False Information Detection on Social Media via a Hybrid Deep Model

Lianwei Wu, Yuan Rao(✉), Hualei Yu, Yiming Wang, and Ambreen Nazir

Lab of Social Intelligence and Complex Data Processing, School of Software, Xi'an Jiaotong University, Xi'an 710049, Shaanxi, China
{stayhungry,lily_yu1994,yimingwang}@stu.xjtu.edu.cn,
raoyuan@mail.xjtu.edu.cn,
Ambreen.nazir@ciitwah.edu.pk

Abstract. There is not only low-cost, easy-access, real-time and valuable information on social media, but also a large amount of false information. False information causes great harm to individuals, the society and the country. So how to detect false information? In the paper, we analyze false information further. We rationally select three information evaluation metrics to distinguish false information. We pioneer the division of information into 5 types and introduce them in detail from the definition, the focus, features, etc. Moreover, in this work, we propose a hybrid deep model to represent text semantics of information with context and capture sentiment semantics features for false information detection. Finally, we apply the model to a benchmark dataset and a Weibo dataset, which shows the model is well-performed.

Keywords: Information credibility evaluation · Rumor detection
Social media · Text classification · False information

1 Introduction

Social media based on users' generating content is represented by Facebook, Twitter, Microblog, WeChat etc., which hastens content generation, propagation and growth. Social media relies on information sharing, real time, interactivity and Diversity of content dissemination and the advantage of randomness of virtual identities, fragmented expression of sentiment, which profoundly transforms the way people access and propagate information. Nevertheless, the informational content conveyed on the social media is not always true and reliable. Recent studies of Gupta et al. [1] show that among the informative tweets, 52% of which are labeled as definitively credible, 35% of which seem credible, and 13% of which are definitively false. Besides, Lazer et al. [2] found that falsehood diffused significantly farther, faster, deeper, and broader than the truth in all categories of information. Experiment shows that the truth is about 6 times as long as falsehood to reach 1500 people. False information without identifying profoundly amplifies negative social emotions and significantly does harm to the

S. Staab et al. (Eds.): SocInfo 2018, LNCS 11186, pp. 323–333, 2018.
https://doi.org/10.1007/978-3-030-01159-8_31

harmony of society and country safety. For instance, Cambridge Analytica found that the events about United States presidential election were affected by fake news. Accordingly, it is one of the crucial and urgent issues that is how to classify false information quickly on social media.

To detect false information on social media, there are a series of studies aiming at detecting different types of false information. On one hand, most studies focus on a single type of false information, which main concentrates upon rumors, for example, early detection of rumors [3–5], rumor-spreading mechanism [7, 8], curbing the spread of rumors [9] et al. Some methods are interested in detecting fake news. For instance, Wang [10] releases a new, real, fake news dataset and proposes a novel model for fake news detection. Moreover, some existing algorithms inclines to detect disinformation on social media [11–13] and Bessi et al. think that disinformation is a conspiracy. Furthermore, there are many existing studies on spam detection [16–18]. On the other hand, a few researchers study several types of incredibility information, which are detecting misinformation and disinformation [20, 34], classifying fake, spam messages and legitimate messages etc. [35, 36].

Although, the existing work has made significant progress in classifying false information on social media. Unfortunately, it is obvious that the above work has several limitations. Firstly, studying on one type or several types of false information is not enough to cover all false information. Secondly, different types of false information have different characteristics. For example, rumors are controversy and unproven statements. However, fake news is unreal information. There is a lack of reasonable way to distinguish the types of false information.

To overcome the limitations mentioned, in this work, we summarize characteristics about different types of information and choose three representative dimensions of information evaluation metrics to divide the information on social media into 5 types, which are true information, rumors, biases, fake news, and spams. Next, we propose a novel hybrid deep model to classify these types of false information. In our method, we employ RCNN model to deeply capture to represent text semantics of information with context and design a model mixed with LSTM and ConvNet to learn text sentiment features. Particularly, sentiment features are the important features of classifying false information further. Lazer et al. [2] found that false rumors inspired replies expressing greater surprise, corroborating the novelty hypothesis, and greater disgust, whereas the truth inspired replies that expressed greater sadness, anticipation, joy, and trust.

