designs.

However, events in social media contain massive-volume and high-noise posts, which need suitable remedy. The massive number of posts of an event is up to tens of thousands. What's worse, misinformation with massive posts means severe influence and damage. To resolve the difficulty of massive volume, we propose an Event2vec module to learn distributed representations of events in social media. Moreover, information in social media is of high noise, i.e., most posts about an event have little contribution to misinformation identification. So, some significant information for misinformation identification may be easily drowned in the high noise posts. To overcome the difficulty of high noise, we incorporate attention mechanism into the Event2vec module. Then we can mine significant posts to obtain better representations of events. Specifically, we propose content and temporal co-attention, which learn importance weights

[3]https://www.nytimes.com/2018/01/13/us/hawaii-missile.html

for content and temporal information of events.

The Event2vec module and the co-attention contribute to learning a good representation of an event. To mine key features from the event representation, deep neural network (DNN) is a good choice. A RNN-based Rumor Detector (RRD) [12] treats text content of posts in an event as a variable-length time series, which can capture the dynamic temporal characteristic during the diffusion process. But a popular event may comprise tens of thousands of posts, back propagation through a great number of time steps of RNN will be computationally ineffective and costly, so RRD only use partial posts from continuous intervals. Thus RRD cannot get stable performance of misinformation identification and practical early detection.

On the one hand, shortcomings of above-mentioned feature-engineering-based and RNN-based methods should be remedied, if we want to further reduce harm of widespread misinformation. On the other hand, some recent studies about CNN architecture have successfully modeled significant semantic features in varieties of fields, e.g., CNN based approaches to speech recognition [13], semantic analysis [14], click-through rate prediction [15], semantic segmentation [16] and reinforcement learning tasks [17]. Different from feature engineering, CNN can not only automatically extract local-global significant features from an input instance but also reveal those high-level interactions. Unlike unchangeably propagating sequential characteristics of RNN, the convolutional architecture and k-max pooling operation in CNN can flexibly extract key features scattered among an input sequence.

In this paper, we propose an ACAMI model for misinformation identification and early detection tasks. The CNN in ACAMI can automatically extract local-global significant features from an input instance and reveal those high-level interactions, so the ACAMI model can flexibly extract key features scattered among one input sequence. We obtain some observations from visualization experiments of what the ACAMI model has learnt, which help better understand human behaviors in social media and more exactly shape real-world social media scenarios for misinformation identification.

The main contributions of this work are as follows:

- We propose a new end-to-end trainable pipeline for misinformation identification, which consists of 1) an unsupervised Event2vec to learn distributed representations of events in social media and 2) convolution networks to automatically obtain key features from distributed representations of both misinformation and true information.
- We are the first to apply content attention and temporal attention to the task of misinformation identification and early detection, which contributes to learning key content and temporal information for each post.
- We demonstrate the robustness of the ACAMI model against massive volume and high noise in misinformation identification and visualize what the proposed model has learnt. Experiments conducted on two typical datasets show that the ACAMI model outperforms the state-of-the-art methods in both misinformation identification and early detection.

The rest of the paper is organized as follows. In Section 2, we review related work and methods of misinformation identification and early detection. Section 3 presents some analyses of the two adopted datasets. Section 4 details the proposed model. In Section 5, we conduct experiments on two typical datasets and compare with several state-of-the-art methods. Section 6 concludes the paper and discusses future work.

2. Related Work

In this section, we review some related works on misinformation identification and early detection. We also introduce related methods of attention mechanism, distributed representations and convolutional neural network.

2.1. Misinformation Identification and Early Detection

Recently, many methods have been put forward for automatic identification of misinformation. The work of [18] analyzes impact and characteristics of hoax articles in Wikipedia and proposes an efficient method to identify these Wikipedia hoaxes. The work of [19] traces misinformation in social media by their propagating characteristics. In social media, some researchers identify misinformation at the post level [2] [4], i.e., classifying a single post as being credible or not based on tweet-based features. Some perform a characterization analysis for the spread of fake images of posts during crisis events [3]. Some identify whether an event belongs to misinformation or true information and extract handcrafted features from the event level [5] [6] [7]. Another work obtains credibility of a post and then aggregates credibility to the event level [8]. Moreover, some other works extract more effective handcrafted features. For instance, the work of [9] [20] takes advantage of ”wisdom of crowds” to identify fake news, i.e., mining opposing voices from conflicting viewpoints. Based on the time series of misinformations lifecycle, the temporal characteristics of social context information are captured in [5] [6]. The work of [10] [11] investigate the web page credibility through users' feedback. Signals tweets are identified from trending misinformation via finding signature text phrases expressing skepticism about factual claims [7]. All the above feature-engineering-based methods fail to cover potentially informative features in dynamic and complicated social media scenarios and shape elaborate high-level interactions among significant features. To overcome these deficiencies, a RNN-based model attempts to capture the dynamic temporal signals in the misinformation diffusion process and incrementally learn both the temporal and textual representations of an event not relying on any handcrafted features [12].

2.2. Attention Mechanism

Attention mechanism is first applied to a visual attention system for scene analysis [21]. The visual attention system selects attended locations in order of decreasing saliency, so that a complex scene can be understood by rapidly selecting saliency locations in a computationally efficient method. In recent years, DNN is getting increasingly popular. Attention mechanism is once again taken out to be integrated into DNN. Hard attention is incorporated into RNN in [22], to attend to different locations within the images one at a time and process them sequentially. The attention mechanism can help control expensive computation independent of the input image size and learn to track items without explicit training signals.

In the field of computer vision, the work of [23] extends the attention-based RNN model to multiple objects detection task that learns to localize and recognize multiple objects despite being given only class labels. For an image caption task, an attention-based model is able to automatically fix its attention on salient objects of an input image while generating the corresponding words of the output sentence [24]. Some employ attention mechanism in a visual question answering task, such as generating question-guided attention to image feature maps for each question [25], a question-guided spatial attention to images for questions of spatial inference [26] and querying an image and inferring the answer multiple times to narrow down the attention to images progressively via stacked attention networks [27]. For fine-grained image classification, an attention-based CNN model improves the performance of which to attend and what to extract without expensive annotations like bounding box or part information [28].

In the field of natural language processing, researchers first introduce attention mechanism to neural machine translation. Based on a primitive encoder-decoder architecture, the work of [29] introduces a soft-global attention to search a source sentence to attend to the most relevant words to predict a target word. Some extend the global and local attention and compare different methods of obtaining attention scores [30]. Moreover, a hierarchical attention mechanism guides layers with a CNN to model text in [31]. The work of [32] proposes self-attention which memorizes key information from self-input without external-guide information. Besides, attention mechanism is introduced into more research issues, such as abstractive text summarization [33], text comprehension task [34] [35] [31], relation classification [36] [37] and text classification [32]. In [38], a novel model for speech recognition is proposed, which incorporates both content-based attention [29] [24] and location-based attention [39].

2.3. Distributed Representations

The idea of distributed representations is to digitize concepts, which is first proposed in [40]. And then we can model digitized concepts with the help of many math and engineering tools, such as stochastic gradient descent [41] and back-propagating [42]. For instance, the work of [42] can learn distributed representations for words via back-propagating. Later on, many works focus on a good language model to learn word embedding, such as [43] [44] [45] [46] [47].

