IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 5, 2019, accepted March 25, 2019, date of publication April 1, 2019, date of current version April 13, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2908495

Learning to Detect Deceptive Opinion Spam: A Survey

YAFENG REN[1,2] AND DONGHONG JI[2,3]

[1]Laboratory of Language Engineering and Computing, Guangdong University of Foreign Studies, Guangzhou 510420, China
[2]Collaborative Innovation Center for Language Research and Services, Guangdong University of Foreign Studies, Guangzhou 510420, China
[3]School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China

Corresponding author: Yafeng Ren (renyafeng@whu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61702121, in part by the National Key Research and Development Program of China under Grant 2017YFC1200500, in part by the Science and Technology Project of Guangzhou under Grant 201704030002, and in part by the Bidding Project of GDUFS Laboratory of Language Engineering and Computing under Grant LEC2018ZBKT004.

ABSTRACT With the development of e-commerce, more and more users begin to post reviews or comments about the quality of products on the internet. Meanwhile, people usually read previous reviews before purchasing online products. However, people are frequently deceived by deceptive opinion spam, which is usually used for promoting the products or damaging their reputations because of economic benefit. Deceptive opinion spam can mislead people's purchase behavior, so the techniques of detecting deceptive opinion spam have extensively been researched in past ten years. In particular, some work based on deep learning has been investigated in last three years for the task. However, there still lack a survey, which can systematically analyze and summarize the previous techniques. To address this issue, this paper first introduces the task of deceptive opinion spam detection. Then, we summarize the existing dataset resources and their construction methods. Third, existing methods are analyzed from two aspects: traditional statistical methods and neural network models. Finally, we give some future directions of the task.

INDEX TERMS Deceptive opinion spam, deceptive review, machine learning, feature engineering, natural language processing, deep learning.

I. INTRODUCTION

In the Web 2.0 era, users can automatically post reviews or comments on e-commerce websites. These user-generated contents are of great value for both consumers and cooperations [1]–[3]. On one hand, consumers can capture some information about a product or service by reading these reviews before purchasing it. On the other hand, the business organizations can adjust their products and marketing strategies by analyzing these reviews. People easily get influenced by reviews information when making a purchasing decision, so positive reviews can bring huge economic benefit and fame for business organizations and individuals. This promotes the generation of deceptive opinion spam (also called deceptive review) [4].

Jindal and Liu first introduced the concept of opinion spam [5], and described three types of reviews:

- *Type 1 (Untruthful Opinions):* Those that deliberately mislead readers by giving undeserving positive reviews to target objects in order to promote the objects or by giving malicious negative reviews to objects in order to damage their reputations.
- *Type 2 (Reviews on Brands Only):* Those that do not comment on the products in reviews specifically for the products but only the brands, the manufacturers or the sellers of the products. These reviews are considered as spam because they are not targeted at specific products.
- *Type 3 (Non-Reviews):* Those that are non-reviews, which contains two main sub-types: advertisements and other irrelevant reviews with no opinions.

The second and third types of opinion spams are called as disruptive opinion spam [5]. These two types of spam pose little threat to people, because human easily identify them. The first type of opinion spam is called as deceptive opinion spam [5]. This type of spam is confusing and difficult to identify, and has attracted more and more research interests

The associate editor coordinating the review of this manuscript and approving it for publication was Tony Thomas.

in recent years. In this paper, we focus on the analysis and discussion about deceptive opinion spam.

Deceptive opinion spam is a type of reviews with fictitious opinions, deliberately written to sound authentic [5], [6]. Two reviews are shown as follows:

- *Review 1:* I have stayed at many hotels travelling for both business and pleasure and I can honestly say that the James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The location is perfect within walking distance to all the great sights and restaurants. Highly recommend to both business travellers and couples.
- *Review 2:* My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful!! The area of the hotel is great, since I love to shop I couldn't ask for more!! We will definitely be back to Chicago and we will for sure be back to the James Chicago.

These two reviews are both from the firstly public dataset in the domain of deceptive opinion spam [6]. The first is a truthful review, and the second is deceptive opinion spam. By only reading two reviews, we can know that it is difficult for human readers to distinguish them. Previous researchers also organized three volunteers to manually annotate the reviews, and only achieved about 60% accuracy [6]. Meanwhile, they found that the accuracy of truthful reviews was significantly higher than that of deceptive opinion spam, and human tended to misjudge deceptive opinion spam as a truthful one.

Deceptive opinion spam is widely distributed [7]. According to statistics, deceptive opinion spam accounts for about 2–6% at Orbitz, Proceline, Expedia and Tripadivisor [8], [9]. Unlike the above sites, there can be a large proportion (up to 14–20%) at Yelp [10]. A large amount of deceptive opinion spam increases people's distrust for online reviews. It is very necessary to design some effective models to identify them automatically [11]. In the past ten years, many methods have been proposed to detect deceptive opinion spam. These methods explored the task from different viewpoints, and promoted the development of this field. However, there are still many challenges for the task. For example, the difficulty in constructing datasets and the poor domain adaption ability of algorithms still have not been solved. In this paper, we summarize and analyze the existing methods and data resources, and give some suggestions for future directions.

At the beginning of 2015, there are two work to preliminarily summarize the previous techniques for identifying deceptive opinion spam [11], [12]. However, these two work has several shortcomings. First, they lack the techniques related to neural networks, especially the rapidly deep learning techniques developed in recent years. Second, they do not systematically analyze and summarize the existing data resources. Third, they fail to guide future research directions. To address these issues, this paper makes a systematic analysis and summary for previous work and data resources, and gives constructive suggestions for future directions.

In this paper, we first introduce the task of deceptive opinion spam detection. Then we summarize the existing data resources and their construction methods. It is followed by the existing methods, containing two types of models, namely, traditional statistical models and neural network models. Finally, we give some future research directions.

II. TASK DEFINITION

The task can be divided into two subtasks. The first is the detection of deceptive opinion spam. The second is the detection of deceptive opinion spammer.

A. DECEPTIVE OPINION SPAM DETECTION

Deceptive opinion spam detection is a text-oriented detection task, and aims to classify a review as spam or non-spam by using the review content itself. It is usually modeled as a binary classification task. Jindal and Liu first proposed the concept of deceptive opinion spam [5], [13], and used a supervised learning method to identify deceptive opinion spam on Amazon reviews. Wu *et al.* proposed to detect deceptive opinion spam based on popularity rankings [14]. Ott *et al.* presented three supervised learning algorithms to detect deceptive opinion spam by integrating the knowledge of psycholinguistic and linguistics [6]. Feng *et al.* verified the connection between deceptive opinion spam and abnormal distributions [15]. Li *et al.* explored generalized approaches for identifying online deceptive opinion spam [16].

B. DECEPTIVE OPINION SPAMMER DETECTION

For deceptive opinion spam detection, analyzing the language phenomenon from the review text only can not effectively identify deceptive opinion spam. Intuitively, increasing the behavior analysis from the review author can improve the detection performance. Deceptive opinion spammer detection is a comprehensive analysis for the review content and the review author. Some preliminary work includes identifying multiple userid of the same user [17], identifying deceptive opinion spammer through behavioral footprint [15], [18], [19], and analyzing the relationship between deceptive reviews and publishers based on graph method [20], [21]. Typically, Yu *et al.* proposed to identify deceptive groups from their conversations [22].

III. DATASETS

This section first summarizes the existing datasets, as shown in Table 1. These datasets can be categorized into four categories based on different construction methods: rule-based method, human-based method, filtering algorithm based method and AMT-based method. In the following section, we will analyze these four types of datasets in details.