Finally, the main contributions of this work are listed as follows:

- The paper firstly divides information into five types on social media, including four types of false information, and true information. We classify false information to evaluate information credibility more comprehensively.
- The paper proposes a hybrid deep model from deep semantics and sentiment features to detect false information, which significantly improves the performance of classifying false information.

2 The Types of Information

2.1 Three Information Evaluation Metrics

From a credible perspective, we can divide information into two aspects, which are true information and false information. Except this information which is confirmed authentically, scientifically, objectively and integrally. Others are false information. There are many obvious differences in false information on the social media. How to distinguish false information? Based on the issue from the three perspectives of the intention of information dissemination and the influence caused by information, the authenticity of information, we choose three representative metrics of information evaluation, which are purposiveness, harmfulness, and credibility to evaluate false information on social media.

Purposiveness is the intention, which can tell whether information dissemination has a strong orientation. We divide purposiveness of information into three types, which are purpose, purposelessness, and unclear purpose. Harmfulness refers to the negative effect of the spread of information on audience. **Harmfulness** intends to do harm to personal safety and to spawn economic loss. **Credibility** means the comprehension reflection of information receivers' subjective trust in the sources of information, objective evaluation on the quantity of information content, information disseminators' degree of trust [27]. We divide information credibility into three types, which are that information is true, information is false, and information needs to be verified.

2.2 The Types of Information

According to three information evaluation metrics and some literature [10, 12, 13, 23, 28, 29], we divide information into true information, rumors, biases, fake news and spams. The table below is a comparison of five types.

On the definition of the five types of information, true information is true, scientific, objective and complete. About the definition of the rumors, Kapferer et al. [32] reckoned that validating the trueness of the rumors needed to satisfy three factors: (1) there were reliable sources of information, (2) individuals expecting to know the truth, (3) information seeming to be true. In consequence, the paper defines that rumors are controversy and unproven statements. The definition of biases is essentially the reflection of cognition and feeling for realistic society. Biases mean exaggerating the facts, being taken out of context, and sampling bias, which affects the stance of social life and the value judgements of the public. Tandoc et al. [33] define fake news is unreal information made on purpose, which has characteristics that are fast speed of spreading, wide range of spreading and the structure of spreading that presents dispersive networks. Karlova et al. [20] also mentioned that fake news has the characteristic of deliberate deception. Spam usually includes casual and useless information, false advertisement, deceptive and trick information in the network. The existence of fake users and zombies is an important reason for spam proliferation.

Based on three information evaluation metrics, the definition of five types of information, taking into account the ease of classification of labelers, the focus of five types of information are different. Rumors focus on credibility of information that has not been confirmed. Biases tend to exaggerate of emotional language expression. Fake news focuses on the purpose of information with deliberate deception. Spam focuses on some advertising information and duplicate information. When we make sure the focus of five types of information are different, this will be five types of information that are mutually exclusive and do not overlap. Additionally, except five types of information, are there other types of information? In the recent literature, we find some researchers study misinformation. The reason why we do not classify misinformation as one type here is that the survival time of false positive information is very short. After misinformation is confirmed, it is either turned into true information, or turned into fake news.

We compare the five types of information as shown in Table 1. On purposiveness, rumors are unproven statements, which have no specific purpose. On harmfulness, different types of information credibility have difference between strong sides and weak sides. On credibility, we use true, to be confirmed, and false, levels to evaluate five types of information. In the next chapter, we will aim at classifying the five types of information.

Table 1. The comparison of the five types of information

	Characteristic	Purposiveness	Harmfulness	Credibility
True Information	True, Scientific, Objective and Complete	Yes	No	True
Rumors	Plausible, To be confirmed	Not Clear	Not too Strong	To be confirmed
Biases	Exaggerated, Fuzzy and damaging	Yes	Very strong	To be confirmed
Fake news	Deliberate deception, Misguided	Yes	Strong	False
Spams	Useless, Confused	Yes	Weak	To be confirmed

3 The Proposed Method

We have already divided information into five types in Sect. 2. In this Section, we propose a novel method to classify them on social media. The paper introduces the details of the method we proposed. Figure 1 shows the network structure of the method. We obtain semantic representation and sentiment information from social media text data. On semantic representation, we use RCNN model to obtain information with context as semantically as possible. On sentiment representation, we use ConvNet model to deeply represent sentiment features. Combined with semantic and sentiment representation, we classify the five types of information on social media. Next, the paper will show the details of the method as follows.

