Detecting Deceptive Review Spam via Attention-Based Neural Networks

Xuepeng Wang[1,2], Kang Liu[1](✉), and Jun Zhao[1,2]

[1] National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China
{xpwang,kliu,jzhao}@nlpr.ia.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. In recent years, the influence of deceptive review spam has further strengthened in purchasing decisions, election choices and product design. Detecting deceptive review spam has attracted more and more researchers. Existing work makes utmost efforts to explore effective linguistic and behavioral features, and utilizes the off-the-shelf classification algorithms to detect spam. But the models are usually compromised training results on the whole datasets. They failed to distinguish whether a review is linguistically suspicious or behaviorally suspicious or both. In this paper, we propose an attention-based neural networks to detect deceptive review spam by distinguishingly using linguistic and behavioral features. Experimental results on real commercial public datasets show the effectiveness of our model over the state-of-the-art methods.

1 Introduction

As increasingly used by customs and businesses, online reviews have formed a booming market. There emerge a large number of websites which provide online review services, such as Amazon, Yelp and TripAdvisor. Positive reviews often mean profits and fame for business and individuals [21]. It has been reported that positive rating scores help the restaurants on Yelp to sell out more products and earn revenue increasing [2,22]. As a result, driven by great commercial profit, more and more business owner begin to hire people to write deceptive positive reviews to promote their own products, and/or post fake negative reviews to discredit their competitors. Such individuals are called review spammers, and the fake reviews are defined as deceptive review spam [10,21]. It is urgent to detect deceptive review spam to maintain the trust of the review host websites.

The earliest academic investigations were carried by Jidal and Liu [10]. They studied 5.8 million reviews and 2.14 million reviewers from Amazon. A large number of duplicate reviews were found indicating that review spam was widespread. Several types of features were proposed and logistic regression was used for model building [10]. The majority of followed work takes it as a binary classification task. The researchers have made utmost efforts to explore effective features to indicate the review spam. For example, Unigram, POS and other linguistic features were explored by Ott et al. [28], Li et al. [19] and Hai et al. [8].

X. Huang et al. (Eds.): NLPCC 2017, LNAI 10619, pp. 866–876, 2018.
https://doi.org/10.1007/978-3-319-73618-1_76

Activity window, extremity of rating and other behavioral features were investigated and applied by Mukherjee et al. [27] and Rayana and Akoglu [29].

So far, previous work has proposed lots of effective approaches. However, researchers mainly focus on feature engineering and just apply the off-the-shelf classification algorithms to detect spam. But exploiting more effective algorithms or models is also significant for this task. Most of the review spam detecting models are usually compromised training results on the whole datasets, over the linguistic and behavioral features. But for the real commercial reviews on the Yelp website, some deceptive reviews are linguistically suspicious, some are behaviorally suspicious[1]. For linguistically suspicious review spam, the behavioral features which seem normal are actually noises for the detection models. But the learnt weight matrices of the traditional detection models are fixed for all the reviews in the datasets. They can not make a special identification for each review. So there needs to find a new way to further distinguishingly utilize the linguistic and behavioral features.

In this paper, we propose an attention-based neural networks by dynamically learning weights for linguistic and behavioral features for each training example. It can learn to distinguish whether each of the review spam is linguistically suspicious or behaviorally suspicious or both. More specifically, we take several effective behavioral features, which were exploited in previous work, as the inputs of the MLP hidden layer in our model. Then we get the behavioral feature vectors from the outputs of the MLP. We employ a convolutional neural network (CNN) to exact the linguistic features of a review, and take the outputs of the CNN as the linguistic feature vectors. Next, we take the behavioral feature vectors as the target hidden states and the linguistic feature vectors as source hidden states. Then an attention function is applied to calculate the score, which indicates how behaviorally suspicious a review is in the given linguistic environment. As well when the linguistic feature vectors are the target hidden states and the behavioral feature vectors are source hidden states, the attention function can also calculate how linguistically suspicious a review is in the given behavioral environment. Then the features vectors are tuned by the calculated scores. We concatenate the outputs of attention layer (the tuned feature vectors) with the original feature vectors. At last, the concatenated vectors go through a softmax layer to make predictions.