A. RULES

The datasets 1–3 of Table 1 are constructed by rules-based methods. Based on Amazon reviews, Jindal and Liu (2008)

TABLE 1. Statistical information of existing datasets.

Method	Num	Source	Volume	Domain	Reference
Rules	1	Amazon	5.8M (reviews), 2.15M (users)	Books, Music, DVD, mProducts	[5]
	2	Amazon	6,819 (reviews), 4,811 (users)	Books	[23]
	3	TripAdvisor/Booking/Agoda	2,848 (reviews)	Hotels	[24]
Human	4	Epinions	6,000 (reviews)	Products	[25]
	5	TripAdvisor	3,000 (reviews)	Hotels	[26]
Filtering	6	Yelp	67,395 (reviews), 38,063 (users)	Hotels, Restaurants	[27]
	7	Yelp	359,052 (reviews), 160,25 (users)	Hotels, Restaurants	[28]
	8	Yelp	608,598 (reviews), 260,277 (users)	Hotels, Restaurants	[28]
	9	Dianping	9,765 (reviews), 9,067 (users)	Restaurants	[29]
AMT	10	TripAdvisor	800 (reviews)	Hotels	[6]
	11	TripAdvisor	1,600 (reviews)	Hotels	[30]
	12	TripAdvisor	3,032 (reviews)	Hotels, Restaurants, Doctor	[16]

found that three types of repetitive or similar reviews were likely to be deceptive opinion spam [5]:

- Different userids on the same product;
- Same userids on different products;
- Different userids on different products.

In actual annotation, they used Jaccard distance to calculate the similarity of the review text for three kinds of repeated reviews, and the text with a similarity greater than 0.9 was labeled as deceptive opinion spam. Finally, Jindal and Liu obtained 55,000 deceptive opinion spam. Different from this method, Fornaciari and Poesio (2014) and Hammad *et al.* (2015) defined a set of rules to construct their datasets, respectively [23], [24].

The rules-based method is not dependent on manual annotation, and the annotation cost is relatively low. It is easy to construct a large number of annotation data, but there is a certain amount of noise. Taking the dataset of Jindal and Liu (2008) as example, because of the misoperation or network connection, there is a phenomenon that the same user has multiple evaluation for the same product with high probability. Therefore, the repeated reviews on the same product or the same user ID are not necessarily deceptive opinion spam, but they are directly labeled as deceptive opinion spam. So this annotation method is still need to discuss [5]. Gilbert *et al.* analyzed about 1 million reviews from Amazon [31], and found that 10–15% reviews were similar to earlier reviews of the product. This shows that some users tend to refer to or directly copy previous reviews when writing a new review.

B. HUMAN

The datasets 4 and 5 of Table 1 are constructed by human-based method. Li *et al.* (2011) summarized 30 rules to identify deceptive opinion spam [25], and annotated the reviews of websites by human. Specifically, they solicited three volunteers (undergraduates) to annotate deceptive opinion spam. Each student independently labeled a review, and determined whether the review was spam or not. To reduce the extent to which the individual human judges were biased, they used the majority voting rule to predict "deceptive" when at least two out of three human judges believed a review to be deceptive. Finally, they obtained the dateset 4. The dataset totally contains 6,000 reviews in which 1,389 reviews are labeled as deceptive opinion spam. Based on similar annotation method, Ren *et al.* (2014) constructed the dataset 5 [26], containing totally 3,000 reviews in which 712 reviews were labeled as deceptive opinion spam.

Although manual annotation is based on certain criteria, it mainly depends on the subjective judgment of human. Because of the strong deception of deceptive opinion spam, the accuracy of artificial recognition is low [30], so there are still a number of mislabeled instances in this type of datasets.

C. FILTERING ALGORITHMS

The datasets 6–9 of Table 1 are constructed by filtering algorithms from Yelp and Dianping, respectively. The filtering algorithm is highly reliable, but these algorithms are all confidential. Mukherjee *et al.* carried out a series of experiments on Yelp dataset [27], and tried to speculate on Yelp filtering algorithm. At the same time, they used lexical features and user behavior features to learn the classifier. Specifically, user behavior characteristics were obtained through the analysis of website publicity and internal data, such as user IP address, geographic location information, network and session log, mouse operation, click behavior and commentator's social behavior on the website. Following the method of Mukherjee *et al.* (2013), Rayana and Akoglu (2015) collected two datasets from Yelp, named as YelpNYC and YelpZip [28]. Li *et al.* also constructed a dataset by Dianping's filtering algorithm [29]. Statistical information shows that the reviews in this dataset contain 85.5 words on average while reviews in Yelp dataset have an average length of 130.6 words.

D. AMT

The datasets 10–12 of Table 1 are constructed by AMT (Amazon Mechanical Turk) based on crowdsourcing platform [6], [16], [30]. Crowdsourcing services can carry out massive data collection. Specifically, it defines the task in the network platform, and pays for online anonymous workers to complete the task. Humans can not precisely distinguish deceptive reviews from existing ones, but they can create deceptive opinion spam as one part of the dataset. Ott *et al.* first accomplished this work by AMT [6], and they set 400 tasks for 20 hotels, in which each hotel contained

20 tasks. Specific task is: if you are a hotel market department employee, for each positive review you wrote for the benefit or hotel development, you may get one dollar. Finally, they collected 400 deceptive opinion spam. Meanwhile, they obtained 400 truthful reviews from TripAdvisor on the same 20 Chicago hotels by removing the reviews on the basis of some constraints. Later, Ott *et al.* explored the problem of negative deceptive opinion spam, and created the first dataset of deceptive opinion spam with negative sentiment reviews by following the method of Ott *et al.* [30]. More recently, Li *et al.* (2014) extended these two datasets into multiple domains (Hotel, Restaurant, Doctor) based on crowdsourcing platform [16].

For the datasets from crowdsourcing platform, it can reflect linguistic and psychological characteristics of deceptive opinion spam to some extent, but the data distribution is different from the distribution in the real world.

E. PERFORMANCE EVALUATION

Deceptive opinion spam detection is usually modeled as a text classification task. The performance evaluation of the model can be divided into two categories according to the distribution of positive and negative instances. For a balanced dataset, the commonly used evaluation indexes include accuracy, precision, recall and F1 score. Accuracy is a comprehensive evaluation of the ability of the algorithm to predicted as positive and negative examples. Precision is an evaluation of the ability of the algorithm to correctly predict positive examples. Recall is an evaluation of the ability of the algorithm to find all positive examples. F1 score is the harmonic value between the precision and recall, and is predictive ability of comprehensive evaluation algorithm for positive examples. For a imbalanced dataset, the commonly used evaluation indexes include ROC curve and AUC (Area Under the receiver operating characteristics Curve).

IV. METHODS

Since Jindal and Liu proposed the concept of deceptive opinion spam [5], the researches have lasted for nearly ten years and a large number of methods have been proposed. In addition to detecting deceptive opinion spam from the review content, there is also some work that focuses on spammer detection from the user behavior. In this section, we will summarize and analyze the existing techniques from two aspects: traditional statistical models and neural network models. Note that traditional statistical models need a large amount of discrete features, while neural network models take the embeddings as input, which can be learned automatically. Because feature engineering is very crucial for traditional statistical models, we first introduce various feature constructions.

A. FEATURE ENGINEERING

Existing features can be divided into two categories: text features and behavior features. In this section, we will introduce and analyze these two types of features in details.

