Learning Contextual Features with Multi-head Self-attention for Fake News Detection

Yangqian Wang[1], Hao Han[1], Ye Ding[2], Xuan Wang[1], and Qing Liao[1,3(✉)]

[1] Harbin Institute of Technology (Shenzhen), Shenzhen 518055, China
1260893592wyq@gmail.com, hanhao@stu.hit.edu.cn, wangxuan@cs.hitsz.edu.cn, liaoqing@hit.edu.cn

[2] Dongguan University of Technology, Dongguan, China
dingye@dgut.edu.cn

[3] Peng Cheng Laboratory, Shenzhen, China

Abstract. Automatic fake news detection has attracted great concern in recent years due to it's tremendous negative impacts on public. Since fake news is usually written to mislead readers, lexical features based methods have great limitations. Previous work has proven the effectiveness of contextual information for fake news detection. However, they ignore the influence of sequence order when extract features from contextual information. Inspired by transformer technique, we propose Contextual Features with Multi-head Self-attention model(CMS) to extract features from contextual information for fake news detection. CMS can automatic capture the dependencies between contextual information and learning a global representation from contextual information for fake news detection. Experimental results on the real-world data demonstrate the effectiveness of the proposed model.

Keywords: Fake news detection · Contextual information · Multi-head self-attention

1 Introduction

In recent years, social media has provided great convenience to the public because of it's more timely and more easier to share and discuss across various social media platforms. However, social media has become an ideal place for fake news propagating due to the lack of supervision mechanism. And the widespread of fake news on social media arise tremendous negative impacts on individual and public. For instances, many commentator believe that the result of the 2016 US presidential election was affected by fake news [1].

In order to mitigate the negative impact of fake news, several fact-checking organizations take enormous labour and times to identify fake news from massive Internet content. The automatic detection of fake news is a critical step in our fight against fake news. Early works for detect fake news often designed a

R. Xu et al. (Eds.): ICCC 2019, LNCS 11518, pp. 132–142, 2019.
https://doi.org/10.1007/978-3-030-23407-2_11

comprehensive sets of hand-crafted features, which is lack of generalization and almost impossible to design all-encompassing features since fake news are usually written across different types. Moreover, we don't have a comprehensive grasp of the linguistic characteristics of fake news yet. There are still great limitations to the methods that only based on text content alone.

On the other hand, news often accompanied with related contextual information, such as speaker of news and there historical data, etc. These contextual information provide more useful information beyond textual content for detecting fake news. In order to learn the high-level feature representation from contextual information, previous works mainly use deep learning methods to automatic capture the useful features. Specifically, contextual information has be seen as a sequence, then Convolutional Neural Network (CNN) based and Recurrent Neural Network (RNN) based methods are applied to capture the dependencies. However, these methods have their own shortcomings. For example, CNN-based methods only can learning local patterns while fails to capture the global feature representation. RNN-based methods usually take the input sequence as an ordered sequence and update the hidden states step by step. Based on this, it's natural for us to ask:

– Will different sequences order lead to quite different results?
– If the first hypothesis is true, then how did we determine the sequence order that can achieve a better performance?

To explore the influence of sequential of contextual information, we conduct a simple experiment based on LIAR dataset [2]. LIAR dataset contains a wealth of contextual information. We utilize a LSTM to encode the contextual information. Obviously, different order of sequence will affect the learning procedure of LSTM. We take followed sequence as examples:

- **S1**: credit history of speaker, context, speaker name, subject of statement, job title of speaker, home state of speaker, party affiliation of speaker.
- **S2**: context, credit history of speaker, speaker name, subject of statement, job title of speaker, home state of speaker, party affiliation of speaker.
- **S3**: subject of statement, speaker name, job title of speaker, party affiliation of speaker, credit history of speaker, home state of speaker, context.
- **S4**: credit history of speaker, context, party affiliation of speaker, speaker name, job title of speaker, home state of speaker, subject of statement.

Figure 1 shows the performance in different orders. As Fig. 1 shows, the performance is affected by the sequential. For example, the detection accuracy get 32.5% when input sequence take S3 while only get 25.7% when take S4. Given the fact that the number of all combinations of sequence order may be very large, it's not wise and inefficient to try all the possible combinations and then choose an optimal sequence manually.

