Hindawi
Complexity
Volume 2021, Article ID 9923374, 11 pages
https://doi.org/10.1155/2021/9923374

Research Article

Fast Detection of Deceptive Reviews by Combining the Time Series and Machine Learning

Minjuan Zhong,[1] **Zhenjin Li**,[2] **Shengzong Liu**,[1] **Bo Yang**,[1] **Rui Tan**,[3] **and Xilong Qu**[1]

[1]*School of Information Technology, Hunan University of Finance and Economics, Changsha 410205, China*
[2]*School of Foreign Language, Hunan University of Finance and Economics, Changsha 410205, China*
[3]*School of Information Management, Jiangxi University of Finance and Economics, Nanchang 330013, China*

Correspondence should be addressed to Zhenjin Li; zhenjinli@yeah.net and Xilong Qu; quxilong@sina.com

Received 17 March 2021; Revised 12 April 2021; Accepted 23 April 2021; Published 30 May 2021

Academic Editor: M. Irfan Uddin

With the rapid growth of online product reviews, many users refer to others' opinions before deciding to purchase any product. However, unfortunately, this fact has promoted the constant use of fake reviews, resulting in many wrong purchase decisions. The effective identification of deceptive reviews becomes a crucial yet challenging task in this research field. The existing supervised learning methods require a large number of labeled examples of deceptive and truthful opinions by domain experts, while the available unsupervised learning methods are inefficient because they depend on the features of reviewers to detect each fake review. Therefore, by focusing on the detection efficiency problem and the limitation of large amount of labeled examples dependence, in this paper, we proposed an effective semisupervised learning approach for detecting spam reviews. Firstly, a time series model of all the reviews of a product is constructed, and then the suspected time intervals are captured based on the burst review increases in these intervals. Secondly, a co-training two-view semisupervised learning algorithm was performed in each captured interval, in which linguistic cues, metadata, and user purchase behaviors were synthetically employed to classify the reviews and check whether they are spam ones or not. A series of numerical experiments on a real dataset acquired from Taobao.com have confirmed the effectiveness of the proposed model, not only reaping benefits in terms of time efficiency and high accuracy but also overcoming the shortcomings of supervised learning methods, which depend on large amounts of labeled examples. And a trade-off balance was obtained between accuracy and efficiency.

1. Introduction

The increasing popularity of e-commerce and rapid development of online shopping have resulted in large numbers of online reviews describing the perception of consumers on many goods and services [1]. Therefore, more and more consumers rely on online product reviews to assess the quality of a product or a service, which influences their purchasing decisions. Obviously, higher numbers of positive reviews induce consumers to purchase certain products, fortifying financial gains for manufactures, while negative reviews prompt consumers to look for alternatives, leading to financial losses. Driven by competition and vested interests, many vendors and retailers try to manipulate online reviews. For example, they tend to post deceptive reviews in an attempt to mislead potential consumers and make them take risky purchasing decisions. In the worst cases, they may employ many spammers or collective spammers to either post glamorized positive reviews with the aim to improve their product reputation or harmful negative reviews to suppress their competitors.

In fact, in recent years, online fraud, including the posting of fake reviews, has become a common phenomenon driven by profits. For instance, a report claimed that ~16% of the Yelp restaurant reviews are fraudulent [2]. In Tripadvisor, Orbitz, Priceline, and Expedia, 2–6% of the reviews

are fake. According to a public survey, the statement of Taobao, "ten stores and nine brushes," has become an open secret. In 2015, the team of Dainping banned more than 630,000 violating accounts and handled more than 19,000 merchants involving releasing more than 20 fake reviews. Obviously, the existence of such fake reviews, which mislead consumers and spread malicious information, has seriously affected the authenticity of consumer reviews, increased the risks of online shopping, and disrupted markets with unfair competition. It has been reported that such deceptive reviews can hardly be recognized by humans. Under such circumstances, it is challenging to detect whether a review is fake or not in real time and to automatically filter fake reviews to help online users evaluate products based on credible reviews rather than suspicious (or fake) ones.

Therefore, to effectively solve these problems, many different approaches have been proposed to identify the authenticity of customer reviews. Currently, there are mainly three techniques for deceptive review detection, supervised learning for labeled data, unsupervised learning for unlabeled data, and semisupervised learning for minimum labeled data [3]. Supervised learning is concerned with the detection of opinion spam as a classification process. Although its detection precision is higher than that of the unsupervised learning technique, an obvious disadvantage is that it lacks reliable labeled training datasets due to the time consumption and difficulty of manually marking thousands of spam reviews [4]. On one hand, normal users cannot effectively identify whether a review is fake or not, and there must be a certain number of false examples in a manually prelabeled dataset. On the other hand, there are often a large number of unlabeled examples. The unsupervised learning method tends to focus on opinion spammers or groups by relying on the behavioral features of reviewers. However, it is difficult to quickly identify each fake review by relying on the features of reviewers.