Distributed representations for concepts at a higher semantic level, such as phase, sentence and paragraph, have received much attention [48] [49] [50]. Semi-supervised and supervised methods are introduced in [51] [52]. Moreover, the work of [53] computes the paragraph embedding through gradient descent, which is unsupervised to obtain more general representations. How to learn distributed representations for concepts at an even higher semantic level, such as an event? In this case, we introduce the attention module to selectively attend to important paragraph text and obtain event representations in a supervised way.

2.4. Convolutional Neural Network

Inspired by biological organization of visual cortex, CNN has been developed for visual object recognition [54]. Hierarchically and increasingly complex features can be constructed by alternating applications of convolutional and pooling layers of CNN. The architectures help model significant semantic features and achieve much improvement in various fields. In speech recognition, CNN has been developed to extract temporal features [13]. Similarly, semantic features from vision information can be guided to image classification and segmentation tasks [16]. Moreover, a general 2D CNN can be extended to a 3D one for 3D image restoration problems [55] and video-based human action recognition [56]. In sentiment prediction and document classification, CNN can be trained to obtain semantic features at the top layer [14] from raw text. CNN can also be employed to other issues, such as click-through rate prediction [15] and reinforcement learning tasks [17]. CNN is usually trained through stochastic gradient descent (SGD), with backpropagation to compute gradients.

This paper is built on our preliminary conference version [57] and the main extensions are detailed as follows.

1) While the previous method in [57] focus on distribution patterns at the dataset scale, we now specifically mine content and temporal importance at the post scale, i.e., mining importance of each post.
2) We are the first to apply content and temporal co-attention to learn representations for events with massive posts in social media via the newly added Event2vec module.
3) More comprehensive experiments, e.g., analyses of attention module, are designed to demonstrate that the attention module is effective, robust and interpretable to resolve the massive volume and high noise difficulties of misinformation identification.
4) A newly added review of methods of misinformation identification, which thoroughly summarize works about attention mechanism in various fields and distributed representation in different semantic levels.

3. Dataset Analysis

Table 1: Statistics of the datasets

Statistic	Twitter	Weibo
# of Users	491,229	2,746,818
# of Posts	1,101,985	3,805,656
# of Events	992	4,664
Avg. # of words / post	10.62	29.04
Avg. # of posts / event	1,111	816
Max # of posts / event	62,827	59,318
Avg. time span / event	1,582.6 Hours	2,460.7 Hours

3.1. Statistics of the Datasets

To empirically evaluate the performance of our methods on misinformation identification, we perform experiments on two typical microblog datasets: Weibo and Twitter datasets[4], which are developed and used by [2] [5] [12]. Ground truth of each event are confirmed from online rumor debunking service, such as Snopes website[5] and Sina community management center. For each event, Twitter API can return search results based on keywords of the Snopes website; the Weibo API can return the original posts, corresponding repost and reply messages about an event.

Details of the two datasets are illustrated in Table 1. An event in social media usually comprises thousands of posts. It should be noted that some events of misinformation contain tens of thousands of posts, whose massive volume means severe influence and damage. For instance, a piece of misinformation about terrorism[6] contains 12,217 posts, which will pose threat to public security and social stability. For practical misinformation identification, models should be still robust even for misinformation with massive posts.

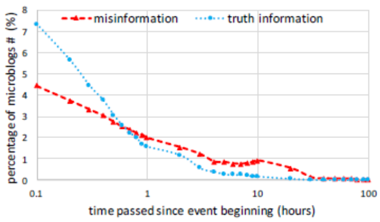


Figure 1: The **long-tailed** distribution of both misinformation and true information in the Weibo dataset in a semi logarithmic coordinate.

[4] http://alt.qcri.org/~wgao/data/rumdect.zip

[5] www.snopes.com

[6] http://www.ibtimes.co.uk/anonymous-hackers-threaten-reveal-identities-1000-ku-klux-klan-members-opkkk-1525758

3.2. Distribution Pattern of Misinformation and True Information

We investigate the data distribution of misinformation and true information in these two datasets, which reveals two patterns of the data distribution.

Take the Weibo dataset as an example, the data distribution is illustrated in Figure 1. Each point represents the percentage of posts during a time window of 0.1 hours at the corresponding time point. The *long-tailed* distribution of both misinformation and true information can be clearly shown even in the semi logarithmic coordinate (otherwise the curves almost coincide with two coordinate axes).

Moreover, we can see that temporal properties usually differ between misinformation and true information. Compared to misinformation, most posts of true information are posted or reposted at the beginning of broadcast and vanish very fast. However, misinformation usually has a relatively larger quantity at the middle phase of an event. This observation inspires us to propose the following Event2vec module, where temporal attention is incorporated to model different temporal properties.

4. Proposed Models

In this section, we propose the ACAMI model. We first introduce the general framework. Then we detail an Event2vec module which can learn distributed representations for events in social media. To investigate how to generate good representations for events with tens of thousands of posts, we incorporate content and temporal co-attention into the Event2vec module.

4.1. General Framework

As illustrated in Figure 2, we will introduce the general framework of the proposed ACAMI model. From the bottom up, there are two submodules as follows.

Using Event2vec to learn distributed representations of events. Similar to Word2vec [48] and Para2vec [53], given a set of events in social

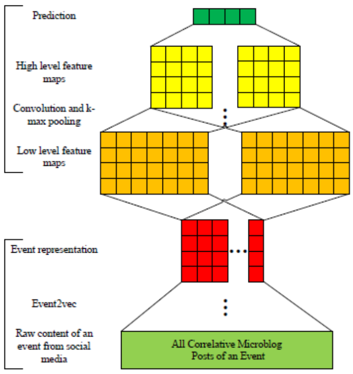


Figure 2: The general framework of the ACAMI model. From the bottom up: learn event representation; extract features from low level to high level with CNN. Event representation learnt by the Event2vec module will not be updated in following training process.

media, we attempt to learn high-quality distributed representations of events. Each event comprises many posts and each post is a paragraph of text with a timestamp. The Event2vec module inputs an event of massive posts and outputs its distributed representation. The formulations of the Event2vec module will be detailed in the next Subsection. Moreover, event representation learnt by the Event2vec module will not be updated in following training process.

Modeling high-level interactions by CNN. A commonly used architecture of CNN comprises convolutional layers, k-max pooling layers and a fully connected layer.

For an input event instance e_i with n phases, each phase is embedded as $\mathbf{g}_i \in \mathbb{R}^d$ and we can get the instance matrix $\mathbf{G} \in \mathbb{R}^{d\times n}$. In the convolutional network, a convolutional layer is obtained by convolution operations of a weight matrix $\mathbf{C} \in \mathbb{R}^{d\times \omega}$ on the activation matrix at the layer below in a row-wise way. Followed by a nonlinearity function applied to the convolution result, an element of a feature map can be obtained as:

$$\mathbf{f}[i] = \tanh\left(\langle \mathbf{G}[:,\ i : i+\omega-1], \mathbf{C}\rangle_F\right)\ , \tag{1}$$

where $\mathbf{G}[:,\ i : i+\omega-1]$ is the i to $(i+\omega-1)$-th columns of $\mathbf{G}$ and the subscript F is the Frobenius inner product, i.e., the summation of products of corresponding elements of both matrices. At last, we take k-max pooling over the feature map $\mathbf{f}$ to capture the most significant features $\mathbf{f}^k_{max}$, i.e., k largest values of the feature map in response to the specific kernel $\mathbf{f}$ and the order of the values in $\mathbf{f}^k_{max}$ stays the same as their original order in $\mathbf{f}$.