1) TEXT FEATURES

Text features include lexicon, grammar, semantic features from the review text and meta-data features about the review. We summarize the following 7 features:

a: BoW (BAG OF WORD)

BoW feature is extensively applied into various tasks of natural language processing [32]–[35]. BoW feature of a text is represented as a word or a continuous number of words in a text, and it is also called n-gram feature. For the detection of deceptive opinion spam, unigram, bigram and trigram are all commonly used features [5], [6], [16], [25], [30]. In some work, the word frequency is also used to be represented as one of the BoW features [13], [30]. Based on different datasets, BoW feature gives obviously different results. For example, it achieves 89.6% accuracy in AMT dataset based on crowdsourcing platform [6], while it only gives 67.8% accuracy on Yelp dataset [27].

b: POS (PART OF SPEECH)

The occurrence or frequency of each POS (Part of Speech) is used as POS feature. Previous studies from computational linguistics showed that the distribution of POS in a text was related to the text type [36], [37], in other words, different types of texts had certain degree of distinction in POS features. In single domain dataset, Ott *et al.* (2011) found that POS feature was not as effective as BoW feature [6]. However, Li *et al.* (2014) showed that POS feature obtained better robustness in cross-domain settings [16].

c: LIWC

LIWC (Linguistic Inquiry and Word Count) software is a popular text analysis tool [38], [39], and is used widely for analyzing the linguistic characteristics from multiple aspects [40]–[43]. It has been used to detect personality traits, study tutoring dynamics and analyze the deception. Ott *et al.* and Li *et al.* (2014) both applied LIWC features into deceptive opinion spam detection [6], [16], and found the performance was lower than that of n-gram features. By integrating LIWC and n-gram features, the performance could be improved.

In particular, LIWC counts and groups the number of instances of nearly 4,500 keywords into 80 psychologically meaningful dimensions, which can be divided into the following four categories [6]:

1) *Linguistic Features:* Functional aspect of a text, e.g., the average number of words per sentence, the rate of misspelling, swearing, etc.
2) *Psychological Features:* Includes all social, emotional, cognitive, perceptual and biological processes, as well as anything related to the time and space.
3) *Personal Features:* Any references to work, leisure, money, religion, etc.
4) *Spoken Features:* Primarily filler and agreement words.

d: STYLOMETRIC

This type of feature mainly contains character-based and word-based lexical features or syntactic features [44].

Lexical features usually include the number of upper case characters and average word length, showing the types of words and characters that a writer tends to use. Syntactic features include the features like the amount of punctuations or the number of function words such as a, the, and of, representing the writing style of a reviewer.

e: SEMANTIC

These features deal with the underlying meaning or concepts of words. Lau *et al.* built a semantic language model for identifying untruthful reviews [7]. Li *et al.* proposed to learn the semantic representation of a text [45], and showed that semantic features could achieve better robustness than n-gram, LIWC and POS features in cross-domain settings. Besides, Kim *et al.* introduced a methodology with frame-based semantic features based on FrameNet [46], which was proved to effectively improve the classification performance in their experiments.

f: DEEP LINGUISTIC FEATURES

Chen *et al.* (2015) introduced two types of deep level linguistic features [47]. The first type was derived from a shallow discourse parser trained on PDTB (Penn Discourse Treebank), which could capture inter-sentence information. The second type was based on the relation between sentiment analysis and spam detection.

g: META-DATA FEATURES

These features contain the information about a review rather than the information on the review text [25], [27], including the review's length, date, time, rating, reviewer id, review id, store id or feedback. Meta-data features have shown to be beneficial in opinion spam detection. Strange or anomalous reviews can be identified using meta-data features, and once a reviewer has been identified as someone who writes spam, it is easy to label all reviews associated with this reviewer as spam. However, these features may not be available in many data sources, which thus limit their utility for the task.

2) BEHAVIOR FEATURES

Behavior features are statistical characteristics of a user's review behavior and his reviews. Behavior features can be extracted from a user's current review and his historical reviews. Some formula and symbol settings are first defined to better introduce the behavior features [19]. Specifically, r represents a review, A represents a set of all users, and a represents a user. $MaxRev(a)$ represents the maximum number of the reviews for a user a, and R_a denotes all reviews which user a posts. Specific features are summarized as follows:

a: MAXIMUM CONTENT SIMILARITY

Similar reviews for different products from the same reviewer has been shown to be a strong indication of a spammer [27]. Cosine similarity is used to measure the content similarity.

$$f_{CS} = max_{r_i, r_j \in R_a, i<j} cos(r_i, r_j) \tag{1}$$

b: MAXIMUM NUMBER OF REVIEWS

It was observed that about 75% of spammers wrote more than 5 reviews on any given day [27]. Therefore, taking into account the number of reviews wrote by a user per day could be beneficial for detecting spammers, since 90% of normal reviewers never created more than one review in one day.

$$f_{MNR} = \frac{MaxRev(a)}{max_{a \in A}(MaxRev(a))} \tag{2}$$

c: BURST CHARACTERISTICS

It was observed that most spammers usually had a short registration time and had a sudden release of reviews. Mukherjee *et al.* (2012) extracted the features of reviewing burstiness by defining activity window [19]. When the time between the latest release time $L(a)$ and the initial release time $F(a)$ was less than a certain threshold ($\tau = 28$), the user might be a spammer.

$$F_{BST}(a) = \begin{cases} 0, L(a) - F(a) > \tau \\ 1, \quad otherwise \end{cases} \tag{3}$$

d: PERCENTAGE OF THE FIRST REVIEW

As consumers tend to read early reviews, and the earlier the spam is released, the greater in impacts consumers [48]. The percentage of the first review collects the percentage of all reviews released by a user as the first review of the corresponding product.

$$f_{RFR} = \frac{|\{r \in R_a : r \; is \; first \; review\}|}{|R_a|} \tag{4}$$

e: REPEATED REVIEWS

Submitting multiple repetitive reviews on the same product is considered to be an abnormal behavior. However, the setting of this feature needs to be discussed. Jindal and Liu (2008) pointed out that the case that same user submitted same reviews for many times should not be regarded as deceptive opinion spam because of the network connection problem or operation errors [5].

f: EXTREME RATING BEHAVIOR

The highest or lowest score is considered as extreme rating, and it is possible that the users praise or vilify products intentionally. For the five-star rating system, scoring one star or five stars is an extreme rating behavior.

g: PERCENTAGE OF POSITIVE REVIEWS

Approximately 85% of spammers write more than 80% of their reviews as positive reviews, thus a high percentage of positive reviews might be an indicator of an untrustworthy reviewer [27].

h: REVIEWER DEVIATION

It is observed that spammers' ratings tend to deviate from the average review rating at a far higher rate than normal reviewers [48], thus identifying user rating deviations is helpful for the detection of dishonest reviewers.

TABLE 2. Comparison of supervised methods in previous work. In the "Dataset" column, A–B means that A represents the construction method, and B denotes the domain.