To address above-mentioned challenges, in this paper, we proposed a novel hybrid model based on deep learning to automatic learning the features for fake news detection. Specifically, a Text-CNN [3] is applied to learn features

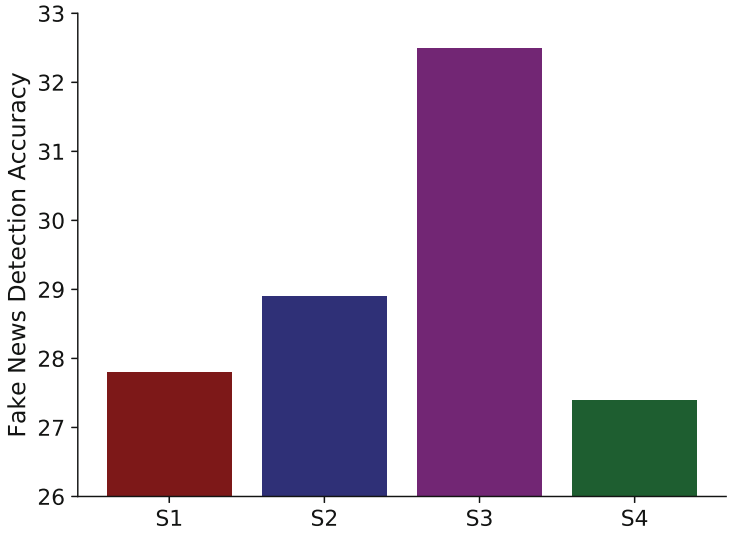

Fig. 1. Performance of LSTM on different order of contextual information.

from textual content. Inspired by transformer technique, we utilize multi-head self-attention to capture the dependencies between contextual information and learning a global feature representation of contextual information to assist in detecting fake news. Self-attention mechanism can ignore the influence of the relative position of each element in the sequence and capture the global representation of sequence. The contribution of this paper can be summarize as follows:

- Through experiments, we confirmed that performance of fake news detection is affected by the sequential of contextual information.
- We proposed a novel CMS model that can ignore the influence of sequential and learning a global representation from contextual information for fake news detection.
- Experiments on real-world fake news dataset demonstrate the effectiveness of the proposed CMS model.

The rest of the paper is organized as follows: Sect. 2 will briefly present the related work. The details of the proposed method will be introduced in Sect. 3. Experimental results are presented in Sects. 4 and 5 conclude this study.

2 Related Work

Fake news detection has attracted great concern in recent years due to it's tremendous negative impacts on public. With the help of big data technology [4–6] and machine learning, automatic fake news detection has made great advance. Early works on detecting fake news mainly focus on designing a complementary set of

hand-crafted features based on linguistic features [7–11]. Unfortunately, we don't have a comprehensive grasp of the characteristics of fake news yet. And this procedure require large efforts to explore the effectiveness of manual features.

Recently, deep learning based methods were proposed to automatic learning the patterns to detect fake news. For example, Ma *et al.* [12] proposed a deep neural network based on RNN to capture the temporal and textual features from rumour posts. Liu and Wu [13] utilize CNN and GRU to capture the useful patterns from user profiles. Multi-Modal fake news detection that integrate the textual feature and visual feature get a certain improvements compared to textual feature only [14,15].

Wang [2] presented a new publicly available dataset for fake news detection in 2017. It included 12.8k manually labeled short statements and each statements accompanied with rich contextual information, i.e., speaker name, job title, political party affiliation, etc. Wang also proposed a Hybrid CNN to detect fake news based this dataset. Based on LIAR dataset, some hybrid deep neural network methods were proposed to facilitate fake news detection. For instances, Long et al. [16] applied LSTM and attention mechanism to capture the patterns and Karimi et al. [17] proposed Multi-Source Multi-Class Fake News Detection (MMFD) models to discriminate different degrees of fakeness. However, above methods ignore the influence of sequence order of contextual information as we discussed in Sect. 1.

Different with above mentioned methods, we proposed a novel model CMS that can ignore the sequential order and capture the global representation from contextual information for fake news detection.