Moreover, the above convolutional and pooling operations can be repeated to yield deeper layers. Finally, there is a fully connected layer and the ultimate output p_{e_i} is obtained via softmax. Here, p_{e_i} is the probability which predicts whether the event e_i belongs to misinformation.

4.2. Event2vec

Note that the Event2vec module in the proposed ACAMI model is different from that in the previous version [57]. The co-attention of the Event2vec module mines content and temporal importance at the post scale, which is more helpful for misinformation identification. The Event2vec module can be formulated as the following two steps.

Splitting all correlative posts of an event into several groups of equal number. We intend to group all correlative posts of an event into a sequence of time windows and extract features through modeling these groups. Why split into several groups? First, an event generally consists of thousands of correlative posts on average and there is huge difference in quantity of events. Moreover, posts during some specific time windows are so relevant that we can treat these neighbor posts as a group which represents a specific event phase. Note that window size of Word2vec is 10, which models semantics of 20 context words[7]. Inspired by the implement suggestion, we also split posts of an event into 20 groups and learn the event representation by modeling representations these context groups, which achieves the best experimental result.

Usually the basic grouping criterion is equally splitting by time or quantity, which means groups are with equal time spans or equal number of posts. Con-

[7]https://code.google.com/archive/p/word2vec/

sidering the long-tailed distribution of social media information illustrated in Figure 1, the first few groups will contain the vast majority of posts if splitting by time, which makes it difficult to learn representations of events well. Moreover, splitting by quantity can be local or global, which means each event is split separately or by globally shared cut-points. The globally equal-quantity grouping method first normalizes timestamps of posts of each event, then obtains globally shared cut-points by equally splitting normalized timestamps of all events into multiple parts [57]. We will adopt both locally and globally equal-quantity grouping methods in the Event2vec module.

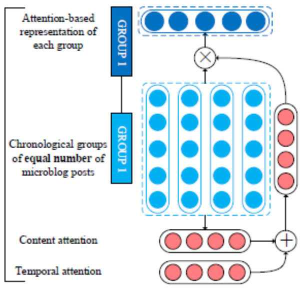


Figure 3: The Event2vec module in ACAMI. From the bottom up: split raw content into chronological groups of equal number of posts; learn a representation of each group via attention mechanism (best viewed in color). Content and temporal co-attention are learnt from content text and timestamp separately.

Learning representation for each group via content and temporal co-attention. The attention module can acquire the importance weights of content and temporal information of each post in a group. This step can be depicted in Figure 3.

First, the paragraph vector [53] is employed to learn representation of each post. Given a post of N words, a word is represented by a column vector $\mathbf{w}_n$ in $\mathbf{W}$ and the post is represented by a column vector $\mathbf{p}_j$ in $\mathbf{D}$. To learn the post representation $\mathbf{p}_j$, we compute

$$\underset{\mathbf{D},\mathbf{W}}{\arg\max} \frac{1}{N} \sum_{n=k}^{N-k} \log p\left(\mathbf{w}_n | \mathbf{w}_{n-k}, \cdots \mathbf{w}_{n+k}\right) \ . \tag{2}$$

The n-th word is predicted via softmax,

$$p\left(\mathbf{w}_n|\mathbf{w}_{n-k},\cdots\mathbf{w}_{n+k}\right)=\frac{\exp(\boldsymbol{\theta}^T\mathbf{x}_n)}{\Sigma_i\exp(\boldsymbol{\theta}^T\mathbf{x}_i)}\ , \tag{3}$$

$$\mathbf{x}_n=h\left(\mathbf{p}_j,\mathbf{w}_{n-k},\cdots,\mathbf{w}_{n+k};\mathbf{D},\mathbf{W}\right)\ , \tag{4}$$

$\boldsymbol{\theta}$ is the softmax parameter and h is a concatenation or average operation. Context words and paragraph memory are leveraged to predict the current word.

We observe that an event may contain tens of thousands of posts, so some significant information for misinformation identification may be easily drowned in the high-noise posts. What's worse, there are many duplicate reposting contents in an event. If we only use Para2vec to capture semantic information of groups of posts, the event representations will mostly focus on those duplicate content. Moreover, early detection of misinformation means using fewer posts of the early stage of an event. How can we still mine key features from fewer posts with lots of noise? Attention mechanism may be a good solution. We propose content attention and temporal attention, which learn importance weights for both content and temporal information of events. I.e., the attention module selectively attends to important content and temporal characteristic of an event.

Based on above observation, content and temporal co-attention will be leveraged to learn representation of each group of posts. Given a group of c posts, we can learn a representation $\mathbf{p}_j\in\mathbb{R}^{d_1}$ for each post and concatenate them to obtain a matrix $\mathbf{M}\in\mathbb{R}^{d_1\times c}$, where d_1 is the dimensionality of the paragraph vector of a post. The attention mechanism will produce a vector $\mathbf{a}$ of attention weights and a weighted representation $\mathbf{g}$ of a group via,

$$\mathbf{B}=\tanh(\mathbf{E}\mathbf{M})\ , \tag{5}$$

$$\mathbf{a}_c=\mathbf{B}^T\mathbf{u}\ , \tag{6}$$

$$\mathbf{a}_t=\mathbf{Y}\mathbf{x}\ , \tag{7}$$

$$\mathbf{a} = softmax(\mathbf{a}_c + \mathbf{a}_t)\ , \tag{8}$$

$$\mathbf{g} = \mathbf{M}\mathbf{a}\ , \tag{9}$$

then we can acquire the input matrix $\mathbf{G}$ of CNN by concatenating those $\mathbf{g}$. Attention weights $\mathbf{a}_c \in \mathbb{R}^c$, $\mathbf{a}_t \in \mathbb{R}^c$ are for content and timestamp of c posts in the group. And $\mathbf{E} \in \mathbb{R}^{d_2 \times d_1}$ is the parameter of a one-layer MLP to get a hidden representation $\mathbf{B}$ of $\mathbf{M}$. Attention parameter $\mathbf{u} \in \mathbb{R}^{d_2}$ can be regarded as high-level semantic representation of "salient information in misinformation", as a similar usage in memory networks [58] [59]. We need to point out that d_2 is a hyper-parameter and the study about tuning d_2 will be presented in Section 5.5. Moreover, $\mathbf{x} \in \mathbb{R}^{n_t}$ is a vector of temporal attention weights of n_t different time intervals,

$$\mathbf{x} = [x_0, x_1, \cdots, x_{n_t-1}]^T, \tag{10}$$

and x_i is for the i-th time interval. The timestamp t of each post can be allocated to a time interval as follows,

$$(interval)_i = \begin{cases} t = 0\ ,\ i = 0; \\ t > (n_t - 2)t_u\ ,\ i = n_t - 1; \\ \vdots \\ (i-1)t_u < t \leq i \cdot t_u\ ,\ else. \end{cases}\ , \tag{11}$$

where $t_u = 3600$ seconds in this work. In addition, each row in $\mathbf{Y}$ is an one-hot vector and $\mathbf{Y}_{ij} = 1$ if the timestamp of the i-th post in the group falls into the j-th time interval.