Approach	Features	Performance metric	Score	Dataset	Reference
SVM	LIWC+Bigrams	Accuracy	89.8%	AMT-TripAdvisor	[6]
SVM	Bigrams	Accuracy	89.6%	AMT-TripAdvisor	[6]
SVM	Deep syntax+unigram	Accuracy	91.2%	AMT-TripAdvisor	[15]
SVM	Behavioral features+ Bigrams	Accuracy	86.1%	Filtering-Yelp	[19]
SVM	n-gram features	Accuracy	86%	AMT-TripAdvisor	[30]
SVM	Stylometric features	F-measure	84%	AMT-TripAdvisor	[44]
SVM	Unigram	Accuracy	83.21%	AMT-TripAdvisor	[49]
LR	Review and reviewer features	AUC	78%	Rules-Amazon	[5]
LR	Text features	AUC	63%	Rules-Amazon	[5]
NB	Review and reviewer features	F-measure	0.9959	Rules-TripAdvisor/Booking/Agoda	[24]
NB	Review and reviewer features	F-score	0.631	Human-Epinions	[25]
NB	Text and behavioral features	F-score	0.686	Human-TripAdvisor	[26]
SAGE	LIWC+POS+Unigram	Accuracy	65%	AMT-TripAdvisor	[16]

i: CHARACTERISTICS OF EARLY REVIEWS

Whether a user's reviews belong to the early reviews or not reflects the behavior characteristics of the user to a certain extent. In order to mislead the consumers, deceptive opinion spam is usually contained in the early reviews of products by spammers.

j: REVIEW LENGTH

For all reviews of a reviewer, the average review length may be an important indication, since about 80% of spammers write longer reviews which are more than 135 words, while an average review length of 92% of reliable reviewers is more than 200 words.

B. TRADITIONAL STATISTICAL MODELS

From the viewpoint of machine learning, traditional statistical models can be divided into supervised learning, unsupervised learning and semi-supervised learning.

1) SUPERVISED LEARNING

Supervised learning method regards the task as a binary classification problem, and uses the labeled data to learn a classifier, then predict whether a review is spam or not. Previous researches using supervised learning method are summarized in Table 2, which mainly adopts three classifiers: Support Vector Machine (SVM), Logistic Regression (LR) and Naive Bayes (NB).

a: SVM

SVM can map feature vectors into high-dimensional space, so as to establish the maximum interval hyperplane, to maximize the distance between different classes [50]. SVM is very effective in solving a small number of instances with nonlinear and high dimensional features. Based on AMT datasets from crowdsourcing platform and Yelp datasets, Ott *et al.* used SVM to identify deceptive opinion spam by integrating n-gram, stylometric and user behavior features, and achieved 83–89.8% accuracy [6], [16], [27], [44]. By integrating deep syntax feature from CFG (Context Free Grammer) parer trees, the accuracy could be improved to 91.2% [15].

b: NB

NB has few parameters because of feature independent assumption, and its learning process is insensitive to data sparseness. Although the features in reality do not fully satisfy conditional independent assumption, NB can still give good classification performance. Li *et al.* (2011) used NB and co-training mechanism to identify deceptive opinion spam by integrating the text and behavior features [25], and achieved 63.1% F1 score in Epinions.com. El-Halees *et al.* proposed to identify deceptive opinion spam in Arabic [24], and used NB to achieve strong performance on tripadvisor.com, booking.com, and agoda.ae.

c: LR

LR is a generalized linear regression model, in which independent variable and logistic probability are linear. LR can be well applied into the problem that independent variable is numeric and dependent variable is discontinuous variable. Jindal and Liu (2008) used this model to explore deceptive opinion spam detection on Amazon dataset [5], [13], and achieved 63–78% AUC value based on different features. Experimental results also showed that LR could give better performance than SVM and NB on Amazon dataset. Ren *et al.* (2015) proposed to find deceptive opinion spam from the viewpoint of correcting the mislabeled instances [51]. Based on LR classifiers, they first divided a dataset into several subsets, and constructed a classifier set of each subset and selected the best one to evaluate the whole dataset. Error variables were defined to compute the probability that the instances have been mislabeled. The mislabeled instances were corrected based on majority and non-objection schemes. Results showed the performance could be improved by correcting the mislabeled instances.

d: SAGM

SAGM (Sparse Additive Generative Model) is a generative Bayesian method [52], which can be regarded as the combination of a topic model and a generalized additive model [53], [54], and the Laplasse prior is used to deal with the sparse distribution of the topical words. Li *et al.* believed

that SAGM could collect a variety of factors (e.g. different fields, experiences or inexperienced, positive or negative), and could predict whether a review was spam or not [16]. Experimental results showed that SAGM was better than SVM in dealing with cross-domain deceptive opinion spam detection.

2) SEMI-SUPERVISED LEARNING

Semi-supervised learning uses a small amount of labeled data and a large number of unlabeled data to learn the classifier. Previous work is mainly divided into two classes: Co-training method and PU learning.

a: CO-TRAINING METHOD

This model is also called two-views method [25], in which each view represents different types of features. This method uses the compatibility and complementarity of multi-views to classify the unlabeled instances and extend the training set. Based on a small number of labeled instances, Li *et al.* (2011) proposed to train two classifiers by using two groups of different features [25], and predicted the categories of unlabeled instances. After extending the labeled dataset, the final classifier was trained by integrating two types of features, which was used to predict the category of unlabeled instances in the test set. In their experiments, they used NB as the classifier. Experimental results showed that the performance of two-views method was better than that of the classifier only trained in a small amount of labeled dataset. Ren *et al.* (2014) proposed to integrate the knowledge from computational linguistics and psycholinguistics [26]. Then supervised-learning method was developed to evaluate the performance of different feature modelings, and to select the best mixed features. Finally, co-training and tri-training methods were designed to exploit a large amount of unlabeled data.

b: PU (POSITIVE UNLABELED) LEARNING

PU learning model built a final classifier based on a small number of positive instances and a large number of unlabeled instances [55], [56]. The basic idea is to find a set of reliable negative instances from the unlabeled data, and then to learn a classifier using EM (Expectation Maximization) or SVM.

PU leaning has been successfully applied into many text classification tasks [57]–[60]. Some researches have investigated PU learning for identifying deceptive opinion spam [29], [49], [61]–[64]. Typically, Li *et al.* proposed a collective positive and unlabeled learning to improve PU learning [29], where they added the positive instances predicted in unlabeled instances into the positive instance set to well train the classifier. Based on some truthful reviews and a large number of unlabeled reviews, Ren *et al.* explored a novel PU learning method [49], called mixing population and individual nature PU learning method, for identifying deceptive opinion spam. First, some reliable negative examples were identified from the unlabeled dataset. Second, some representative positive examples and negative examples were generated by integrating latent dirichlet allocation and K-means. Third, all spam examples (easily mislabeled) were clustered into different groups based on dirichlet process mixture model, and two schemes (population nature and individual nature) were mixed to determine the category label of spy examples. Finally, multiple kernel learning was used for building the final classifier. Results showed that the proposed method outperformed previous PU learning methods.

In addition to the above two types of semi-supervised methods, Hai *et al.* proposed to exploit the relatedness of multiple opinion spam detection tasks and available unlabeled data to address the scarcity of labeled opinion spam data [65]. They first used a multi-task learning method based on LR, to boost the learning for a task by sharing the knowledge contained in the training signals of other related tasks. To leverage the unlabeled data, they introduced a graph Laplacian regularizer into each base model. Then they proposed a SMTL-LLR (Semi-supervised Multi-Task Learning method via Laplacian regularized Logistic Regression) model to improve the detection performance. Besides, Rout *et al.* explained how semi-supervised learning methods were used for identifying deceptive opinion spam [66].

3) UNSUPERVISED LEARNING

Because of the difficulty of constructing accurately the labeled datasets, supervised learning is not always applicable. Unsupervised learning provides a solution, as it doesn't require labeled data.