In summary, the contributions of this work are as follows:

- In stead of focusing on feature engineering as the most previous work did, we turn to find a more effective algorithm to tackle the deceptive review spam detecting task.
- We proposed an attention-based model neural networks by distinguishingly utilizing the linguistic and behavioral features for detecting each review spam. It learns dynamic weights for each training example. Compared with previous models, it can learn that whether a deceptive review is linguistically suspicious or behaviorally suspicious or both.

[1] https://www.yelp-support.com/article/What-is-Yelp-s-recommendation-software?.

– The experiments carried on the real commercial public datasets show that, the proposed model preforms more effectively than the state-of-the-art work, in both hotel and restaurant domains.

2 Related Work

Detecting review spam is a more difficult task than detecting other forms of spam, such as email spam [3], web search engine spam [7], blog spam [14] and tagging spam [15]. The deceptive review spam detection problem was firstly explored by Jindal and Liu [10]. They analysed 5.8 million reviews and 2.14 reviewers from the popular Amazon.com. They showed how widespread the problem of fake reviews was. Then they built their own dataset, and simply use near-duplicate reviews as examples of deceptive reviews. Several linguistic and behavioral features were proposed and logistic regression was applied for detection. Most followed work has made major efforts to discover suspicious clues and design effective features.

The Work Exploiting Linguistic Features. The first dataset of gold-standard deceptive review spam was released by Ott et al. [28] with employing crowd-sourcing through the Amazon Mechanical Turk. At the same time, they investigate the effectiveness of psychological and linguistic clues on identifying review spam. Several writing features were explored by Harris [9]. Then they applied several human- and machine-based assessment methods on the features. Feng et al. [5] focused on the syntactic stylometry in the review spam problem. Li et al. [19] was interesting exploring the general difference of language usage between deceptive and truthful reviews. Moreover, Li et al. [18] investigated the positive-unlabeled learning problem with unigrams and bigrams features. Kim et al. [13] analysed the semantic frame features in the deceptive review texts.

The Work Exploiting Behavioral Features. The reviewers' rating behavioral features were investigated by Lim et al. [20]. Jindal et al. [11] found several unusual review patterns which can represent suspicious behaviors of reviews. Li et al. [16] proposed a two-view semi-supervised method based on behavioral features. Feng et al. [6] focused on describing the distributions of reviewer's unusual behaviors. Xie et al. [34] applied the abnormal temporal patterns of reviewers to detect singleton reviews at resellerratings.com. Mukherjee et al. [24] studied a principal method to model the spamicity of reviewers. The behavioral feature of review co-occurrence was found by Fei et al. [4] in review bursts. By analysing the review at Dianping.com, Li et al. [17] found the temporal and spatial patterns in reviewers' footprints. [12] also investigated the temporal features of the reviews at Yelp websites. Moreover, the experiments carried by Mukherjee et al. [27] proved that reviewers' behavioral features are more effective than reviews' linguistic features on the realistic commercial datasets. Wang et al. [31] investigated the reviewers' behaviors in the online store review graph. Akoglu et al. [1]

exploit the network effect among reviewers and products. There is also some work that detected review spam by combining using the linguistic and behavioral features. Mukherjee et al. [26] proved the effectiveness of the combination of linguistic features and behavioral features. Besides, Rayana and Akoglu [29] utilized lots of clues from review text, reviewers' behaviors and the review graph structure to make a collective review spam detection.

The Work Detecting Review Spammers. The previous work referred above are mainly focusing on detecting review spam. There were also some work exploring detecting the review spammers. Wang et al. [32] identified online store review spammers via social graph. Another work [25] researched the group spamming activity. This work was the first attempt to solve the problem of review spammers from a group collaboration between multiple spammers. In this paper, we focus on detecting deceptive review spam by utilizing linguistic and behavioral features.