5. Experiments

In this section, we first present several compared methods and experimental settings used in our proposed method. Then we report experimental results of misinformation identification and early detection on two typical datasets. Moreover, we research into the robustness of the proposed ACAMI against massive volume and high noise in misinformation identification. We then discuss the influence of the number of posts to the performance of misinformation identification. We also conduct some visualization experiments which help apparently illustrate what the proposed model has learnt against high noise.

5.1. Experimental Settings

Several methods are used for empirical comparison with ours:

(1) **RRD** proposes a longest continuous intervals algorithm to construct input instances of a GRU model. The enhanced GRU hidden layer conduce to obtain high-level interactions of features [12].

(2) **SVM-TS** is a linear SVM classifier that uses Time-Series structures to model the variation of social context features and these handcrafted features are extracted based on contents, users and propagation patterns [6].

(3) **DT-Rank** is a Decision-Tree-based Ranking model to identify trending rumors through ranking the clustered disputed factual claims based on statistical features [7]. **DTC** is a Decision Tree Classifier modeling information credibility [2].

(4) **SVM-RBF** is a SVM-based model with the RBF kernel [60].

(5) **RFC** is a Random Forest Classifier with three parameters to fit the temporal tweets volume curve [5].

(6) **CAMI** is our preliminary conference work [57], using CNN to model distribution pattern of misinformation.

In all experiments, we randomly choose 10% of the dataset for model tuning and the rest 90% are randomly assigned to a 3:1 ratio for training and test. Similar to [12], we report the *Accuracy*, *Precision*, *Recall* and *F1-score* of these methods to measure the performance of misinformation identification.

For the proposed ACAMI, we apply a CNN architecture with two layers in this work, which is implemented with Theano[8]. The parameters of ACAMI are set as $n_t = 32$, $t_u = 3600$, the dimensionality of the paragraph vector $d_1 = 50$,

[8] http://deeplearning.net/software/theano/

attention dimensionality $d_2 = 20$, the numbers of feature maps m and filter width w of two layers of CNN are set as $m = [6, 4], w = [8, 5]$ for the Weibo dataset, $m = [3, 2], w = [8, 5]$ for the Twitter dataset.

Table 2: Results of misinformation identification on both Weibo and Twitter datasets. (Class **M**: *Misinformation*; Class **T**: *True Information*)

Method	Class	Weibo				Twitter			
		Accuracy	Precision	Recall	F_1	Accuracy	Precision	Recall	F_1
DT-Rank [7]	M	0.732	0.738	0.715	0.726	0.681	0.711	0.698	0.704
	T		0.726	0.749	0.737		0.647	0.662	0.655
SVM-RBF [60]	M	0.818	0.822	0.812	0.817	0.715	0.698	0.809	0.749
	T		0.815	0.824	0.819		0.741	0.610	0.669
DTC [2]	M	0.831	0.847	0.815	0.831	0.718	0.721	0.711	0.716
	T		0.815	0.847	0.830		0.715	0.725	0.720
RFC [5]	M	0.849	0.786	0.959	0.864	0.728	0.742	0.737	0.740
	T		0.947	0.739	0.830		0.713	0.718	0.716
SVM-TS [6]	M	0.857	0.839	0.885	0.861	0.745	0.707	**0.864**	0.778
	T		0.878	0.830	0.857		0.809	0.618	0.701
RRD [12]	M	0.910	0.876	**0.956**	0.914	0.757	0.732	0.815	0.771
	T		0.952	0.864	0.906		0.788	0.698	0.771
CAMI [57]	M	0.933	0.921	0.945	0.933	0.777	0.744	0.848	0.793
	T		0.945	0.921	0.932		0.820	0.705	0.758
ACAMI	M	**0.948**	**0.940**	0.952	**0.946**	**0.803**	**0.781**	0.806	**0.794**
	T		**0.956**	**0.944**	**0.950**		**0.824**	**0.800**	**0.812**

5.2. Results of Misinformation Identification

The results of all methods are illustrated in Table 2. We can see that the performance ranking of misinformation identification methods is as follows, ACAMI, CAMI, RRD, SVM-TS, RFC, DTC, SVM-RBF and DT-Rank. Compared with DNN-based methods, the performance of other methods is relatively poor. These methods using handcrafted features or rules may not adapt to shape dynamic and complicated scenarios in social media. In contrast, DNN-based methods, ACAMI, CAMI and RRD, can learn high-level interactions among deep latent features, which contribute to model real-world scenarios.

Comparing those conventional methods, DT-Rank uses a set of regular expressions selected from signal posts containing skeptical enquiries. But not all posts in both Twitter and Weibo datasets involve these skeptical enquiries. These selected expressions are insufficient to conclude the information credibility. Moreover, SVM-TS and RFC incorporate the temporal structure into conventional models, which helps outperform other compared methods like SVM-RBF and DTC. So, we can see that modeling these temporal features is workable and effective.

For these DNN-based methods, the CAMI model obtains significant improvement over RRD. Despite the fact that both models learn deep latent features from a sequence of groups of posts, a trained GRU model possesses a constant recurrent transition matrix, which induces unchangeable propagations of sequence signals between every two consecutive time windows. However, in real-world scenarios, social media is so dynamic and complicated that the above constant recurrent transition matrix of the RRD model has its limitation to shape an adequate misinformation identification model. Furthermore, the above RRD model cannot get stable performance of misinformation identification due to the incomplete usage of input information. While key features of both misinformation and true information can appear at any part of an input sequence and may be dropped by RRD. The convolutional architecture and k-max pooling operation in the CAMI model, in contrast, can flexibly extract key features scattered among an input sequence. We will demonstrate it by the following visualization experiment.

In regard to the CAMI and ACAMI models, the ACAMI model surpasses the CAMI model in terms of all the evaluation metrics on both datasets. There is big difference between time distributions of misinformation and true information, so the CAMI model extracts more accurate and effective features based on the time distribution to gain better performance than previous methods. However, the attention mechanism of the ACAMI model can weigh the importance of every post at a finer scale than the CAMI model. Moreover, the ACAMI model can directly consider content and timestamp of a post. While the CAMI model will ignore some groups that are relatively unimportant on average, even if these groups contain a bit of important posts. Again, we will demonstrate this by the following visualization experiment.

Moreover, the accuracy on Twitter is significantly lower than that on Weibo. There may be two reasons. Firstly, a post in Twitter usually contains less words than a post in Weibo, comparing the average number of words per post in Table 1. Posts in Twitter is comparatively less informative than posts in Weibo, so it more difficult identify misinformation in Twitter with fewer words. Secondly, an event in Twitter usually contains more posts than an event in Weibo, comparing the average or maximum number of posts per event in Table 1. It is more difficult to attend to a few key posts from greater volumes of posts, which will degrade the performance of misinformation identification in Twitter.