3 Methodology

3.1 Problem Definition

Let $X = \{x_1, x_2, ..., x_N\}$ be a set of news, and each news can be denoted as $x = \{s, c\}$, where s, c represent the statement and contextual information of the news item, respectively. Each statement is consisted of several words, which can be denoted as $s = \{w_1, w_2, ..., w_n\}$, where n is the length of statement. The representation of contextual information can be denoted as $c = \{c_1, c_2, ..., c_t\}$, where t is the number of features in contextual information, such as speaker name, subject of statement, etc. $Y = \{y_1, y_2, ..., y_N\}$ denotes the corresponding label. Our goal is using the news set X and the corresponding label set $Y$to learn a multi-classification model $\mathcal{F}$, which can predict the fakeness for unlabeled news.

3.2 Model Overview

Figure 2 depicts the architecture of CMS. CMS consists of two main components, a module for extracting the linguistic feature of news statement, and a module for capturing the contextual dependencies. Specifically, CMS is composed of three parts as followed:

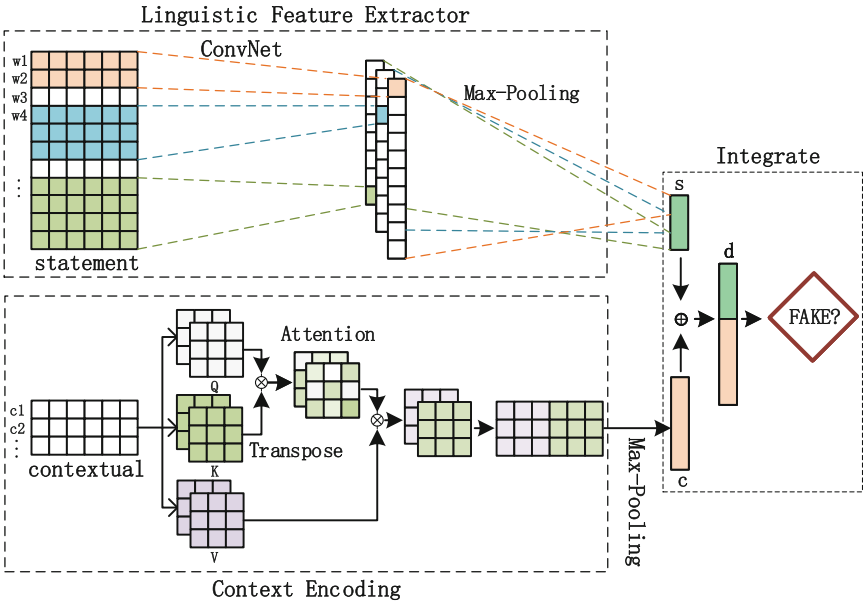

Fig. 2. The structure of CMS.

- Linguistic Feature Extractor: We use a Text-CNN to extract linguistic representation from statement. Statement is represented as word embedding matrix through word embedding layer and then Text-CNN is applied to learning an representation vector $\boldsymbol{s}$.
- Context Encoding: To extract the global representation of contextual information and ignore the influence of the sequence order in contextual information, we use multi-head self-attention mechanism to encode the contextual information and outputs a high-level representation vector, denoted as $\boldsymbol{c}$.
- Integrate: The outputs of the two above components are concatenated and finally obtain the hidden representation of news, denoted as $\boldsymbol{d}$, which will fed into a fully connected layer for classification.

Linguistic Feature Extractor. We use Text-CNN [3] to extract linguistic features from statement. Statement is represented as a matrix of word embeddings denoted as $\boldsymbol{W} \in \mathbb{R}^{n \times e}$, where e is the dimensions of the word embedding. Text-CNN usually applied multiple kernel with multiple filters to extract features with different granularities. A filter $\boldsymbol{f} \in \mathbb{R}^{l \times e}$ is convolved with word embedding matrix and obtain a corresponding feature map $\boldsymbol{P} \in \mathbb{R}^{n-l+1}$, where l is the length of the filter. For each $p_j \in \boldsymbol{P}$, the filter operation can be formulated as follows:

$$p_j = \mathcal{C}(f \odot W_{j:j+l-1}) \tag{1}$$

where $W_{j:j+l-1}$ stand for l consecutive words starting with the j-th word in the statement, and $\mathcal{C}(\cdot)$ is a non-linear activation function, such as $ReLU$. We use

max-pooling on each feature map to get the most important information:

$$v = \text{Max-Pooling}(P) \tag{2}$$

Now, we get the corresponding feature for one particular filter. The complementary feature for statement can be obtained by concatenating all pooling feature vectors:

$$\boldsymbol{s} = [v_1, v_2, ..., v_k] \tag{3}$$

where k is the number of filters.