3 The Proposed Model

In this section, we further explain our attention-based neural networks in detail as shown in Fig. 1. As we referred in Sect. 1, most of the previous work focuses on exploiting effective features and just applies the off-the-shelf classification models to detect spam. Although the model can learn to identify the deceptive reviews, the trained models are usually compromised results over the whole

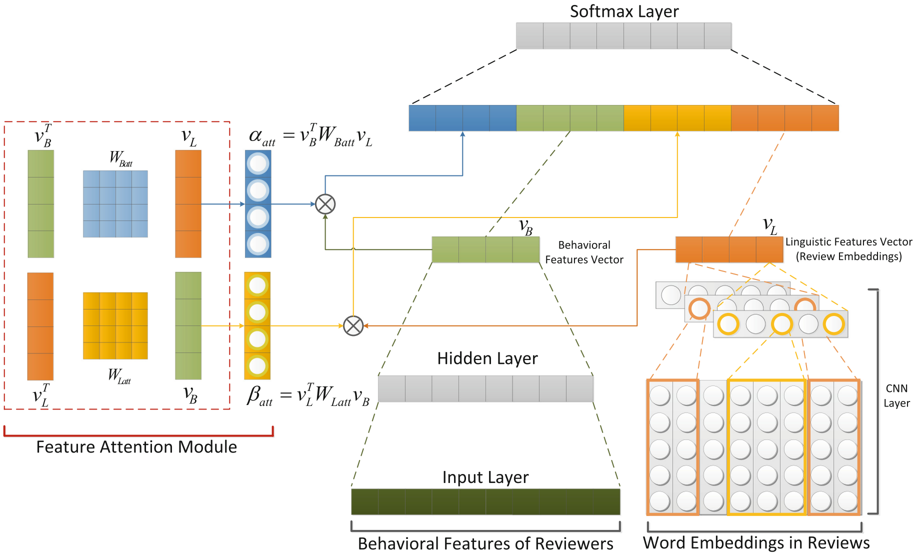
Softmax Layer
v_B^T
W_{Batt}
v_L
$\alpha_{att} = v_B^T W_{Batt} v_L$
v_B
Behavioral Features Vector
v_L
Linguistic Features Vector (Review Embeddings)
CNN Layer
Hidden Layer
Input Layer
v_L^T
W_{Latt}
v_B
$\beta_{att} = v_L^T W_{Latt} v_B$
Feature Attention Module
Behavioral Features of Reviewers
Word Embeddings in Reviews

Fig. 1. Illustration of our model.

datasets. It can figure out whether a review is deceptively suspicious, but it can not distinguish whether the review is linguistically suspicious or behaviorally suspicious. We find that some spammers on the website post reviews without any elaborate disguise, we can identify them by linguistic features. For example, there are lots of exclamation sentences contained in the negative spam for defaming [29]. However, some crafty spammers are good at writing plausible reviews with abundant experiences [27]. We have to figure out the suspicious behaviors in their activities. So we propose a novel attention-based neural networks. Compared with the previous work, it can learn dynamic weights for each review in the datasets, and further distinguish the suspiciousness category of the review spam by the feature attention mechanism.

3.1 The Feature Extraction Module

As shown in Fig. 1, we employ a MLP layer to extract behavioral feature vectors v_B from the inputs of effective behavioral features F_B. The output of the MLP layer is calculated as

$$v_B = \tanh\left(W_B F_B + b_B\right), \tag{1}$$

where $W_B \in \mathbb{R}^{D_B \times D_o}$, D_B is the dimension of the behavior feature inputs, D_o is the dimension of the MLP's outputs.

To extract the linguistic feature vectors, we adopt a convolutional neural network with word embeddings $e\left(w_i\right) \in \mathbb{R}^{Dw}$. Compared with the discrete manual features used in previous work and the RNN model, Ren and Zhang [30] have proved that the CNN can capture complex global semantic information and perform more effectivelys. We set n filter weight matrices $\widehat{W} = \{W_1, W_2, \ldots W_n\}$. Then we get the linguistic feature vectors v_L from the outputs of each filter utilizing a max pooling layer.

3.2 The Feature Attention Module

As shown in Fig. 1, we construct a feature attention module to learn how linguistically suspicious the spam is in the given behavioral environment, and how behavioral suspicious the spam is in the given linguistically environment. During model training we calculate the behavioral attention score α_{att} of the review spam by

$$\alpha_{att} = v_B^T W_{Batt} v_L, \tag{2}$$

where the behavioral attention matrix $W_{Batt} \in \mathbb{R}^{D_B \times D_L}$, D_L is the dimension of the linguistic feature vectors v_L. Here v_L is the source vectors and v_B is the target vectors. Then the linguistic attention score β_{att} is calculated as

$$\beta_{att} = v_L^T W_{Latt} v_B, \tag{3}$$

where the linguistic attention matrix $W_{Latt} \in \mathbb{R}^{D_L \times D_B}$. For the non-spam review, we also set two attention matrix W'_{Batt} and W'_{Latt}, and calculate the attention score same as the Eqs. 2 and 3.