5.3. Effects of Temporal and Content Attention

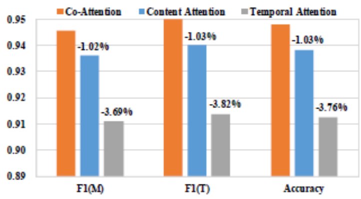


Figure 4: The ablation study of the proposed ACAMI with only temporal attention, only content attention and temporal-content co-attention using different metrics (Best viewed in color). The numbers on the columns indicate relative decrease of individual attention against co-attention. M: Misinformation; Class T: True Information.

Though extensive experiments are conducted to demonstrate effectiveness of the proposed ACAMI method, it is also interesting to compare the effects of temporal and content attention for the contributions to misinformation identification, respectively. Therefore, we do ablation study of the proposed ACAMI with only temporal attention, only content attention and temporal-content co-attention, whose results in the Weibo dataset is reported in Figure 4 and similar results are also achieved in the Twitter dataset. Comparing with performance of temporal-content co-attention, the performance of only content attention and only temporal attention decrease 1.02%, 1.03%, 1.03% and 3.69%, 3.82%, 3.76% in F1(M), F1(T), Accuracy, respectively.

From the results in Figure 4, we can draw the following two conclusions. Firstly, both temporal and content attention, that is, the timestamps and content information of posts are very significant for misinformation identification. Secondly, the content attention makes a relatively greater contribution to identifying misinformation than the temporal attention.

5.4. Grouping Methods

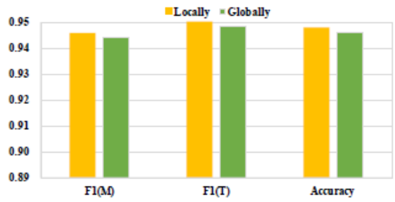
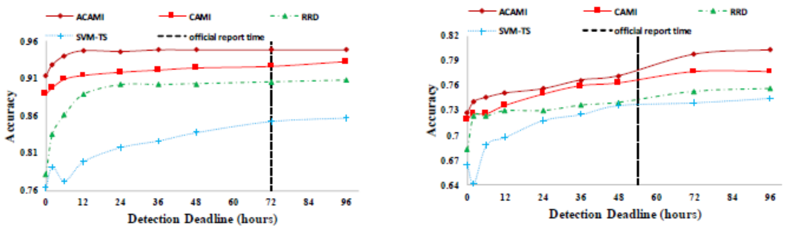


Figure 5: The performance of the proposed ACAMI with locally and globally equal-quantity grouping methods using different metrics (Best viewed in color). M: Misinformation; Class T: True Information.

As described in Section 4.2, we adopt both locally and globally equal-quantity grouping methods in the Event2vec module, we want to investigate how two grouping method will influence the performance of the proposed ACAMI, whose results in the Weibo dataset is reported in Figure 5 and similar results are also achieved in the Twitter dataset. We can see that the proposed ACAMI with either locally or globally equal-quantity grouping method achieves almost the same performance in all metrics.

The globally equal-quantity grouping method can consider the global temporal distribution of information in social media, which has proven to be effective in the previous work [57]. So, the attention mechanism in the Event2vec module may help reduce the gap between two grouping methods. It should be noted that globally equal-quantity grouping method is much more complicated that the locally equal-quantity one. Therefore our proposed attention-based Event2vec module can simplify the grouping method of the previous work [57]. For the sake of simplicity, we can just adopt the locally equal-quantity grouping method, that is, divide all correlative posts of an event into equivalent amount.



(a) Weibo dataset (b) Twitter dataset

Figure 6: Early detection of misinformation of four most competitive methods on both Weibo and Twitter datasets. The official report time is the average reporting time over misinformation and announced by the debunking services like Snopes and Sina community management center.

5.5. Early Detection of Misinformation

In order to evaluate performance of early detection of compared methods, we set a series of detection deadlines and only use posts from the initial broadcast to corresponding deadlines during the test process.

Four most competitive methods are for comparison, ACAMI, CAMI, RRD and SVM-TS. Moreover, conventional early detection tasks count on official announcements, which is the average reporting time over misinformation and announced by the debunking services like Snopes and Sina community management center. So, we take official report time as a reference.

Performance of the CAMI and ACAMI models versus the above methods with various deadlines are illustrated in Figure 6. The CAMI and ACAMI models can reach relatively high accuracy at a very early time while other methods will take a longer time to achieve good performance. Furthermore, accuracy of the CAMI and ACAMI models take a strong lead at any phase. Only in this way can the CAMI and ACAMI models shot misinformation at first appearance and achieve more practical early detection.

The accuracy of most methods will experience a conspicuous climbing during the first few hours and then rise with different growth rates, convergence rates and convergence accuracies. For instance, accuracy curve of SVM-TS climbs slowly at early phase and gradually converge to a relatively low accuracy. Moreover, its accuracy curve still fluctuates after the official report time. While the accuracy curve of RRD climbs rapidly at early phase and converges to a much higher accuracy on a much earlier deadline than that of SVM-TS.

Most state-of-the-art methods for early detection, such as RRD and SVM-TS, usually follow the intuitive paradigm to model time series features in sequences of posts. But these time-series-based models are not qualified for practical early detection due to the *conflict* between the models and the task. Take RRD as an example. On the one hand, the input sequence should be long enough to embody these possibly existing dynamic temporal signals to be captured by RRD [12]. On the other hand, the practical early detection means limited input sequence can be used. The limited input sequence may not cover required dynamic temporal signals. So RRD may not be suitable for early detection of misinformation in some cases. Nonetheless, convolutional and max pooling operations of the CAMI model can flexibly extract key features even from a limited input sequence, which make the CAMI model more effectively applied to early detection of misinformation. Moreover, the ACAMI model can attend to every post within each group, at a finer scale than the CAMI model, which helps further improve the performance of early detection.

Besides, the proposed ACAMI model can achieve a slightly better performance than the CAMI model. An event may contain tens of thousands of posts and many posts share duplicate reposting content. Moreover, early detection of misinformation means using fewer posts of the early stage of an event. Attention mechanism in the ACAMI model can help still mine key features from fewer posts with lots of noise. The content attention and temporal attention learn importance weights for both content and temporal information of events which selectively attend to important content and temporal characteristic of an event.

Table 3: The detailed mapping between *Group* and *Post#* (We divide events of both Weibo and Twitter datasets into 5 parts (i.e., Group1-5) by the number of posts.)

Post#	Group1	Group2	Group3	Group4	Group5
Weibo	<100	100-200	200-400	400-1000	>=1000
Twitter	<20	20-50	50-100	100-500	>=500

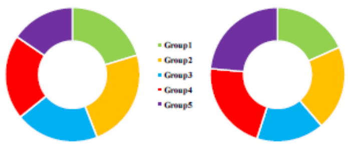
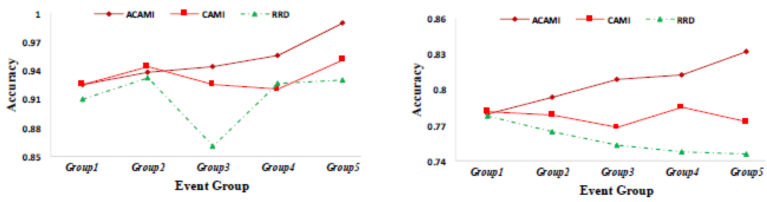


Figure 7: The proportion (best viewed in color) of Group1-5 on two datasets: Weibo dataset (left) and Twitter dataset (right).