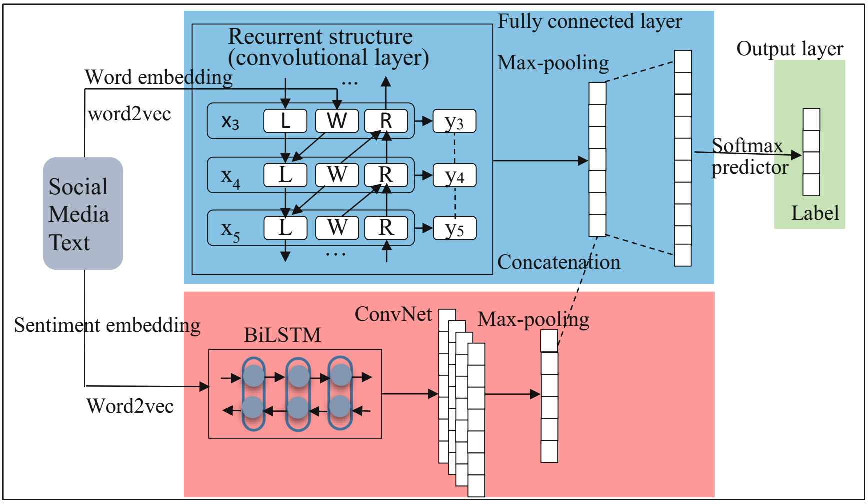

Fig. 1. The method we proposed for classifying false information (Color figure online)

3.1 Learning Text Semantic Representation from Microblog

The light blue part in Fig. 1 uses the RCNN model to describe the process of learning text semantic representation from single microblog. The RCNN model is essentially CNN model, which uses recurrent structure as the convolutional layer of CNN model. On the input of the model, we use a tool of Chinese word segmentation, which is named Jieba[1] to split text of the microblog and then remove stop words by Chinese stop words list firstly. Then we use pre-trained 50-dimensional word2vec embeddings from Chinese Wikipedia corpus. Embeddings obtained serve as the input of the RCNN model. In recurrent structure, L, R represents the semantics of all left side and right side contexts respectively, and W represents the word embedding of a specific word. The schemas of calculation of L and R are shown in Eqs. (1) and (2). We illustrate the equations with examples, we suppose that the input sentence of the recurrent structure is "Wisconsin is on pace to double the number of layoffs this year" and W_{x3}, W_{x4}, W_{x5} means "pace", "to", and "double" respectively. Then L_{x4} represents the semantics of the left side context "Wisconsin is on pace", R_{x4} represents the semantics of the right side context "double the number of layoffs this year". Next, x_4 is the concatenation of the left side context vector L_{x4}, the word embedding W_{x4} and the right side context vector R_{x4}. As shown in Eq. (3), the recurrent structure learns more semantic representation from text of microblog, compared to conventional neural network models, which only use a fixed window. We apply *tanh* activation function to x_i and send the result to the next layer y_i, which is a latent semantic vector.

$$L_{x4} = L_{x3} + W_{x5} \tag{1}$$

$$R_{x4} = R_{x5} + W_{x5} \tag{2}$$

$$\mathrm{x}_4 = [\mathrm{L}_{\mathrm{x}4}; \mathrm{W}_{\mathrm{x}4}; \mathrm{R}_{\mathrm{x}4}] \tag{3}$$

The results y_1, y_2, …, y_n of recurrent structure contain a lot of repeated contextual information. To reduce the repeated contextual information to prevent overfitting and decrease the input size of the next layer to cut down calculated amount, we apply a max-pooling that converts texts with various lengths into a fixed-length vector and send the result to next layer.

3.2 Learning Sentiment Representation from Microblog

Sentiment features are important features to distinguish false information. To obtain the sentiment features of microblogs on Sina weibo better, we construct a ConvNet model to learn sentiment representation. As shown in the light red part of Fig. 1, on the input of sentiment model, we build Chinese sentiment lexicon[2], negative word lexicon[3] and degree adverb lexicon[4] to extract sentiment words, negative words and degree adverb words from microblogs respectively. Like the input on 3.1 section, we use pre-trained 50-dimensional word2vec embedding from Wikipedia to represent the words. We extracted and formed sentiment embedding, which is named sentiment-word2vec. Then the concatenation inputs a convNet layer to capture sentiment features better. Finally, we use max-pooling layer to extract a fix-length vector to attain more representative sentiment features from the output of the convNet layer. Additionally, the results of output is sent to next layer to wait for merging.