Raymond *et al.* (2012) explored an unsupervised model [7] and developed SLM (Semantic Language Model) to identify deceptive opinion spam. This model followed a assumption that if semantic content of a review was close to another one, it was likely that the two reviews were duplicates and thus were labeled as deceptive opinion spam. In their experiments, they built a dataset based on Amazon reviews. They first identified the reviews as spam with a cosine similarity above some threshold, and then manually confirmed them. Conversely, the reviews that did not have a cosine similarity above a certain threshold with any other reviews were kept as truthful reviews and not manually reviewed. Finally, the dataset contained 54,618 reviews, in which 6% were labeled as spam. SLM was used to assign a spamminess score to each instance. Using this score, they were able to achieve 0.9987 AUC score, which outperformed SVM on the same dataset with 0.5571 AUC. They argued that experimental results showed that SLM was effective for this task, and that unsupervised methods could achieve a high detection accuracy for duplicate spam reviews.

Ren *et al.* took full account of the psychological state of the author of deceptive opinion spam [67], and thought that there must be some differences on language structure and emotional polarity between deceptive opinion spam and truthful reviews. Then, they defined the features related to the review text and use genetic algorithm for feature selection. Finally, they used two clustering algorithm to identify deceptive opinion spam. Sedighi *et al.* explored a decision tree method to identify deceptive opinion spam from trustworthy ones [68].

They utilized unsupervised representation learning along with traditional feature selection methods to select appropriate features and evaluate them. The proposed model could take data correlation into consideration to select suitable features.

C. NEURAL NETWORK MODELS

Neural network models have extensively been applied into many NLP tasks [69]–[77]. Compared with traditional statistical models, neural network models have several advantages. First, neural network models have great non-linear fitting capabilities due to the comparable depth of network architecture. Second, neural network models learn intrinsic features from raw data fully automatically even without any substantial human efforts to carefully construct patterns. Third, neural network models with the well-trained word embeddings can effectively capture the syntactic structures of a text or semantic relations between context words in a more scalable way. The most representative neural models contain CNN (Convolutional Neural network), RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory), and Recursive Neural Network, etc. For deceptive opinion spam detection, some preliminary work has been done by using neural network models. These work largely focuses on CNN and RNN.

a: CNN

CNN is a special type of feed-forward neural network originally employed in the field of computer vision. Convolutional layers in CNN play the role of feature extractor, which extracts local features as they restrict the receptive fields of the hidden layers to be local. Such a characteristic is useful for the classification tasks in NLP. Li *et al.* (2015) took word vector as inputs, and used CNN to learn the semantic representation, which was directly as the features to identify deceptive opinion spam [45]. Experimental results verified the effectiveness of CNN, which was more robust in cross-domain settings. Meanwhile, CNN achieved better results than LSTM in mixed dataset. Zhao *et al.* optimized CNN by embedding the word order characteristics in its convolutional layer and pooling layer, which made CNN more suitable for deceptive opinion spam detection [78].

Recently, Wang *et al.* proposed an attention-based neural network model to identify deceptive opinion spam by dynamically learning weights for linguistic and behavioral features for each training example. In their model, they used a CNN to exact the linguistic features [79]. Later, they proposed a neural network model to detect review spam from the viewpoint of the cold-start problem [80]. They adopted CNN to learn to represent a new reviewer's review with jointly embedded textual and behavioral information. More recently, Zhang *et al.* (2018) proposed DRI-RCNN (Deceptive Review Identification by Recurrent Convolutional Neural Network) to identify deceptive opinion spam by using word contexts and deep learning [81]. The basic idea was that deceptive opinion spam and truthful reviews were written by authors without and with real experience, respectively, so the review authors should have different contextual description on target objectives.

b: RNN

RNN uses internal ''memory'' to process a sequence of inputs, and is widely used for processing sequential information. Theoretically, RNN can make use of the information in arbitrarily long sequences, but in practice, standard RNN is limited to looking back only a few steps due to the vanishing gradient or exploding gradient problem. Researchers have developed more sophisticated types of RNN to avoid the shortcomings of standard RNN such as Bidirectional RNN, deep bidirectional RNN, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), etc.

Ren and Zhang (2016) empirically explored a neural network model to learn document-level representation for detecting deceptive opinion spam [82]. Given a document, the model learned first sentence representations with a CNN, which were combined using a gated RNN to model discourse information and to yield a document vector. Finally, the document representation was used directly as features to identify deceptive opinion spam. Results on in-domain and cross-domain settings showed that the proposed model outperformed the traditional discrete models. More recently, Wang *et al.* investigated LSTM to detect spammers [83]. They used a real case of fake review in Taiwan, and compared the analytical results of the current study with results of previous literature. They found that LSTM was more effective than SVM for detecting fake reviews.

c: OTHERS

Wang *et al.* (2016) proposed to learn the representations of reviews in a data-driven manner instead of heavily relying on expert's knowledge to identify deceptive opinion spam [84]. A tensor network model was built on the relation generated from two patterns, and a tensor factorization algorithm was used to learn the vector representations of reviews and products. Afterwards, they concatenated the review text, the embedding of a review and the reviewed product as final representation of a review. Then, a classifier was applied to detect deceptive opinion spam. Experimental results showed that the proposed method learned more robust review representations.

Previous techniques failed to consider comprehensive features of entities such as review, reviewer, product and group of reviewers simultaneously. Noekhah *et al.* (2018) proposed a novel Multi-Iteration Network Structure which considered the most effective features along with inter- and intra-relationships between entities on Amazon. Experimental results proved the proposed model could improve the detection performance by reducing the false noise [85]. Besides, Aghakhani *et al.* adopted Generative Adversarial Networks (GANs) for identifying deceptive reviews [86].

TABLE 3. Experimental results of different features on AMT dataset.

Features	Accuracy (%)
Human (Majority)	58.1
POS	73.0
LIWC	76.8
Unigram	88.4
Bigram	89.6
LIWC+Bigram	89.8
Shallow syntax + Unigram	88.6
Deep syntax	90.4
Deep syntax + Unigram	91.2
Compatibility features + Syntax + Unigram	91.3
Bigram + LIWC + Sentiment	88.6
Bigram + LIWC + Syntactic	89.1
Bigram + LIWC + Syntactic + Discourse Parsing	89.5

Dong *et al.* presented an end-to-end trainable unified model to leverage the appealing properties from Autoencoder and random forest [87].

D. COMPARISON AND ANALYSIS

We show experimental results of different features on AMT dataset constructed by corwdsourcing platform. As shown in Table 3, Ott *et al.* (2011) used SVM with different features to identify deceptive opinion spam [6], and achieved 73.0% accuracy by using POS features, and 76.8% accuracy with psycholinguistic features (LIWC), respectively. By integrating LIWC and Bigram, the model achieved 89.8% accuracy. Later, Feng *et al.* (2012) investigated syntactic stylometry for deceptive opinion spam detection [15]. They demonstrated that the features driven from Context Free Grammar (CFG) parse trees consistently improved the detection performance. Specifically, deep syntactic features achieved 90.4% accuracy. When syntactic features were combined with Unigram features, the performance of 91.2% accuracy could be obtained. Feng and Hirst (2013) proposed using profile compatibility to identify deceptive opinion spam [88]. They defined two types of compatibility between product profiles, and designed a methodology to tackle them by extracting aspects and associated descriptions from reviews. By integrating the profile alignment compatibility features with the features (Deep syntactic + Unigram) of Feng *et al.* (2012), they achieved 91.3% accuracy. More recently, Chen *et al.* (2015) introduced two types of deep level linguistic features for identifying deceptive opinion spam. The features of the first type were derived from a shallow discourse parser trained on Penn Discourse Treebank (PDTB), which could capture inter-sentence information. The second type was based on the relationship between sentiment analysis and spam detection. They used a 5-fold nested cross validation for evaluation. By integrating sentiment features into Bigram and LIWC features, the model achieved 88.6% accuracy. By integrating syntactic features into Bigram and LIWC features, the model achieved 89.1% accuracy. The model achieved 89.5% accuracy by integrating four types of features. This showed the effectiveness of deep level linguistic features for the task.