(a) Weibo dataset

(b) Twitter dataset

Figure 8: The performance of three most competitive models on five different event group of both Weibo and Twitter datasets.

5.6. Robustness against Massive Volume

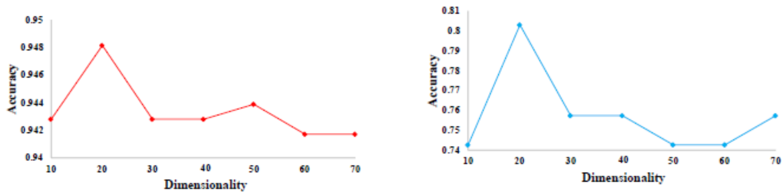
Similar to Tweet Index[9], Microblog Event Index (MEI) here is referred to as the number of microblog posts of an event. In this subsection, we want to discuss the influence of MEI to the performance of misinformation identification. Because we should check whether models are still robust to misinformation with massive posts, which usually means severe influence. We split the events in the test set into five groups based on MEI and compare the performance on each group among three most competitive methods, ACAMI, CAMI and RRD. To be specific, we first present why and how these five event groups are divided. Then we will detail the analyses based on the performance of the three models.

Intuitively, it is relatively difficult to identify whether an event is misinformation or not if the event contains massive posts. In an extreme circumstance, if an event comprises tens of thousands of posts, some significant information may be easily drowned in the information flood. Moreover, we learn representations based on Para2vec, which is unsupervised and learns from context. So, it is challenging to learn a good representation for an event with massive posts. Therefore, we divide the events into five groups (i.e., Group1-5) based on MEI. The grouping criteria is shown in Table 3. For instance, an event whose MEI falls within the scope of 200 and 400 belongs to Group3 in the Weibo dataset. On account of different distributions of Weibo and Twitter datasets, the scope of MEI may be different. For simplicity, groups are roughly equidistributed. In this way, metrics (such as *Accuracy*) computed based on the same group size are comparable. The proportions of Group1-5 are depicted in Figure 7.

The performance on these five different event groups is shown in Figure 8. Here we only compare the three competitive methods, ACAMI, CAMI and RRD. From Figure 8, we can see that the performance of the three models on Group1 is closer than other groups. However, as the MEI increases, the performance curve of the CAMI and RRD models fluctuates a lot. While the

[9]https://blog.twitter.com/engineering/en_us/a/2014/building-a-complete-tweet-index.html

performance of the ACAMI model is more robust as MEI increases. Moreover, the ACAMI model acquires better performance than CAMI and RRD on the Group5 of the highest MEI. If we want to develop a practical system for misinformation identification, we should check models' robustness to misinformation with massive posts. Because massive volume may mean severe influence and some models may fail. Compared with the CAMI model, the attention module in the ACAMI model plays a key role in extracting significant information from so many posts of an event.



(a) Weibo dataset (b) Twitter dataset

Figure 9: The performance of the proposed ACAMI model with different attention dimensionality d_2.

5.7. Attention Dimensionality

The parameters in the attention module are as follows, $\mathbf{E} \in \mathbb{R}^{d_2 \times d_1}$, $\mathbf{u} \in \mathbb{R}^{d_2}$. It seems that we can fine tune the hyper-parameters d_1 and d_2. But d_1 is also the dimensionality of the paragraph vector of a post in the Para2vec module. In order to best capture the distributed semantic representation of a paragraph of text, we first need to fine tune the dimensionality d_1 of the paragraph vector, as suggested in [61]. So when we improve the following attention module, we only fine tune the hyper-parameter d_2 and keep the hyper-parameter d_1 unchanged.

Here, we refer to the hyper-parameter d_2 as the attention dimensionality. We report performance of the proposed ACAMI model with different attention dimensionality d_2. From Figure 9, we can see that the performance truly fluctuates a lot with the attention dimensionality d_2. Moreover, the ACAMI model can achieve the best performance when the attention dimensionality d_2 is set around 20 for both Weibo and Twitter datasets.

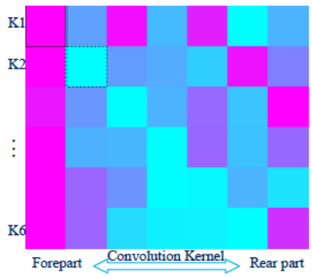


Figure 10: Visualization of convolutional kernels from the first convolutional layer (better viewed in color and rows). Each row represents a convolution kernel of size 7 and there are kernels (termed K1, K2, $\cdots$, K6) from 6 feature maps. Colors varying from bright blue (dashed line box) to bright red (black box) map values from low to high, representing the response intensity of kernels with respect to the input.

5.8. *Visualizing the CAMI and ACAMI Models*

The visualization experiments of the CAMI and ACAMI models attempt to demonstrate the following things. *First*, we can observe that key features scatter among an input sequence but not focus on a fixed part of sequences. *Second*, the CAMI model can flexibly extract these scattered key features. *Third*, the attention mechanism can further improve the robustness of the ACAMI model against high noise.

Visualizing convolutional kernels. We obtain all convolutional kernels from the first convolutional layer of a learnt CAMI model. With regard to a kernel matrix $\mathbf{W} \in \mathbb{R}^{d \times \omega}$ corresponding to a specific feature map, we sum all the rows into a row vector $\mathbf{v}_i \in \mathbb{R}^{\omega}$. Suppose there are m feature maps, we can stack these row vectors, $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_m$, into a visualization matrix $\mathbf{V} \in \mathbb{R}^{m \times \omega}$ and then plot it in a checkerboard which is illustrated in Figure 10. Taking the adopted one-dimension convolution into consideration, each row in the visualization figure illustrates general response of a corresponding kernel with respect to the input sequence.

From Figure 10, we can see that the forepart of the input usually obtains relatively stronger response than the rear part. After all, main description of misinformation and most relative replies may locate at the forepart. Only using partial posts from continuous intervals, the RRD model may not make the best of key features. These observations show that the CAMI model can flexibly extract key features scattered among an input sequence.

Visualizing saliency maps. Inspired by visualizing work in computer vision [62] [63], we plan to visualize key features grabbed by the CAMI model. In a feedback pass during test process, we compute the gradient of a class label value with respect to the input embedding matrix. More concretely, for a test instance, we perform a feedforward pass to obtain the output value and corresponding class label. Then we treat the class label value as loss and implement back propagation algorithm to acquire the gradient matrix of the class label value with respect to the input embedding matrix. Finally, we can get the most salient part of the input instance from the gradient matrix.

The top part of Table 4 demonstrates extracted salient posts of an identified misinformation about "Donald Trump Said Republicans Are the Dumbest Group of Voters", in which many questioning and denial signals can be observed in corresponding groups of posts. Such groups with indicating signals could be flexibly grabbed by the CAMI.

Visualizing the attention module of the ACAMI model. Similar to the above visualization of saliency maps, we implement back propagation algorithm to acquire the gradient matrix for the ACAMI model. Statistics in Section III reveal that some misinformation contains up to tens of thousands of posts. But most users simply accept and repost the misinformation, i.e. most posts about an event are high noise to misinformation identification. So we visualize the CAMI model and the ACAMI model to illustrate what the proposed models has learnt against high noise.