In conclusion, we integrate the output results of 3.1 and 3.2 section into a concatenation. Then the concatenation enters full-connected layer and subsequently passes a softmax predictor to obtain the label of different types of information mentioned in Sect. 2.2.

4 Experiments

4.1 Experimental Settings

At the beginning, on the dataset of our experiments, we split Weibo dataset we built into 70% training set and 30% testing set. Specifically, the Weibo dataset contains five types of information, 70% of which are extracted as training set in the dataset of information in every type and the remaining is testing set.

Secondly, on experimental evaluation metrics, we choose several representative evaluation metrics for our experiments, which are **accuracy**, **recall** and **F1-score**. **Accuracy** is a necessary evaluation metric for all most classification experiments. **Recall** is coverage that measures how many positive examples are divided correctly. **F1-score** is the harmonic average of accuracy and recall. Generally, for above the evaluation metrics, the larger the values of evaluation metrics are, the better the effectiveness will be.

What's more, on the hyper-parameter settings of the baseline dataset and the Weibo dataset we built. We use 50-dimensional word2vec to embedding text of microblogs as the input of the method. We set the convolution window size to be 4 and filter sizes to be 100. On the calculation of loss functions, the last layer of the network firstly processes through softmax and then uses cross-entropy to calculate loss. We set the learning rate of the optimal gradient descent to be 10^{-2} and batch size to be 50.

4.2 Data

Here, we introduce two multi-type of information datasets, which are benchmark dataset – LIAR dataset [10] and Weibo dataset we built.

LIAR Dataset: LIAR dataset is a new human-labeled benchmark dataset from POLITIFACT.COM, which includes 12,836 short statements labeled for truthfulness, subject, context/venue, speaker, state, party, and prior history. LIAR dataset contains six fine-grained labels for the truthfulness ratings: pants-fire, false, barely-true, half-true, mostly-true, and true. The distribution of labels in the LIAR dataset is relatively well-balanced: except for 1,050 pants-fire cases, the instances for all other labels range from 2,063 to 2,638. Table 2 shows some random snippets from LIAR dataset.

Table 2. The LIAR dataset statistics

Dataset statistics	Num.
Training set size	10269
Validation set size	1284
Testing set size	1283
Avg. statement length (tokens)	17.9

Weibo Dataset: Considering public multi-type information datasets lack of annotation, especially, there is a blank in Chinese information credibility field. Consequently, we construct multi-type information dataset – the Weibo dataset based on Sina weibo, one of the biggest social platforms in China.

The overall sketch of the Weibo dataset is shown in Table 3. We collect more than 40 thousand microblogs that consist of 9600 true information, 8000 rumors, 8000 biases, 8000 fake news, and 8000 spams. The average microblog length in different types of information is different. Here the average microblog length refers to the number of words of single microblog. The average length of fake news is the longest, which is 115. Oppositely, spam is the shortest, which is only 93. Based on the dataset we built, we conduct a series of experiments to verify the performance of our method.

Table 3. The overall sketch of the Weibo dataset

The types of information	True information	Rumors	Biases	Fake news	Spams
Microblogs	9600	8000	8000	8000	8000
AVG. microblog length	104	111	98	115	93
MAX forwarding volume	224.5K	213.6K	232.7K	220.1K	150.3K

4.3 Results

We compare our method with other baseline methods on classifying false information. Table 4 shows the performance of different methods, which can outperform the state-of-the-art methods. Then we analyze the experimental results based on several methods with different inputs.

Table 4. Results of comparison with different methods for different dataset

	LIAR dataset			Weibo dataset		
	Accuracy	Recall	f1-score	Accuracy	Recall	F1-score
LR + all	0.263	0.532	0.342	0.359	0.624	0.456
SVM + all	0.271	0.564	0.354	0.375	0.645	0.474
CNN + all	0.273	0.486	0.355	0.368	0.677	0.477
RCNN + text	0.304	0.557	0.391	0.402	0.701	0.511
Our method	0.337	0.597	0.431	0.433	0.749	0.549

Comparison with Methods

All means text semantic features, and sentiment features are spliced together as a concatenation.