Therefore, the motivation of this paper includes two aspects. First, the outbreak of fake reviews has a burst characteristic. For example, there can suddenly be many reviewers for a certain brand on e-commerce websites. To seize more opportunities, some sellers hire spammers to post deceptive reviews. When spammers start to post fake reviews for a certain product, the number of reviews for this product significantly increases in a short period. Thus, capturing suspicious time intervals would be beneficial to both the high detection accuracy and time efficiency due to the narrowing of the detection range. Second, although there are fewer labeled examples in the real world, large numbers of unlabeled examples can be used to improve the learning performance. Based on the small numbers of labeled examples, is it possible to build a classifier to minimize the reliance on manually labeled training sets.

The contributions of this paper are as follows:

(1) By focusing on the recognition efficiency, a rapid detection method was proposed. For all the reviews of a certain product, a time series model was constructed based on their release time, and suspicious time intervals were captured from the burst increases in the number of reviews. As the detection interval of the reviews is narrowed, the efficiency of identifying deceptive reviews is greatly improved.

(2) Effectively solving the training set dependence problem: a co-training two-view semisupervised learning algorithm was applied to the reviews of the suspicious time intervals, where language clues of the content reviews, metadata features, and user purchase behaviors were synthetically employed to distinguish fake reviews from real ones.

(3) Extensive experiments have verified the effectiveness of the proposed model. Compared with Atefeh's work [5], higher detection accuracy was obtained without a large number of labeled training sets. In addition, the proposed method showed to be comparable with the method proposed by Zhang et al. [6]. The detection of the suspected intervals significantly improved the review detection efficiency, and a trade-off balance was obtained between the detection accuracy and efficiency.

The rest of this paper is organized as follows. Section 2 reviews the studies in the area of detecting deceptive reviews. In Section 3, we introduced the proposed method for the fast detection of fake reviews. The experimental settings and result analysis were illustrated in Section 4, and the conclusions and future prospects were mentioned in Section 5.

2. Related Works

This section briefly discusses the existing works for review spam detection. Depending on the types of the used data, Jitendra summarized a previous detection approach as supervised, unsupervised, and semisupervised learning [3].

2.1. Supervised Learning Techniques. Typically, opinion spam identification is mapped as a binary classification process [7]. Based on the positive and negative examples annotated by people, researchers extracted discriminant features and fed them into a bunch of traditional classification models, such as Naïve Bayes (NB), K-nearest neighbor (KNN), and support vector machine (SVM), to train classifiers. Then, they predicted unlabeled reviews to be fake or genuine [8]. In the text content, duplicate and near-duplicate reviews were considered to be important clues for the task of fake review detection [9, 10], as repeat reviews could maximize economics and time efficiency for the spammers. Jindal and Liu made a pioneering effort to solve the review spam problem. Their technique is based on comparing reviews to find duplicate ones and consider them as fake ones. They also trained classifiers to accurately recognize such reviews. Lin et al. categorized duplicate spam into three types, including repeat reviews from the same reviewer on different products, repeat reviews from different reviewers on the same product, and repeat reviews from different reviewers on several products.

However, such type of review spam has a limited scope as spammers normally change their own reviews to avoid being

detected by automated modules. Thus, other text-linguistic features, such as average review length, percentage of first-person pronouns, ratio of subjective or object words, and extreme sentiment tendencies, have also received much attention. Rupesh and Singh [11] fully considered the writing style features, such as the type of punctuation mark and part-of-speech (POS). On the basis of them, they applied sequential minimal optimization (SMO), decision tree, and Naïve Bayes algorithm for performing fake review identification. The experimental results showed good performance. Hafifiz et al. [12] combined character n-grams and word n-grams as a feature for detecting positive and negative deceptive opinions using a regression analysis classifier. Banerjee et al. [13] distinguished deceptive reviews from genuine ones based on the following four linguistic clues: understand ability, level of detail, writing style, and cognition indicators. Based on these features, opinion spams were identified by using a logistic regression model.