Next we calculate the weighted feature vectors as

$$v'_B = \alpha_{att} v_B, \tag{4}$$

$$v'_L = \beta_{att} v_L, \tag{5}$$

Then the concatenation of the weighted feature vectors and the feature vectors is taken as the inputs of the softmax layer.

$$v = [v'_B : v_B : v'_L : v_L] \tag{6}$$

$$o = W_{sft} v + b_{sft}, \tag{7}$$

where $W_{sft} \in \mathbb{R}^{2*(D_B+D_l) \times D_{sft}}$, D_{sft} is the output dimension of the linear layer in softmax layer. The category prediction probability is calculated as

$$p\left(c_i \mid \theta\right) = \frac{\exp\left(o_i\right)}{\sum_{j=1}^{n_o} \exp\left(o_j\right)}, \tag{8}$$

where c_i is the prediction category, n_o is the number of categories, $\theta = [W_B, b_B, \widehat{W}, W_{Batt}, W_{Latt}, W'_{Batt}, W'_{Latt}, W_{sft}, b_{sft}]$. Finally, our training objective is to minimize the cross-entropy loss over plus a l_2-regularization term,

$$\mathcal{L}\left(\theta\right) = -\sum_{i=1}^{N} \log\left(c_i \mid \theta\right) + \frac{\lambda}{2} \left\|\theta\right\|^2 \tag{9}$$

We use Adam algorithm to minimize the loss function in Eq. 9. We initial all the matrix and vector parameters with uniform samples in $(\sqrt{6\left(r+c\right)}, \sqrt{6\left(r+c\right)})$, where r and c are the numbers of rows and columns of the matrices. For the word embeddings, we initial them with the vectors of 200-dimensions which are trained on Yelp review datasets [27], using the CBOW model proposed by Mikolov et al. [23].

When the model identifies the review spam in the testing datasets, we take the maximum conditional probabilities respectively calculated through W_{Batt}, W_{Latt}, W'_{Batt} and W'_{Latt} as the prediction labels.

4 Experiments

4.1 Datasets and Evaluation Metrics

Datasets: To evaluate the effectiveness of our model, we conduct the publicly released datasets which contain the realistic commercial reviews from the Yelp website. The datasets were widely used in the work of Mukherjee et al. [26], Mukherjee et al. [27], Rayana and Akoglu [29] and Wang et al. [33]. There are also other publicly available datasets for experiments. But some of them [10,20,34] are human labelled, and have been proved not to be reliable by Ott et al. [28]. Some of them [28] are generated by crowd sourcing, which have been proved not fully reflecting the realistic characteristics of the commercial review spam by Mukherjee et al. [27]. The statistics of the Yelp datasets used in this paper are listed in Table 1.

Table 1. Yelp labeled dataset statistics.

Domain	Hotel	Restaurant
Fake	802	8368
Non-fake	4876	50149
% fake	14.1%	14.3%
# reviews	5678	58517
# reviewers	5124	35593

Evaluation Metrics: We select precision (P), recall (R), F1-Score (F1) and accuracy (A) as metrics.

4.2 Our Model v.s. The State-of-the-Arts Work

In this paper, we compare our attention-based neural networks with the state-of-the-arts work to test the effectiveness. One of the compared work is presented by Mukherjee et al. [26]. Mukherjee et al. [26] analysed the reviews at the Yelp websites and proposed eight effective statistical behavioral features (e.g., the Activity Window, the Percentage of Positive Reviews). They also proved that the bigram is more effective than other previous linguistic features (e.g., POS, Deep Syntax and Information Gain) in detecting the realistic commercial deceptive review spam. Then they applied SVM and naïve Bayes respectively on the behavioral and linguistic features, and got the best performance with SVM. Another compared work is accomplished by Wang et al. [33]. To collectively utilize the global information in the review system, they proposed eleven asymmetric relations between reviewers and products. Then they learnt the representations of reviews by the tensor decomposition algorithm in a low dimension feature space. They proved that the leant representations are more effective than the traditional statistic features. In fact, their representations (i.e. the concatenation of reviewer embeddings and product embeddings) can be regarded as a kind of behavioral feature vectors. They also took the bigram as the linguistic features. Same with Mukherjee et al. [26], Wang et al. [33] applied the SVM on their learnt behavioral feature vectors and linguistic features to detect deceptive review spam. For fair experimental comparison, we apply our model respectively on the behavioral features proposed by Mukherjee et al. [26], and the behavioral features vectors learnt by Wang et al. [33]. For our model, we set the window size of the CNN filters to 2 for extracting linguistic features from bigram word embeddings. Besides, we set the number of convolution matrices to 30, D_B to 100, λ to $0.1E-6$. All the hyper-parameters are tuned by grid search on the development dataset.