Context Encoding. In order to capture the global representation of contextual information without regard of the order of features, we employ a variant of transformer [18] as the core of the module.

Usually, the credibility of news is context-dependent, and the interactions between features in contextual information are important to fake news detection. Self-attention is used to capture the informative interactions between features in contextual information. In our case, the self-attention process is based on the assumption that all features in contextual information are closely related and disorder. The output of an attention function is a map of a query and a set of key-value pairs, where the query, keys, values, and output are all vectors. The weighted sum of values are the output, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key. In practice, a set of queries, keys and values are packed together into matrices $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$, respectively. Contextual information is represented as a matrix of word embedding denoted as $\boldsymbol{C} \in \mathbb{R}^{t \times e}$. $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ are three different subspace matrix of $\boldsymbol{C}$, which is get by linear projection. So the process of self-attention with contextual information can be formulate as:

$$Attention\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{4}$$

where d_k is the dimension of vector in $\boldsymbol{Q}$. Now, a new representation of contextual information is got through a self-attention head. For the purpose of allowing the module to obtain high-quality representation of features, multi-head self-attention is used to drawing features interactions with different subspace of $\boldsymbol{C}$ jointly:

$$MultiHead\left(\boldsymbol{C}\right) = Concat\left(Attention_1, ..., Attention_H\right) \tag{5}$$

where H is the number of heads. After that, a fully connected layer is followed and the output size of the full connected network is equal to $\boldsymbol{C}$. In particular, the above compute process can be repeated K times, and the output of previous layer is the input of next layer. Taking a max-pooling operation on the output of last layer, denoted as $\boldsymbol{C'} \in \mathbb{R}^{t \times e}$, and obtain a global representation of contextual information:

$$\boldsymbol{c} = \text{Max-Pooling}(\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_e) \tag{6}$$

where $\boldsymbol{r}_i$ is the i-th columns in $\boldsymbol{C'}$. So the final representation of contextual information $\boldsymbol{c} \in \mathbb{R}^e$.

Integrate. The final feature representation of news, i.e., $\boldsymbol{d}$, can be obtained by concatenating $\boldsymbol{s}$ and $\boldsymbol{c}$. After that, $\boldsymbol{d}$ is fed into a fully connected layer followed by a *softmax* activation function to predict the label of news:

$$\boldsymbol{l} = softmax(\boldsymbol{W}^T \boldsymbol{d} + b) \tag{7}$$

where $\boldsymbol{l} \in \mathbb{R}^M$, M is the number of class. The loss function can be formulated as:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{y}_i \log \boldsymbol{l}_i + \frac{\lambda}{2} \|\Theta\|_2^2 \tag{8}$$

where $\boldsymbol{y}_i$ is the corresponding ground-truth, Θ denote the parameters of CMS, and λ is the trade-off coefficient of L_2 regularizer.

4 Experiment

4.1 Dataset

In order to evaluate the performance of CMS, we conducted a series of comparative experiments in a real-world dataset LIAR published in [2]. LIAR dataset contains a total of 12.8k manual labeled short statements, and each statement contains rich contextual information. Table 1 show the details of LIAR dataset.

Table 1. Details of *LIAR* dataset

Contextual information	speaker, speaker's job, state, party, context, credit history of speaker, subject of statement
Label	pants-fire, false, barely-true, half-true, mostly-true, true

4.2 Experimental Setting

We use pytorch[1] to implement the proposed model. We use 300-dimension pretrained word2vec embedding [19] to initialize the word embeddings. We utilize batch size 64, and Adam optimizer [20] is adopted as the optimizer with learning rate 0.0001. L_2 regularization coefficient is $1e^{-4}$. In Linguistic Feature Extractor, we take (2,3,4) as kernel size and each kernel has 50 filters. As for Contextual Encoding component, the number of layers and heads is 3, 2, respectively. We use the average accuracy of 10 trials as the performance metric.