LR + All. We use **All** features as the input of logistic regression (LR) model to classify false information. Additionally, **SVM** is Support Vector Machines. **CNN** is convolutional neural network.

RCNN + Text. We use text semantic features as the inputs of the RCNN model to classify false information. The RCNN Lai et al. [18] proposed is a text classification model, which outperforms the state-of-the-art methods, particularly on document-level datasets.

RCNN + Text + Sentiment. We use the combination of RCNN and Text, sentiment features to evaluate multiple types information.

Our Method. In our method, obtaining text semantic features with context as much as possible by RCNN model, and getting sentiment features by convNet, are integrated as a concatenation as input of our method to use CNN model to evaluate multi-type information.

Results

The experimental results are shown in Table 4.

Contrasting LR + All, SVM + All, CNN + All and our method. These methods have the same inputs, which are All. The experimental results show that our method has the better performance than other classical methods. It proves that our method uses RCNN model to capture more text semantic information with context and use convNet to extract more text sentimental features.

Contrasting RCNN + Text, RCNN + Text + Sentiment and our method. Three methods are based on CNN improved model – RCNN model. However, for different inputs, three methods show different experimental performance. The experimental results of RCNN + Text + Sentiment method are superior to RCNN + Text, which indicates that sentiment features are favorable distinguishing features for classifying false information. Meantime the performance of our method is obviously optimal in the three methods. The experiments of our method firmly confirm constructing text semantic representation and sentiment representation to classify false information is effective.

4.4 Discussion

The method proposed uses RCNN model to represent text semantic information with context as much as possible, and represent sentiment features by convNet prominently, which achieves great success on classifying false information for evaluating information credibility, even though, there still exists many issues to solve in the future. There are a couple of examples.

- On text sentiment representation of microblogs, our sentiment representation of sentiment words, negative words and degree adverb words, is slightly shallow because we have not considered the relationship of the words and their context.
- Another, there is also a problem that the volume of the Weibo dataset, including 5 types of information microblogs based on Sina weibo, is insufficient. In the future, we will continue to crawl different types of microblogs and label them to expand the Weibo dataset.

5 Conclusions

In this work, to classify false information for evaluating information credibility on social media, we propose a novel method, which captures contextual text semantic information with the recurrent structure of RCNN model, and learns sentiment representation with convNet model. On a benchmark dataset and a Weibo dataset collected from Sina weibo, the experiment demonstrates that the method we proposed outperforms the state-of-the-art methods.

References

1. Gupta, A., Kumaraguru, P., Castillo, C., Meier, P.: TweetCred: real-time credibility assessment of content on Twitter. In: Aiello, L.M., McFarland, D. (eds.) SocInfo 2014. LNCS, vol. 8851, pp. 228–243. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-13734-6_16
2. Lazer, D., Baum, M., Benkler, Y., Berinsky, A.J., Greenhill, K.M.: The science of fake news. Science **359**(6380), 1094–1096 (2018)
3. Zhao, Z., Resnick, P., Mei, Q.: Enquiring minds: early detection of rumors in social media from enquiry posts. In: Proceedings of WWW, pp. 1395–1405 (2015)
4. Wu, K., Yang, S., Zhu, K.: False rumors detection on Sina Weibo by propagation structures. In: Proceedings of the 31st ICDE, pp. 651–662 (2015)
5. Ma, J., Gao, W., Wong, K.: Detect rumors in microblog posts using propagation structure via kernel learning. In: Proceedings of ACL (2017)