For comparison's purposes, we do the same as in the CAMI model and show extracted salient posts of the same identified misinformation about "Donald Trump Said Republicans Are the Dumbest Group of Voters". Apart from posts

Table 4: Extracted salient posts. The table is divided into two parts: the top part represents salient posts extracted by the CAMI model; the bottom part represents *extra* salient posts extracted by the ACAMI model.

time window #1 of CAMI	what???? IS IT TRUE? probably faked I doubt the Trump2016 folks do
time window #2 of CAMI	untrue... False, darn it. Didn't think so... it pays to fact check
time window #6 of CAMI	this is false Fake. False. Deceitful. but no proof exists that he said this... Just another graphic created by a pundit
Extra posts by ACAMI	it is just another scam FYI, Alert !!!!! #Dipshidiot! Nasty, is it true

in the top part of Table 4, the ACAMI model still acquires extra significant information in the bottom part of Table 4. Why the CAMI model misses some key information? Because the CAMI model is only at the group scale not the post scale and only extracts key features of relatively important groups on average. And there are some groups which are relatively unimportant as a whole but indeed contain some key posts. The content attention and temporal attention in the ACAMI model learn importance weights for both content and temporal information of events which selectively attend to important content and temporal characteristic of an event. So the ACAMI model can weigh the importance of each post within a group and attend to key features in a finer post scale that may be ignored by the CAMI model. That is to say, the ACAMI model is more robust against high noise with the help of the attention module thus achieves a better performance.

6. Conclusion

In this paper, we have proposed the ACAMI model for both misinformation identification and early detection tasks. Moreover, we propose an Event2vec method to learn representations for events with massive posts in social media. Besides, content and temporal co-attention can help still mine key content and temporal features from thousands of posts with high noise and simplify the grouping procedure in the proposed models. Extensive experiments on two typical social media datasets have demonstrated the effectiveness of the ACAMI model than both conventional feature-engineering-based methods and a RNN-based method. We also illustrate temporal properties of information in social media and visualize what the proposed model can capture, which will help shape more exact real-world social media scenarios for misinformation identification. Then we can better accomplish the task of misinformation identification and early detection.

In the future, we may incorporate cause and effect relationship among misinformation and trending issues into the proposed models. Acquiring all-round understanding of misinformation in social media, we can build a more effective, robust and interpretable model.

References

References

[1] K. K. Kumar, G. Geethakumari, Detecting misinformation in online social networks using cognitive psychology, Human-centric Computing and Information Sciences 4 (1) (2014) 1.

[2] C. Castillo, M. Mendoza, B. Poblete, Information credibility on twitter, in: WWW, 2011, pp. 675–684.

[3] A. Gupta, H. Lamba, P. Kumaraguru, A. Joshi, Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy, in: WWW, 2013, pp. 729–736.

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

[5] S. Kwon, M. Cha, K. Jung, W. Chen, Y. Wang, Prominent features of rumor propagation in online social media, in: ICDM, 2013, pp. 1103–1108.

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

[7] Z. Zhao, P. Resnick, Q. Mei, Early detection of rumors in social media from enquiry posts, in: WWW, 2015, pp. 1395–1405.

[8] Z. Jin, J. Cao, Y.-G. Jiang, Y. Zhang, News credibility evaluation on microblog with a hierarchical propagation model, in: ICDM, 2014, pp. 230–239.

[9] Z. Jin, J. Cao, Y. Zhang, J. Luo, News verification by exploiting conflicting social viewpoints in microblogs, in: AAAI, 2016, pp. 2972–2978.

[10] K. D. Giudice, Crowdsourcing credibility: The impact of audience feedback on web page credibility, Proceedings of the American Society for Information Science and Technology 47 (1) (2010) 1–9.

[11] S. Y. Rieh, G. Y. Jeon, J. Y. Yang, C. Lampe, Audience-aware credibility: From understanding audience to establishing credible blogs., in: ICWSM, 2014.

[12] J. Ma, W. Gao, P. Mitra, S. Kwon, B. J. Jansen, K.-F. Wong, M. Cha, Detecting rumors from microblogs with recurrent neural networks, in: IJCAI, 2016, pp. 3818–3824.

[13] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, G. Penn, Applying convolutional neural networks concepts to hybrid nn-hmm model for speech recognition, in: ICASSP, 2012, pp. 4277–4280.

[14] N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences, in: ACL, 2014, pp. 655–665.

[15] Q. Liu, F. Yu, S. Wu, L. Wang, A convolutional click prediction model, in: CIKM, 2015, pp. 1743–1746.

[16] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: CVPR, 2017.

[17] A. Tamar, S. Levine, P. Abbeel, Y. Wu, G. Thomas, Value iteration networks, in: NIPS, 2016, pp. 2146–2154.

[18] S. Kumar, R. West, J. Leskovec, Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes, in: WWW, 2016, pp. 591–602.

[19] L. Wu, H. Liu, Tracing fake-news footprints: Characterizing social media messages by how they propagate, in: WSDM, 2018.

[20] L. Wu, J. Li, X. Hu, H. Liu, Gleaning wisdom from the past: Early detection of emerging rumors in social media, in: SDM, 2017, pp. 99–107.

[21] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11) (1998) 1254–1259.

[22] V. Mnih, N. Heess, A. Graves, et al., Recurrent models of visual attention, in: NIPS, 2014, pp. 2204–2212.

[23] J. Ba, V. Mnih, K. Kavukcuoglu, Multiple object recognition with visual attention, in: ICLR, 2015.

[24] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: ICML, 2015, pp. 2048–2057.

[25] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, R. Nevatia, Abc-cnn: An attention based convolutional neural network for visual question answering, arXiv preprint arXiv:1511.05960.

[26] H. Xu, K. Saenko, Ask, attend and answer: Exploring question-guided spatial attention for visual question answering, in: ECCV, 2016, pp. 451–466.

[27] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: CVPR, 2016, pp. 21–29.

[28] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, Z. Zhang, The application of two-level attention models in deep convolutional neural network for fine-grained image classification, in: CVPR, 2015, pp. 842–850.

[29] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: ICLR, 2015.

[30] T. Luong, H. Pham, C. D. Manning, Effective approaches to attention-based neural machine translation, in: EMNLP, 2015, pp. 1412–1421.

[31] W. Yin, S. Ebert, H. Schütze, Attention-based convolutional neural network for machine comprehension, arXiv preprint arXiv:1602.04341.

[32] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: NAACL, 2016, pp. 1480–1489.

[33] A. M. Rush, S. Chopra, J. Weston, A neural attention model for abstractive sentence summarization, in: EMNLP, 2015.

[34] B. Dhingra, H. Liu, W. W. Cohen, R. Salakhutdinov, Gated-attention readers for text comprehension, in: ACL, 2017.

[35] R. Kadlec, M. Schmid, O. Bajgar, J. Kleindienst, Text understanding with the attention sum reader network, in: ACL Workshop, 2016.

[36] L. Wang, Z. Cao, G. de Melo, Z. Liu, Relation classification via multi-level attention cnns, in: ACL, 2016.