[1] https://pytorch.org/.

4.3 Performance Comparison

To demonstrate the effectiveness of CMS, we compare the following methods:

- **Random**: Randomly selecting the class of a test sample.
- **Majority**: Majority method choose a label that is the greater part of total dataset. In our experiment, we labeled each test sample as *half-true*.
- **Text-CNN**: Text-CNN model only use statement to classify news.
- **Hybrid-CNN**: Text-CNN is applied to learning textual feature from statements, and a CNN followed by a Bi-LSTM was applied to extract feature from contextual information.
- **LSTM-Attention**: Two LSTM was applied to extract features, one for textual context with attention, and another for contextual information.
- **MMFD**: CNN-LSTM was proposed to extract features from textual content and contextual information. Then the features were summed with weights by attention mechanism.

It is worth noting that the above-mentioned methods, Hybrid-CNN and LSTM-Attention use different combinations of contextual information and select a best performance manually, while our proposed model can be fully automated to achieve this goal. Table 2 shows the performance of all methods. As we can see, Text-CNN achieve worst performance since it only based on textual content of statement. Hybrid-CNN, LSTM-Attention and MMFD get a certain improvements compared to Text-CNN. Beyond the mentioned contextual information, MMFD also add the verdict reports generated by experts in *politifact.com* to further improve the detection accuracy. Even though, our proposed model still gets 6.5% improvements on detection accuracy compared to MMFD. Compared to LSTM-Attention and Hybrid-CNN, CMS can ignore the sequence order and capture the global features from contextual information and finally get 45.3% accuracy.

Table 2. Performance of detecting fake news.

Method	Accuracy(%)
Majority	20.8
Random	17.4
Text-CNN	21.7
Hybrid-CNN	27.4
MMFD	38.8
LSTM-Attention	41.5
CMS	45.3

4.4 Effectiveness of CMS

To evaluate the effectiveness of CMS, we conduct a comparison experiment to illustrate the superiority of CMS. We take several sequence orders and apply different methods to extract features from these sequence order. For the sake of fairness, we keep the textual features extractor as Text-CNN to all compared methods. Figure 3 shows the detailed results. The input sequence keeps consistent with Sect. 1. It is clearly show that the performance of Hybrid-CNN, LSTM-Attention and MMFD is severely constrained by the sequence order. For example, all compared methods achieve highest accuracy when adopt sequence S3 and get 32.5%, 40.0%, 41.5%, respectively. Such results may indicate that sequence S3 is a more appropriate order when extract features from contextual information. However, the performance can be very bad when those methods meets some not so appropriate sequence order. For example, LSTM-Attention only get 25.7% accuracy when the input sequence is S4.

In contrast, CMS always maintaining a better performance no matter how the sequence order is. Table 3 shows the detail results. As we can see, other methods

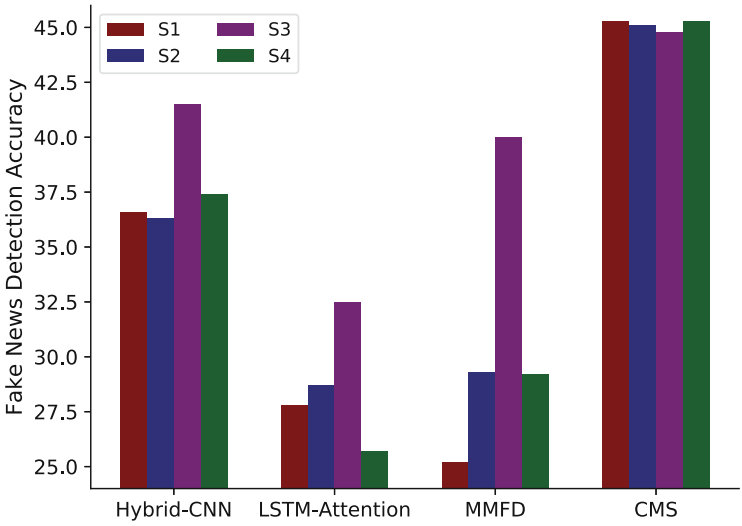

Fig. 3. The performance of CMS and compared methods on different sequence orders.