6. Ma, J., Gao, W., Wei, Z., Lu, Y., Wong, K.: Detect rumors using time series of social context information on microblogging websites. In: Proceedings of CIKM, pp. 1751–1754. ACM (2015)
7. Hu, Y., Pan, Q., Hou, W., He, M.: Rumor spreading model with the different attitudes towards rumors. Phys. A **502**, 331–344 (2018)
8. Wang, B., Chen, G., Fu, L.: DRIMUX: dynamic rumor influence minimization with user experience in social networks. Proc. TKDE **29**(10), 2168–2181 (2017)
9. Hamidian, S., Diab, M.: Rumor identification and belief investigation on Twitter. In: Proceedings of NAACL-HLT, pp. 3–8 (2016)
10. Wang, W.: "Liar, Liar Pants on Fire": a new benchmark dataset for fake news detection. In: Proceedings of ACL, pp. 422–426 (2017)
11. Imran, M., Castillo, C., Diaz, F., Vieweg, S.: Processing social media messages in mass emergency: survey summary. In: Proceedings of WWW, pp. 507–511 (2018)
12. Kahne, J., Bowyer, B.: Educating for democracy in a partisan age: confronting the challenges of motivated reasoning and misinformation. Am. Educ. Res. J. **54**(1), 3–34 (2017)
13. Bessi, A., Coletto, M., Davidescu, G.: Science vs. conspiracy: collective narratives in the age of misinformation. PLoS ONE **10**(2), 1–17 (2015)
14. Faris, R., Roberts, H., Etling, B., et al.: Partisanship, Propaganda, and Disinformation: Online Media and the 2016 US Presidential Election. Berkman Klein Center Research Publication (2017)
15. Marwick, A., Lewis, R.: Media Manipulation and Disinformation Online. Data & Society Research Institute, New York (2017)
16. Sedhai, S., Sun, A.: Semi-supervised spam detection in twitter stream. IEEE Trans. Comput. Soc. Syst. **5**(1), 169–175 (2018)
17. Kim, S, Chang, H., Lee, S., Yu, M., Kang, J.: Deep semantic frame-based deceptive opinion spam analysis. In: Proceedings of CIKM, pp. 1131–1140. ACM (2015)
18. Hai, Z., Zhao, P., Cheng, P., Yang, P.: Deceptive review spam detection via exploiting task relatedness and unlabeled data. In: Proceedings of EMNLP, pp. 1817–1826 (2016)
19. Chen, C., Wang, Y., Zhang, J.: Statistical features-based real-time detection of drifted Twitter spam. TIFS **12**(4), 914–925 (2017)
20. Volkova, S., Jang, J.: Misleading or falsification: inferring deceptive strategies and types in online news and social media. In: Proceedings of WWW, pp. 575–583 (2018)
21. Rajdev, M., Lee, K.: Fake and spam messages: detecting misinformation during natural disasters on social media. In: Web Intelligence and Intelligent Agent Technology, vol. 1, pp. 17–20. IEEE (2015)
22. Lai, S., Xu, L., Liu, K.: Recurrent convolutional neural networks for text classification. Proc. AAAI **333**, 2267–2273 (2015)
23. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of WWW, pp. 675–684. ACM (2011)
24. Mikolov, T., Chen, K., Corrado, G.: Efficient estimation of word representations in vector space. Comput. Sci. (2013)
25. Nal, K., Grefenstette, E., Blunsom, P.: A convolutional neural network for modelling sentences. In: Proceedings of ACL, pp. 655–665 (2014)
26. Quoc, L., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of ICML, pp. 1188–1196 (2014)
27. Tseng, S., Fogg, B.: Credibility and computing technology. Commun. ACM **42**(5), 39–44 (1999)

28. Popat, K., Mukherjee, S., Strötgen, J.: Where the truth lies: explaining the credibility of emerging claims on the web and social media. In: Proceedings of WWW, pp. 1003–1012 (2017)
29. Viviani, M., Pasi, G.: Credibility in social media: opinions, news, and health information—a survey. Wiley Interdiscip. Rev. Data Min. Knowl. Disc. **7**(5) (2017)
30. Kwon, S., Cha, S., Jung, K.: Rumor detection over varying time windows. PLoS ONE **12**(1), e0168344 (2017)
31. Choi, Y., Seo, Y., Yoon, S.: E-WOM messaging on social media: social ties, temporal distance, and message concreteness. Internet Res. **27**(3), 495–505 (2017)
32. Kapferer, J.: Rumors: Uses, Interpretation and Necessity. Routledge, London (2017)
33. Tandoc, J., Lim, Z., Ling, R.: Defining "fake news" a typology of scholarly definitions. Digit. J. **6**(2), 137–153 (2018)
34. Zhang, A., Ranganathan, A., Metz, S.: A structured response to misinformation: defining and annotating credibility indicators in news articles. In: Proceedings of WWW, pp. 603–612 (2018)
35. Ruchansky, N., Seo, S., Liu, Y.: CSI: a hybrid deep model for fake news detection. In: Proceedings of CIKM, pp. 797–806. ACM (2017)
36. Li, H., Fei, G., Wang, S.: Bimodal distribution and co-bursting in review spam detection. In: Proceedings of WWW, pp. 1063–1072 (2017)