Other researchers also addressed review spam detection based on the recent achievements in deep learning and neural networks. Li et al. [14] used convolutional neural networks to learn the semantic representation of review text and then provided a new model named sentence weighted neural network (SWNN) for the recognition of fake reviews. In the text representation learning, considering the sentence weight to better represent the text semantics, the model obtained the term weight based on the KL-divergence method and calculated the sentence weight based on this, so as to learn document-level vector representation. Experimental data show that the SWNN model is more robust than traditional classification models on cross-domain issues. Zhang et al. [15] proposed a novel approach for identifying deceptive reviews, where a word vector with six components is obtained, and the contextual local information of each word is captured by adopting a recurrent convolutional neural network.

2.2. Unsupervised Learning Techniques. Due to the unavailability of properly labeled datasets, the supervised learning method is not always preferable. Alternative unsupervised learning methods for which no labeled data are required tend to focus on individual or group spammers by relying on the features of behavioral reviewers. For example, Liu and Pang [16] believed that the behavior distribution of opinion spammers is very different from that of normal users. For this reason, they proposed an unsupervised review deviation-based framework model to identify spammers, where a series of review and author-level behavior deviation features are introduced, respectively, to measure the spamicity, such as content duplication, review length, the purity of sentiment in review, and extreme rating. Meanwhile, they especially performed the aspect-based opinion deviation mining to model latent content deviation. Ji et al. [17] formulated a variety of individual spam indicators to measure the spamicity of reviewers to isolate candidate spammer groups. Dong et al. [18] proposed an unsupervised topic-sentiment joint probabilistic model (UTSJ) based on a latent Dirichlet allocation (LDA) model, and Zhong et al. [19] proposed an unsupervised method for identifying opinion spammers. In their work, they calculated a reputation score value using weight parameters that turn the contributions of both content-based characteristics and behavior-based characteristics and captured suspected spammers based on a threshold value. Furthermore, in David et al. [20], the authors studied those spammers attempting to manipulate average ratings for product. Different from previous work, they mined reviewer behavior using only ratings without resorting to text-based analysis. They believed those reviewers who have posted a large number of reviews that do not meet the average rating are considered to be spammers with a high probability. In addition, Shaalan et al. [21] conducted detection on singleton reviews, which is one time reviews. They observed that genuine opinions are usually directed uniformly toward important aspects of entities, while spammers attempt to counter the consensus toward these aspects to cover their malicious intent by adding more text but on less important aspects. So, fine-grained opinion representations from review texts were first learned by the unsupervised deep aspect-level sentiment model, and then a long short-term memory (LSTM) network was trained to track the evolution of opinions to identify spam instances. In Atefeh et al. [5], authors proposed a robust review spam detection system wherein suspicious time intervals were captured from time series of reviews by a pattern recognition technique, and the rating deviation, content-based factors, and activeness of reviewers are employed efficiently. The system can reap benefit in terms of time efficiency due to the narrowing of the detection range.

2.3. Semisupervised Learning Techniques. Due to the scarcity of deceptive reviews and the difficulty of manually labeling reviews, semisupervised techniques have attracted the attention of more researchers in dealing with this problem.

For example, Rout et al. [22] demonstrated and evaluated four popular semisupervised techniques: co-training algorithm, expectation maximization algorithm, label propagation and spreading, and positive-unlabeled (PU) learning for spam classification. In their work, four features, namely, sentiment polarity, parts-of-speech tags, linguistic inquiry word count (LIWC), bigram frequency, were extracted from the gold-standard dataset designed by Amazon Mechanical Turk (AMT) [23] and incorporated into the learning model. The experimental data showed that better performance was achieved by using PU learning. However, the limitation is the absence of metadata information and the application of their techniques in real-world datasets.

Li et al. [24] distinguished between spam and truthful reviews on a large number of unlabeled data by performing a co-training algorithm. The co-training semisupervised algorithm is a two-view (review and reviewer-centric) and bootstrapping method and uses a small number of labeled examples to label large numbers of unlabeled data. Zhang et al. [6] proposed the CoSpa algorithm based on SVM benchmark classifier to implement co-training for

identifying spam reviews. They used two types of representations for each review: one is the shallow lexical terms derived from the textual content and the other is the probabilistic context free grammar (PCFG) rules derived from a deep syntax analysis of the review. On the basis of it, the statistical difference between the fake and real reviews was observed by using Wilcoxon signed-rank, and two strategies, CoSpa-C and CoSpa-U, were performed using classifier training. The experimental results demonstrated the effectiveness of the two algorithms in spam review identification.

In addition, a PU-learning algorithm [25, 26] was introduced to employ annotation work for a large number of unlabeled reviews. Donato et al. [25] distinguished fake reviews from sentiment polarity perspectives into positive and negative fake reviews, and they used improved PU-learning methods for effective identification. Ren et al. [26] proposed a novel approach called mixing population and individual nature. The framework first uses a Dirichlet process mixture model (DPMM) to cluster the spy samples in an unlabeled dataset. Then, two schemes (population nature and individual nature) are mixed to determine the category label of the spy examples. Finally, multiple kernel learning is presented to build the final classifiers for fake review detection.