[37] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, B. Xu, Attention-based bidirectional long short-term memory networks for relation classification, in: ACL, 2016, p. 207.

[38] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, Y. Bengio, Attention-based models for speech recognition, in: NIPS, 2015, pp. 577–585.

[39] A. Graves, Generating sequences with recurrent neural networks, arXiv preprint arXiv:1308.0850.

[40] G. E. Hinton, Learning distributed representations of concepts, in: Proceedings of the eighth annual conference of the cognitive science society, Vol. 1, Amherst, MA, 1986, p. 12.

[41] G. E. Hinton, S. T. Roweis, Stochastic neighbor embedding, in: NIPS, 2003, pp. 857–864.

[42] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors, Cognitive Modeling 5 (3) (1988) 1.

[43] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, Journal of Machine Learning Research 3 (Feb) (2003) 1137–1155.

[44] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, P. Kuksa, Natural language processing (almost) from scratch, Journal of Machine Learning Research 12 (Aug) (2011) 2493–2537.

[45] E. H. Huang, R. Socher, C. D. Manning, A. Y. Ng, Improving word representations via global context and multiple word prototypes, in: ACL, 2012, pp. 873–882.

[46] A. Mnih, G. Hinton, Three new graphical models for statistical language modelling, in: ICML, ACM, 2007, pp. 641–648.

[47] A. Mnih, G. E. Hinton, A scalable hierarchical distributed language model, in: NIPS, 2009, pp. 1081–1088.

[48] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: NIPS, 2013, pp. 3111–3119.

[49] J. Mitchell, M. Lapata, Composition in distributional models of semantics, Cognitive Science 34 (8) (2010) 1388–1429.

[50] A. Yessenalina, C. Cardie, Compositional matrix-space models for sentiment analysis, in: EMNLP, 2011, pp. 172–182.

[51] R. Socher, J. Pennington, E. H. Huang, A. Y. Ng, C. D. Manning, Semi-supervised recursive autoencoders for predicting sentiment distributions, in: EMNLP, 2011, pp. 151–161.

[52] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts, et al., Recursive deep models for semantic compositionality over a sentiment treebank, in: EMNLP, Vol. 1631, 2013, p. 1642.

[53] Q. V. Le, T. Mikolov, Distributed representations of sentences and documents., in: ICML, 2014, pp. 1188–1196.

[54] Y. Le Cun, Y. Bengio, Word-level training of a handwritten word recognizer based on convolutional neural networks, in: IAPR, Vol. 2, 1994, pp. 88–92.

[55] V. Jain, J. F. Murray, F. Roth, S. Turaga, V. Zhigulin, K. L. Briggman, M. N. Helmstaedter, W. Denk, H. S. Seung, Supervised learning of image restoration with convolutional networks, in: ICCV, 2007, pp. 1–8.

[56] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (1) (2013) 221–231.

[57] F. Yu, Q. Liu, S. Wu, L. Wang, T. Tan, A convolutional approach for misinformation identication, in: IJCAI, 2017, pp. 3818–3824.

[58] S. Sukhbaatar, J. Weston, R. Fergus, et al., End-to-end memory networks, in: NIPS, 2015, pp. 2440–2448.

[59] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, R. Socher, Ask me anything: Dynamic memory networks for natural language processing, in: ICML, 2016, pp. 1378–1387.

[60] F. Yang, Y. Liu, X. Yu, M. Yang, Automatic detection of rumor on sina weibo, in: SIGKDD Workshop on Mining Data Semantics, 2012, p. 13.

[61] S. Lai, K. Liu, S. He, J. Zhao, How to generate a good word embedding, IEEE Intelligent Systems 31 (6) (2016) 5–14.

[62] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034.

[63] C. Vondrick, A. Khosla, T. Malisiewicz, A. Torralba, Hoggles: Visualizing object detection features, in: ICCV, 2013, pp. 1–8.

Feng Yu received his B.S. degree in electrical engineering and automation from Harbin Institute of Technology, China, in 2015. He is a PhD student in Center for Research on Intelligent Perception and Computing (CRIPAC) at National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. His research interests are in machine learning, data mining, information retrieval, and behavior analysis in cyberspace. He has published several papers in the areas of data mining, information retrieval and knowledge management at international conferences, such as TIST, IJCAI, SIGIR and CIKM.

Qiang Liu received his B.S. degree in electronic science from Yanshan University, China, in 2013. He is currently working toward the Ph.D. degree in Center for Research on Intelligent Perception and Computing (CRIPAC) at National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA) and the University of Chinese Academy of Sciences (UCAS), Beijing, China. His research interests include machine learning, data mining, user modeling and information credibility evaluation. He has published several papers in the areas of data mining and information retrieval at international journals and conferences, such as IEEE TKDE, AAAI, SIGIR, CIKM and ICDM.

Shu Wu received his B.S. degree from Hunan University, China, in 2004, M.S. degree from Xiamen University, China, in 2007, and his Ph.D. degree from Department of Computer Science, University of Sherbrooke, Quebec, Canada, all in computer science. He is an Associate Professor in Center for Research on Intelligent Perception and Computing (CRIPAC) at National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA). He has published more than 20 papers in the areas of data mining and information retrieval at international journals and conferences, such as IEEE TKDE, IEEE THMS, AAAI, ICDM, SIGIR, and CIKM. His research interests include data mining, information retrieval and recommendation systems.

Liang Wang received both the BEng and MEng degrees from Anhui University in 1997 and 2000, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA) in 2004. From 2004 to 2010, he was a research assistant at Imperial College London, United Kingdom, and Monash University, Australia, a research fellow at the University of Melbourne, Australia, and a lecturer at the University of Bath, United Kingdom, respectively. Currently, he is a full professor of the Hundred Talents Program at the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published in highly ranked international journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence and IEEE Transactions on Image Processing, and leading international conferences such as CVPR, ICCV, and ICDM. He is a senior member of the IEEE, and an IAPR Fellow.

Tieniu Tan received his BSc degree in electronic engineering from Xi'an Jiaotong University, China, in 1984, and his MSc and PhD degrees in electronic engineering from Imperial College London, U.K., in 1986 and 1989, respectively. In October 1989, he joined the Computational Vision Group at the Department of Computer Science, The University of Reading, Reading, U.K., where he worked as a Research Fellow, Senior Research Fellow and Lecturer. In January 1998, he returned to China to join the National Laboratory of Pattern Recognition (NLPR), Institute of Automation of the Chinese Academy of Sciences (CAS), Beijing, China, where he is currently Professor and former director (1998-2013) of the NLPR and Center for Research on Intelligent Perception and Computing (CRIPAC), and was Director General of the Institute (2000-2007). He was also Vice President of the Chinese Academy of Sciences (2015-2016). His current research interests include biometrics, image and video understanding, and information content security. Dr. Tan is a Fellow of CAS, TWAS (The World Academy of Sciences for the advancement of science in developing countries), IEEE and IAPR, and an International Fellow of the UK Royal Academy of Engineering. He is or has served as Associate Editor or member of editorial boards of many leading international journals including IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions on Automation Science and Engineering, IEEE Transactions on Information Forensics and Security, IEEE Transactions on Circuits and Systems for Video Technology, Pattern Recognition, Pattern Recognition Letters, Image and Vision Computing, etc.