We also show experimental results of different features on Yelp dataset constructed by filtering algorithm. As shown in Table 4, Bigram features yield 67.8% accuracy on Yelp dataset. By adding POS Bigrams features, the performance can be improved 1–2% in accuracy, and achieve the best performance (68.1% accuracy). Only using POS Unigram, the detection accuracy is very low (55.6% accuracy). Based on the above observation, we can say that LIWC makes little contribution compared with Unigram and Bigram features.

TABLE 4. Experimental results of different features on Yelp dataset, which is constructed by filtering algorithm.

Features	Accuracy (%)
Unigram	66.9
Bigram	67.8
Unigram + Bigram	67.8
Bigram + LIWC	67.8
POS Unigram	55.6
Bigram+ POS Bigram	68.1
Bigram + Deep syntax	67.6

The datasets from the review websites have different data characteristics compared with the reviews based on crowdsourcing platform. Deceptive opinion spam and truthful reviews from Yelp tend to be consistent in language characteristics, while deceptive opinion spam from crowdsourcing platform and truthful reviews have great differences in language. Based on Table 3 and 4, we can know SVM with Unigram features on Yelp dataset only achieves 66.9% accuracy, which is far below the accuracy of 88.4% on AMT dataset constructed by crowdsourcing platform. This tells us that detecting real-life deceptive opinion spam (67% accuracy) in the commercial setting of Yelp is significantly harder than detecting crowdsourced deceptive opinion spam (90% accuracy). Mukherjee *et al.* (2013) analysed the reason from a psychological viewpoint [27], and pointed out that the Yelp website's deceptive opinion spam tried to make opinion spam as convincing as a truthful review. However, deceptive opinion spam from Amazon crowdsourcing platform in which the purpose of the Turker was to make money and they might lack a real purchase experience. For AMT dataset constructed by crowdsourcing platform, n-gram features give a high accuracy, indicating that deceptive opinion spam and truthful reviews are quite different in language structures by only imagining to make reviews.

V. DISCUSSIONS AND FUTURE DIRECTIONS

In previous sections, we present an overview of machine learning techniques that have been used in deceptive opinion spam detection. We can know that previous models largely focused on supervised learning techniques. However, supervised learning usually requires a labeled dataset, which can be difficult to acquire in the area of deceptive opinion spam detection. Based on the section of data resources, it can be observed that the commonly used datasets in previous studies are created by crowdsourcing platform, due to the difficulty of labeling. Evaluating classifiers based on these datasets can be problematic, as it has been observed that they are not necessarily representative of deceptive opinion spam in the real world.

Previous studies has laid a solid foundation for deceptive opinion spam detection. Looking forward to the future, the following research directions are worthy of attention and discussion.

(1) A serious problem in this task is lack of standard datasets. Previous literatures and experiments show that manual annotation of deceptive opinion spam is very difficult. Thus, most of previous researches use the datasets constructed by crowdsourcing platform, resulting in that the performance of proposed methods may not be measured accurately to some extent. Accordingly, the construction of gold datasets is still a problem that needs to be solved urgently.

(2) The domain adaption problem of the model needs to be effectively solved. Like other NLP tasks, in the field of deceptive opinion spam detection, it is faced with the problem of lacking annotation datasets, but how to apply the model trained in the source field to the target field is an important research direction. Previous work mainly focuses on a single field, while the study of domain adaption has not been carried out in depth. Li *et al.* (2014) trained the model in hotel domain, and tested in restaurant and doctor domains [16]. Experimental results showed that the accuracy and F1 score droped seriously compared with the performance in the same filed. So more in-depth exploration and research are needed for cross-domain deceptive opinion spam detection.

(3) Semi-supervised and unsupervised methods need to be explored more. Because large-scale datasets are difficult to obtain accurately, a meaningful research direction is how to effectively use a large number of unlabeled data in the real world. At present, semi-supervised methods used in this task mainly focus on Co-training and PU learning methods. The classification performance based on these algorithms is not satisfactory. So designing more effective semi-supervised methods is an important research direction. Taking full account of the difficulty of constructing a dataset for deceptive opinion spam, it is particularly important to design unsupervised learning algorithms to identify deceptive opinion spam. However, existing unsupervised methods largely adopt heuristic rules, so the exploration space of unsupervised methods is very large in this task.

(4) The neural network models need to be explored more in this task. Recently, neural network models have been used to learn semantic representation for NLP tasks [89]–[94], achieving highly competitive results. However, neural network models have not been well used for this task. Potential advantages of using neural networks for spam detection are two-fold. First, neural models use dense hidden layers or automatic combinations, which can capture complex global semantic information that is difficult to express using traditional discrete manual features. Second, neural networks take distributed word embeddings as inputs, which can be trained from a large-scale raw text, thus alleviating the sparsity of annotated data to some extent. For the task, most of existing methods focus on traditional statistical models, which need to design a large amount of discrete manual features. So designing effective neural network algorithms will be a heated research topic for this task.

REFERENCES

[1] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, nos. 1–2, pp. 1–135, 2008.
[2] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lect. Hum. Lang. Technol.*, vol. 5, no. 1, pp. 1–167, 2012.
[3] E. Fitzpatrick, J. Bachenko, and T. Fornaciari, *Automatic Detection of Verbal Deception*. San Rafael, CA, USA: Morgan & Claypool, 2015.
[4] N. Jindal and B. Liu, "Analyzing and detecting review spam," in *Proc. IEEE Int. Conf. Data Mining*, Oct. 2007, pp. 547–552.
[5] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proc. Int. Conf. Web Search Data Mining*, 2008, pp. 219–230.
[6] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics*, 2011, pp. 309–319.
[7] R. Y. K. Lau, S. Y. Liao, R. C. Kwok, K. Xu, Y. Xia, and Y. Li, "Text mining and probabilistic language modeling for online review spam detection," *ACM Trans. Manage. Inf. Syst.*, vol. 2, no. 4, pp. 1–30, Dec. 2011.
[8] M. Ott, C. Cardie, and J. Hancock, "Estimating the prevalence of deception in online review communities," in *Proc. Int. Conf. World Wide Web*, 2012, pp. 201–210.
[9] V. López, S. D. Río, J. M. Benítez, and F. Herrera, "Cost-sensitive linguistic fuzzy rule based classification systems under the mapreduce framework for imbalanced big data," *Fuzzy Sets Syst.*, vol. 258, pp. 5–38, Jan. 2015.
[10] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting burstiness in reviews for review spammer detection," in *Proc. 7th Int. AAAI Conf. Weblogs Social Media*, 2013, pp. 175–184.
[11] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. A. Najada, "Survey of review spam detection using machine learning techniques," *J. Big Data*, vol. 2, no. 1, p. 23, 2015.
[12] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey," *Expert Syst. Appl.*, vol. 42, no. 7, pp. 3634–3642, 2015.
[13] N. Jindal, B. Liu, and E. P. Lim, "Finding unusual review patterns using unexpected rules," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 1549–1552.
[14] G. Wu, D. Greene, B. Smyth, and P. Cunningham, "Distortion as a validation criterion in the identification of suspicious reviews," in *Proc. Workshop Social Media Anal.*, 2010, pp. 10–13.
[15] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2012, pp. 171–175.
[16] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a general rule for identifying deceptive opinion spam," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 1566–1576.
[17] T. Qian and B. Liu, "Identifying multiple userids of the same author," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1124–1135.
[18] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review spam detection via temporal pattern discovery," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2012, pp. 823–831.
[19] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *Proc. Int. Conf. World Wide Web*, 2012, pp. 191–200.
[20] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Identify online store review spammers via social review graph," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 1–61, 2012.
[21] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Review graph based online store review spammer detection," in *Proc. IEEE Int. Conf. Data Mining*, 2011, pp. 1242–1247.
[22] D. Yu, Y. Tyshchuk, H. Ji, and W. Wallace, "Detecting deceptive groups using conversations and network analysis," in *Proc. Annu. Meeting Assoc. Comput. Linguistics Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 857–866.
[23] T. Fornaciari and M. Poesio, "Identifying fake Amazon reviews as learning from crowds," in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2014, pp. 279–287.
[24] A. M. El-Halees and A. A. Hammad, "An approach for detecting spam in arabic opinion reviews," *Int. Arab J. Inf. Technol.*, to be published.
[25] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 2488–2493.