The results of compared experiments are shown in Table 2. We first compare our attention-based neural networks with the work of Mukherjee et al. [26] on the same eight statistical behavioral features and bigrams (Table 2(a,b) rows 1, 3). Our model results in around 2.5% improvement in F1 and 2.1% improvement in A at the hotel domain, and results in around 1.8% improvement in F1 and

Table 2. SVM classification results across behavioral features linguistic features (bigrams here) by Mukherjee et al. [26], the classification results achieved by our model without attention mechanism using the features in Mukherjee et al. [26] (Our_model_noAtt_M), and the results achieved by our model with attention mechanism using the features in Mukherjee et al. [26] (Our_model_withAtt_M); the SVM classification results across bigrams and behavioral feature vectors learnt by Wang et al. [33], the classification results achieved by our model without attention mechanism using the features in Wang et al. [33] (Our_model_noAtt_W), the classification results achieved by our model with attention mechanism using the features in Wang et al. [33] (Our_model_withAtt_W). All the results here are 5-fold CV results. Both the training and testing use balanced data (50:50). Improvements of our model are statistically significant with $p < 0.005$ based on paired t-test.

Features	P	R	F1	A	
Mukherjee et al. [26]	82.8	86.9	84.8	85.1	1
Our_model_noAtt_M	85.4	86.5	86.0	85.9	2
Our_model_withAtt_M	86.3	88.5	**87.3**	**87.2**	3
Wang et al. [33]	84.2	89.9	87.0	86.5	4
Our_model_noAtt_W	86.8	88.5	87.6	87.5	5
Our_model_withAtt_W	88.1	89.7	**88.9**	**88.8**	6

(a) Hotel

P	R	F1	A	
84.5	87.8	86.1	86.5	1
85.9	87.5	86.7	86.6	2
87.4	88.4	**87.9**	**87.8**	3
86.8	91.8	89.2	89.9	4
88.9	91.3	90.1	90.0	5
89.4	93.0	**91.2**	**91.0**	6

(b) Restaurant

1.3% improvement in A at the restaurant domain. These results show that, compared with directly applying the off-the-shelf classification algorithm on the features, our model make a more effective performance with the bi-directional attention mechanism, to distinguish the suspicious type of review spam. Then we compared our model to the work of Wang et al. [33] on their learnt behavioral feature vectors and bigrams (Table 2(a,b) rows 4, 6). Our model results in around 1.9% improvement in F1 and 2.3% improvement in A at the hotel domain, and results in around 2.0% improvement in F1 and 1.1% improvement in A at the restaurant domain. It proves that our model is more effective than the method in Wang et al. [33] as well. This is probably because of that the feature attention module can learn how behavioral suspicious each review spam is, when given the corresponding linguistic features, and vice versa.

4.3 The Effectiveness of the Feature Attention Module

To further evaluate the effectiveness of our feature attention module, we compared our model with attention module to that without attention module (Table 2(a,b) rows 2, 3, 5, 6). When we move out the attention module, our model performs slightly better than Mukherjee et al. [26] and Wang et al. [33] in some domain metrics. Specifically, it performs 0.1% better in A at the restaurant domain (Table 2(b) rows 1, 2). And it performs 0.6% better in F1 at the hotel domain (Table 2(a) rows 4, 5). But some improvements are relatively obvious. For example, it performs 1.2% better in F1 at the hotel domain (Table 2(a) rows 1, 2). This indicates that the model only with the MLP and CNN module can hardly do a robust performance. When we add the attention module in our model, the experimental results show that the attention mechanism help to perform 1.5% better in F1 and 1.2% better in A at both domain in average (Table 2(a,b) rows 2, 3, 5, 6). It proves that the attention module actually helps to identify deceptive review spam by distinguishing the suspicious type of review.