Table 3. Difference of sequence order.

Method	Best accuracy(%)	Worst accuracy(%)	Difference(%)
Hybrid-CNN	41.5	36.3	5.2
LSTM-Attention	32.5	25.7	6.8
MMFD	40.0	25.2	14.8
CMS	45.3	44.8	0.5

have a huge gap between the best sequence order and worst sequence order, while CMS only have 0.5% difference. Hence, our proposed model not only achieve better performance, but also keeps stable on accuracy. This further illustrate that CMS can neglect the order influence and capture the global patterns effectively.

5 Conclusion

In this paper, we proposed a novel hybrid model based on deep neural network to automatic learning the useful features from textual content and contextual information. Inspired by transformer technique, we applied multi-head self-attention to extract features from contextual information. Multi-head self-attention mechanism can ignore the distance and guarantee that each contextual information will be considered when capture the dependencies from contextual information. Experimental results further demonstrate the effectiveness of CMS method.

Acknowledgements. This work is partly supported by National Key Research and Development Program of China (2017YFB0802204) and National Natural Science Foundation of China (No.U1711261).

References

1. Allcott, H., Gentzkow, M.: Social media and fake news in the 2016 election. J. Econ. Perspect. **31**(2), 211–236 (2017)
2. Wang, W.Y.: "liar, liar pants on fire": a new benchmark dataset for fake news detection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 422–426 (2017)
3. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, 25–29 October 2014, Doha, Qatar, A meeting of SIGDAT, A Special Interest Group of the ACL, pp. 1746–1751 (2014)
4. Kashlev, A., Lu, S., Mohan, A.: Big data workflows: a reference architecture and the dataview system. Serv. Trans. Big Data **4**, 1–19 (2017)
5. Zhang, L.J., Zeng, J.: 5c, a new model of defining big data. Serv. Trans. Big Data **4**, 48–61 (2017)
6. Sinanc, D., Demirezen, M., Sagiroglu, S.: Evaluations of big data processing. Serv. Trans. Big Data **3**, 44–54 (2016)
7. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 675–684. ACM (2011)
8. Feng, S., Banerjee, R., Choi, Y.: Syntactic stylometry for deception detection. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers, vol. 2, pp. 171–175. Association for Computational Linguistics (2012)
9. Yang, F., Liu, Y., Yu, X., Yang, M.: Automatic detection of rumor on sina weibo. In: Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, vol. 13. ACM (2012)

10. Liao, Q., Wang, W., Han, Y., Zhang, Q.: Analyzing the influential people in sina weibo dataset. In: 2013 IEEE Global Communications Conference, GLOBECOM 2013, Atlanta, GA, USA, 9–13 December 2013, pp. 3066–3071 (2013)
11. Liao, Q., Guan, N., Zhang, Q.: Logdet divergence based sparse non-negative matrix factorization for stable representation. In: 2015 IEEE International Conference on Data Mining, ICDM 2015, Atlantic City, NJ, USA, 14–17 November 2015, pp. 871–876 (2015)
12. Ma, J., et al.: Detecting rumors from microblogs with recurrent neural networks. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, USA, pp. 3818–3824 (2016)
13. Liu, Y., Wu, Y.F.B.: Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
14. Wang, Y., et al.: EANN: event adversarial neural networks for multi-modal fake news detection. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 849–857. ACM (2018)
15. Jin, Z., Cao, J., Guo, H., Zhang, Y., Luo, J.: Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In: Proceedings of the 25th ACM International Conference on Multimedia, pp. 795–816. ACM (2017)
16. Long, Y., Lu, Q., Xiang, R., Li, M., Huang, C.R.: Fake news detection through multi-perspective speaker profiles. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pp. 252–256 (2017)
17. Karimi, H., Roy, P., Saba-Sadiya, S., Tang, J.: Multi-source multi-class fake news detection. In: Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, USA, 2018, pp. 1546–1557 (2018)
18. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, pp. 6000–6010 (2017)
19. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, 2–4 May 2013, Workshop Track Proceedings (2013)
20. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015)