[26] Y. Ren, D. Ji, and L. Yin, "Deceptive reviews detection based on semi-supervised learning algorithm," *J. Sichuan Univ. (Eng. Sci. Ed.)*, vol. 46, no. 3, pp. 62–69, 2014.
[27] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "What yelp fake review filter might be doing?" in *Proc. 7th Int. AAAI Conf. Weblogs Social Media*, 2013, pp. 409–418.
[28] S. Rayana and L. Akoglu, "Collective opinion spam detection: Bridging review networks and metadata," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 94–985.
[29] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao, "Spotting fake reviews via collective positive-unlabeled learning," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 899–904.
[30] M. Ott, C. Cardie, and J. T. Hancock, "Negative deceptive opinion spam," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2013, pp. 497–501.
[31] E. Gilbert and K. Karahalios, "Understanding deja reviewers," in *Proc. ACM Conf. Comput. Supported Cooperat. Work*, 2010, pp. 225–228.
[32] Z. Wang, Z. Wu, R. Wang, and Y. Ren, "Twitter sarcasm detection exploiting a context-based model," in *Proc. Int. Conf. Web Inf. Syst. Eng.*, 2015, pp. 77–91.
[33] T. Qian, Y. Zhang, M. Zhang, Y. Ren, and D. Ji, "A transition-based model for joint segmentation, pos-tagging and normalization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1837–1846.
[34] Y. Ren, J. Deng, and D. Ji, "Twitter normalization via 1-to-n recovering," in *Proc. Int. Conf. Web Inf. Syst. Eng.*, 2016, pp. 19–34.
[35] L. Chen, B. Chen, Y. Ren, and D. Ji, "Long short-term memory RNN for biomedical named entity recognition," *BMC Bioinf.*, vol. 18, no. 1, pp. 462–493, 2017.
[36] P. Rayson, A. Wilson, and G. Leech, "Grammatical word class variation within the British National Corpus Sampler," *Lang. Comput.*. 2002, pp. 295–306.
[37] D. Biber and R. Quirk, "Longman grammar of spoken and written English," *Tesol Quart.*, vol. 34, no. 4, pp. 787–788, 1999.
[38] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The development and psychometric properties of LIWC2015," *Austin*, vol. 29, no. 11, pp. 1020–1025, 2007.
[39] A. Abbasi, Z. Zhang, D. Zimbra, H. Chen, and J. F. Nunamaker, "Detecting fake Websites: The contribution of statistical learning theory," *Mis Quart.*, vol. 34, no. 3, pp. 435–461, 2010.
[40] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," *J. Artif. Intell. Res.*, vol. 30, pp. 457–500, Nov. 2007.
[41] W. L. Cade, B. A. Lehman, and A. Olney, "An exploration of off topic conversation," in *Proc. Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 669–672.
[42] R. Mihalcea and C. Strapparava, "The lie detector: Explorations in the automatic recognition of deceptive language," in *Proc. Meeting Assoc. Comput. Linguistics Int. Joint Conf. Natural Lang. Process.*, 2009, pp. 309–312.
[43] A. Vrij, S. Mann, S. Kristen, and R. P. Fisher, "Cues to deception and ability to detect lies as a function of police interview styles," *Law Hum. Behav.*, vol. 31, no. 5, pp. 499–518, 2007.
[44] S. Shojaee, M. A. A. Murad, A. B. Azman, N. M. Sharef, and S. Nadali, "Detecting deceptive reviews using lexical and syntactic features," in *Proc. Int. Conf. Intell. Syst. Design Appl.*, 2014, pp. 53–58.
[45] L. Li, W. Ren, B. Qin, and T. Liu, "Learning document representation for deceptive opinion spam detection," in *Proc. 14th China Nat. Conf. Comput. Linguistics*, 2015, pp. 393–404.
[46] S. Kim, H. Chang, S. Lee, M. Yu, and J. Kang, "Deep semantic frame-based deceptive opinion spam analysis," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 1131–1140.
[47] C. Chen, H. Zhao, and Y. Yang, "Deceptive opinion spam detection using deep level linguistic features," in *Proc. 4th CCF Conf. Natural Lang. Process. Chin. Comput.*, 2015, pp. 465–474.
[48] E. P. Lim, V. A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 939–948.
[49] Y. Ren, D. Ji, H. Zhang, and L. Yin, "Deceptive reviews detection based on positive and unlabeled learning," *J. Comput. Res. Develop.*, vol. 52, no. 3, pp. 639–648, 2015.
[50] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[51] Y. Ren, J. I. Donghong, L. Yin, and H. Zhang, "Finding deceptive opinion spam by correcting the mislabeled instances," *Chin. J. Electron.*, vol. 24, no. 1, pp. 52–57, 2015.
[52] J. Eisenstein, A. Ahmed, and E. P. Xing, "Sparse additive generative models of text," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 1041–1048.
[53] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
[54] T. Hastie, *Generalized Additive Models*. London, U.K.: Chapman & Hall, 1990.
[55] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially supervised classification of text documents," in *Proc. 19th Int. Conf. Mach. Learn.*, 2002, pp. 387–394.
[56] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building text classifiers using positive and unlabeled examples," in *Proc. IEEE Int. Conf. Data Mining*, 2003, pp. 179–186.
[57] C. Elkan and K. Noto, "Learning classifiers from only positive and unlabeled data," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 213–220.
[58] X. Li, P. S. Yu, B. Liu, and S. K. Ng, "Positive unlabeled learning for data stream classification," in *Proc. SIAM Int. Conf. Data Mining*, 2009, pp. 257–268.
[59] H. Zhao, Z. Lu, and P. Poupart. (2015). "Self-adaptive hierarchical sentence model." [Online]. Available: https://arxiv.org/abs/1504.05070
[60] Y. Xiao, B. Liu, J. Yin, L. Cao, C. Zhang, and Z. Hao, "Similarity-based approach for positive and unlabelled learning," in *Proc. Int. Joint Conf. Artif. Intell.*, 2011, pp. 1577–1582.
[61] Y. Ren, D. Ji, and H. Zhang, "Positive unlabeled learning for deceptive reviews detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 201–210.
[62] D. H. Fusilier, R. Guzman-Cabrera, M. Montes-Y-Gómez, and P. Rosso, "Using pu-learning to detect deceptive opinion spam," in *Proc. 4th Workshop Comput. Approaches Subjectivity, Sentiment Social Media Anal.*, 2013, pp. 38–45.
[63] P. Rosso, "Detecting positive and negative deceptive opinions using PU-learning," *Inf. Process. Manage.*, vol. 51, no. 4, pp. 433–443, 2015.
[64] J. K. Rout, S. Singh, S. K. Jena, and S. Bakshi, "Deceptive review detection using labeled and unlabeled data," *Multimedia Tools Appl.*, vol. 76, no. 3, pp. 1–25, 2017.
[65] Z. Hai, P. Zhao, P. Cheng, P. Yang, X. L. Li, and G. Li, "Deceptive review spam detection via exploiting task relatedness and unlabeled data," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1817–1826.
[66] J. Rout, A. Dalmia, K. K. R. Choo, S. Bakshi, and S. Jena, "Revisiting semi-supervised learning for online deceptive review detection," *IEEE Access*, to be published.
[67] Y. Ren, L. Yin, and D. Ji, "Deceptive reviews detection based on language structure and sentiment polarity," *J. Frontiers Comput. Sci. Technol.*, vol. 8, no. 3, pp. 313–320, 2014.
[68] Z. Sedighi, H. Ebrahimpour-Komleh, and A. Bagheri, "RLOSD: Representation learning based opinion spam detection," in *Proc. Intell. Syst. Signal Process.*, Dec. 2018, pp. 74–80.
[69] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *J. Mach. Learn. Res.*, vol. 12 pp. 2493–2537, Aug. 2011.
[70] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.
[71] N. Kalchbrenner, E. Grefenstette, and P. Blunsom. (2014). "A convolutional neural network for modelling sentences." [Online]. Available: https://arxiv.org/abs/1404.2188
[72] Y. Jing, "Research of deceptive opinion spam recognition based on deep learning," East China Normal Univ., Shanghai, China, Tech. Rep., 2014.
[73] Y. Ren, Y. Zhang, M. Zhang, and D. Ji, "Context-sensitive Twitter sentiment classification using neural network," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 215–221.
[74] Y. Ren, R. Wang, and D. Ji, "A topic-enhanced word embedding for Twitter sentiment classification," *Inf. Sci.*, vol. 369, pp. 188–198, Nov. 2016.
[75] Q. Dou *et al.*, "Automatic detection of cerebral microbleeds from MR images via 3D convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1182–1195, May 2016.
[76] Y. Ren and D. Ji, "Neural networks for deceptive opinion spam detection: An empirical study," *Inf. Sci.*, vols. 385–386, pp. 213–224, Apr. 2017.
[77] D. Zhang, L. Zou, X. Zhou, and F. He, "Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer," *IEEE Access*, to be published.