- [**REVIEW EXAMPLE**][2] AMAZING!!!! I've been to quite a few gastronomy driven restaurants...some have been mind blowing...some not. But Alinea was beyond! ...But how can you really give justice t this whole presentation? I CAN'T!!! ...I'd have to say this was the most amazing place I've ever eaten at! Our waiters ranged from normal to pretentious...but whatever...the food was amazing. The presentation..amazing...the decor (especially when you walk into the hallway from the street.... amazing...the attention to detail....amazing! I would definitely be back because this place is freakin AMAZING!!!!!
- Behavioral Attention Score: 0.1537; Behavioral Features: RL = 0.78, RC = 0.35, AW = 0.41, PR = 0.60; Linguistic Attention Score: 0.9727.

4.4 The Attention Spam Example in Datasets

To further present the effect of our attention-based neural networks, we list an attention deceptive review spam example during testing our model with bidirectional attention mechanism on the features used by Mukherjee et al. [26] at the restaurant domain. As shown in the above review example, the behavioral features seem very normal, for example the behavioral feature Review Length with the normalization value 0.78 indicates that it is a long review. Mukherjee et al. [27] found that the average number of words per non-spam review is relatively longer than that of spam review. But when we turn to the context of the review, we find that it contains lots of exclamation points and all-capital words. It describes the restaurant in a strongly promoting mood. So the review is very suspicious on linguistic features.

In this review example, the behavioral features are noises for the traditional detection models. Inversely, there are other deceptive reviews which are behaviorally suspicious and seem normal in linguistic features. The linguistic features are noises for them. But the learnt weight matrices of the traditional detection models are fixed for all the reviews in the datasets. The models are actually compromised training results. They fail to make a special identification for each review. So our model adopts the feature attention module to learn dynamic weights (attention score α_{att} and β_{att}) for linguistic and behavioral features for each training example. Indeed, as shown in [**REVIEW EXAMPLE**] the linguistic attention score learnt by our model is larger than the behavioral attention score. It indicates that, for this linguistically suspicious review spam, our model has dynamically paid more attention to the linguistic features than the behavioral features.

[2] An attention deceptive review spam example during testing our model with bidirectional attention mechanism on the features used by Mukherjee et al. [26] at the restaurant domain.

5 Conclusion

We introduced a neural network framework with attention mechanism for detecting deceptive review spam. The attention mechanism can learn dynamic weights for linguistic and behavioral features for each training sample. The proposed model not only achieves state-of-the-art performance, but also shows the importance of linguistic and behavioral features according to the weights provided by the attention mechanism. Extensive experiments show that our model outperforms all baseline models and achieves precision, recall, and F-value. In the future, we will explore more effective methods for the task.

Acknowledgments. This work was supported by the Natural Science Foundation of China (No. 61533018), the National Basic Research Program of China (No. 2014CB340503) and the National Natural Science Foundation of China (No. 61502493).

References

1. Akoglu, L., Chandy, R., Faloutsos, C.: Opinion fraud detection in online reviews by network effects. In: ICWSM 2013, vol. 13, pp. 2–11 (2013)
2. Anderson, M., Magruder, J.: Learning from the crowd: regression discontinuity estimates of the effects of an online review database. Econ. J. **122**, 957–989 (2012)
3. Carreras, X., Marquez, L.: Boosting trees for anti-spam email filtering. arXiv preprint cs/0109015 (2001)
4. Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R.: Exploiting burstiness in reviews for review spammer detection. In: ICWSM 2013. Citeseer (2013)
5. Feng, S., Banerjee, R., Choi, Y.: Syntactic stylometry for deception detection. In: ACL 2012, pp. 171–175. Association for Computational Linguistics (2012)
6. Feng, S., Xing, L., Gogar, A., Choi, Y.: Distributional footprints of deceptive product reviews. In: ICWSM 2012 (2012)
7. Gyongyi, Z., Garcia-Molina, H.: Web spam taxonomy. In: AIRWeb 2005 (2005)
8. Hai, Z., Zhao, P., Cheng, P., Yang, P., Li, X.L., Li, G.: Deceptive review spam detection via exploiting task relatedness and unlabeled data. In: EMNLP 2016, pp. 1817–1826 (2016)
9. Harris, C.: Detecting deceptive opinion spam using human computation. In: AAAI 2012 (2012)
10. Jindal, N., Liu, B.: Opinion spam and analysis. In: WSDM, pp. 219–230. ACM (2008)
11. Jindal, N., Liu, B., Lim, E.P.: Finding unusual review patterns using unexpected rules. In: CIKM 2010, pp. 1549–1552. ACM (2010)
12. Santosh, K.C., Mukherjee, A.: On the temporal dynamics of opinion spamming: case studies on yelp. In: WWW, pp. 369–379 (2016)
13. Kim, S., Chang, H., Lee, S., Yu, M., Kang, J.: Deep semantic frame-based deceptive opinion spam analysis. In: CIKM 2015, pp. 1131–1140. ACM (2015)
14. Kolari, P., Java, A., Finin, T., Oates, T., Joshi, A.: Detecting spam blogs: a machine learning approach. In: AAAI 2006, vol. 21, p. 1351. AAAI Press/MIT Press, Menlo Park/Cambridge (2006)
15. Koutrika, G., Effendi, F.A., Gyöngyi, Z., Heymann, P., Garcia-Molina, H.: Combating spam in tagging systems. In: AIRWeb 2007, pp. 57–64. ACM (2007)