[78] S. Zhao, Z. Xu, L. Liu, and M. Guo. (2017). "Towards accurate deceptive opinion spam detection based on word order-preserving CNN." [Online]. Available: https://arxiv.org/abs/1711.09181v2
[79] X. Wang, K. Liu, and J. Zhao, "Detecting deceptive review spam via attention-based neural networks," in *Proc. 6th Conf. Natural Lang. Process. Chin. Comput.*, 2017, pp. 866–876.
[80] X. Wang, K. Liu, and J. Zhan, "Handling cold-start problem in review spam detection by jointly embedding texts and behaviors," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 366–376.
[81] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, "DRI-RCNN: An approach to deceptive review identification using recurrent convolutional neural network," *Inf. Process. Manage.*, vol. 54, no. 4, pp. 576–592, 2018.
[82] Y. Ren and Y. Zhang, "Deceptive opinion spam detection using neural network," in *Proc. 26th Int. Conf. Comput. Linguistics*, 2016, pp. 140–150.
[83] C.-C. Wang, M.-Y. Day, C.-C. Chen, and J.-W. Liou, "Detecting spamming reviews using long short-term memory recurrent neural network framework," in *Proc. 2nd Int. Conf. E-Commerce, E-Bus. E-Government*, 2018, pp. 16–20.
[84] X. Wang, K. Liu, S. He, and J. Zhao, "Learning to represent review with tensor decomposition for spam detection," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 866–875.
[85] S. Noekhah, N. B. Salim, and N. H. Zakaria, "A novel model for opinion spam detection based on multi-iteration network structure," *Adv. Sci. Lett.*, vol. 24, no. 2, pp. 1437–1442, 2018.
[86] H. Aghakhani, A. Machiry, S. Nilizadeh, C. Kruegel, and G. Vigna. (2018). "Detecting deceptive reviews using generative adversarial networks." [Online]. Available: https://arxiv.org/abs/1805.10364v1
[87] M. Dong, L. Yao, X. Wang, B. Benatallah, C. Huang, and X. Ning, "Opinion fraud detection via neural autoencoder decision forest," *Pattern Recognit. Lett.*, to be published.
[88] V. W. Feng and G. Hirst, "Detecting deceptive opinions with profile compatibility," in *Proc. 6th Int. Joint Conf. Natural Lang. Process.*, 2013, pp. 338–346.
[89] K. Cho *et al.* (2014). "Learning phrase representations using rnn encoder-decoder for statistical machine translation." [Online]. Available: https://arxiv.org/abs/1406.1078
[90] M. Denil, A. Demiraj, N. Kalchbrenner, P. Blunsom, and N. de Freitas. (2014). "Modelling, visualising and summarising documents with a single convolutional neural network." [Online]. Available: https://arxiv.org/abs/1406.3830
[91] K. S. Tai, R. Socher, and C. D. Manning. (2015). "Improved semantic representations from tree-structured long short-term memory networks." [Online]. Available: https://arxiv.org/abs/1503.00075
[92] Y. Ren, Y. Zhang, M. Zhang, and D. Ji, "Improving Twitter sentiment classification using topic-enriched multi-prototype word embeddings," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 3038–3044.
[93] L. Fei, M. Zhang, G. Fu, and D. Ji, "A neural joint model for entity and relation extraction from biomedical text," *BMC Bioinf.*, vol. 18, no. 1, p. 198, 2017.
[94] Y. Ren, D. Ji, and H. Ren, "Context-augmented convolutional neural networks for Twitter sarcasm detection," *Neurocomputing*, vol. 308, pp. 1–7, Sep. 2018.

YAFENG REN received the Ph.D. degree with the Computer School, Wuhan University, China, 2015. He was a Postdoctoral Research Fellow with the Singapore University of Technology and Design from 2015 to 2016. He is currently an Associate Professor with the Guangdong University of Foreign Studies. He has been working on deceptive opinion spam detection over the past ten years, and has published ten related papers in journals and conferences, including AAAI, EMNLP, and COLING. His research interests include sentiment analysis and opinion mining, web mining, and bioinformatics.

DONGHONG JI received the B.S., M.S., and Ph.D. degrees from the Computer School, Wuhan University, China, in 1988, 1991, and 1995, respectively. He was a Postdoctoral Research Fellow with Tsinghua University from 1995 to 1998. From 1998 to 2008, he was a Researcher with the Institute Infocomm Research of Singapore. He is currently a Professor and a Ph.D. Supervisor with the Computer School, Wuhan University. His interests include natural language processing, machine learning, and data mining.