16. Li, F., Huang, M., Yang, Y., Zhu, X.: Learning to identify review spam. In: IJCAI 2011, vol. 22, p. 2488 (2011)
17. Li, H., Chen, Z., Mukherjee, A., Liu, B., Shao, J.: Analyzing and detecting opinion spam on a large-scale dataset via temporal and spatial patterns. In: AAAI (2015)
18. Li, H., Liu, B., Mukherjee, A., Shao, J.: Spotting fake reviews using positive-unlabeled learning. Computación y Sistemas **18**, 467–475 (2014)
19. Li, J., Ott, M., Cardie, C., Hovy, E.: Towards a general rule for identifying deceptive opinion spam. In: ACL 2014, pp. 1566–1576. Association for Computational Linguistics (2014)
20. Lim, E.P., Nguyen, V.A., Jindal, N., Liu, B., Lauw, H.W.: Detecting product review spammers using rating behaviors. In: Proceedings of 19th CIKM, pp. 939–948. ACM (2010)
21. Liu, B.: Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press, Cambridge (2015)
22. Luca, M.: Reviews, reputation, and revenue: the case of Yelp.com. In: Harvard Business School NOM Unit Working Paper (12–016) (2011)
23. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS 2013, pp. 3111–3119 (2013)
24. Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., Ghosh, R.: Spotting opinion spammers using behavioral footprints. In: SIGKDD. ACM (2013)
25. Mukherjee, A., Liu, B., Glance, N.: Spotting fake reviewer groups in consumer reviews. In: WWW, pp. 191–200. ACM (2012)
26. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.: Fake review detection: classification and analysis of real and pseudo reviews. Technical report UIC-CS-2013-03 (2013)
27. Mukherjee, A., Venkataraman, V., Liu, B., Glance, N.S.: What yelp fake review filter might be doing? In: ICWSM (2013)
28. Ott, M., Choi, Y., Cardie, C., Hancock, T.J.: Finding deceptive opinion spam by any stretch of the imagination. In: ACL 2011, pp. 309–319 (2011)
29. Rayana, S., Akoglu, L.: Collective opinion spam detection: bridging review networks and metadata. In: SIGKDD 2015, pp. 985–994. ACM (2015)
30. Ren, Y., Zhang, Y.: Deceptive opinion spam detection using neural network. In: COLING 2016, pp. 140–150. The COLING 2016 Organizing Committee (2016)
31. Wang, G., Xie, S., Liu, B., Yu, P.S.: Review graph based online store review spammer detection. In: ICDM, pp. 1242–1247. IEEE (2011)
32. Wang, G., Xie, S., Liu, B., Yu, P.S.: Identify online store review spammers via social review graph. TIST **3**(4), 61 (2012)
33. Wang, X., Liu, K., He, S., Zhao, J.: Learning to represent review with tensor decomposition for spam detection. In: EMNLP 2016. Association for Computational Linguistics (2016)
34. Xie, S., Wang, G., Lin, S., Yu, P.S.: Review spam detection via temporal pattern discovery. In: KDD 2012, pp. 823–831. ACM (2012)