In summary, the supervised learning method heavily depends on a large number of training data, while unsupervised learning techniques are inefficient because they rely on the features of reviewers to detect each fake review. To solve the problem of identification efficiency and data dependence, an effective semisupervised learning approach for detecting spam reviews is proposed, in which the time series model combining with the co-training two-view semi-supervised learning algorithm is performed. The subsequent experiment results show that the proposed algorithm not only avoids data dependence problem but also achieves a trade-off balance between accuracy and efficiency.

3. Fast Detection Model of Deceptive Reviews Based on Collaborative Training

In this section, we aim to propose a system capable of effectively detecting spam reviews. The basic idea is that any burst increase in the number of reviews for a certain product implies significant evidence of spam attacks. Therefore, the system firstly constructs a time series of the number of reviews for a certain product to capture any suspicious intervals and distinguishes between deceptive and genuine reviews using a two-view semisupervised co-training learning method.

3.1. Time Series Model Construction. Generally, there are often multiple sellers on the Taobao website for a certain product. Each store has a set of reviews that are ranked in ascending order based on the release time. Supposing that $R(s) = \{r_1,...,r_n\}$ represents a set of reviews for a product, $TS(s) = \{ts_1,...,ts_n\}$ refers to the set of release times, where n is the total number of reviews, ts_i is the corresponding release time for the ith review r_i, and $ts_i \le ts_j$.

For a certain product, a time series graph of each store is constructed according to the review release time. Thus, the x-axis represents the release time of the review, and the y-axis shows the number of reviews. A time window size (Δt) is defined to divide the element of time into equal intervals. Various sizes, such as one week, ten days, and half a month, were considered as Δt in our work to obtain the most accurate results. The time interval of the review was defined as $I = [t_0, t_0 + \Delta t]$, where I_n refers to the nth time interval. Given a time window size, the number of reviews f denotes the sum of reviews in this time interval. The specific formula was defined as follows:

$$\begin{aligned} I_n &= \left[t_0 + (n-1)\Delta t, t_0 + n\Delta t\right], \\ I &= \bigcup_{n=1}^{N} I_n, \\ f(I_n) &= \sum_{ts_j \in I_n} r_j. \end{aligned} \tag{1}$$

We believe that the burst patterns of reviewing a product imply significant evidence of spam attacks. As a matter of fact, when spammers receive a task, they immediately post fake reviews for a product. The number of reviews sharply increases in those certain intervals because these spam reviews are added to the usual truthful reviews in the interval and create a burst. Therefore, any peaks or extremely fast-growing trends are observed and monitored in the times-series graph. Then, the suspicious time intervals, across which fake reviews are investigated, are captured.

3.2. Feature Analysis of Fake Reviews. Certain features can give a hint on the existence of spamicity. Therefore, we presented these features from the lingual, metadata, and author levels as follows.

3.2.1. Linguistic Features of Review Text

(1) Content Duplication. The intuition is that writing one's imagined experience is difficult. Hence, spammers may choose to copy other existing reviews or make minor changes to them and then post them for the same or other products. Whether a target review is a duplicate review or not can be affirmed by measuring its similarity to other reviews.

Generally, spammers often express their feelings and experiences from multiple perspectives. Therefore, the main content of a review can be reflected using the sentiment evaluation unit, which consists of target-opinion pairs. The fine-grained mining of a sentiment evaluation unit and performing a cosine similarity calculation based on it can determine whether a target review is a duplicate review or not. In this study, a language technology platform (LPT) toolkit (developed by Harbin Institute of Technology) (http://ltp.ai/) was used to extract a sentiment evaluation unit from a review. Then, cosine similarity [27] was adopted to compute the content similarities between them.

(2) Self-Reference. The self-reference of a review is that a reviewer expresses their experiences of all the aspects of a product in the form of the first person. Its purpose is to emphasize that the review is their true experience. For example, we often notice such following reviews: "After reading a lot of very negative review on this album, I thought I should hate this album, but I love it and it's hard to say why, "I love her lyrics and I love the songs and what can I say she has a unique voice which I think is hot and very cool!!!," etc. However, there are some sentences that use "you," "you should.....," etc. to recommend or guide other consumers. Thus, we think such reviews are very suspicious. Therefore, self-reference has an implicit role in deceptive review detection. The ratio of the first-person pronouns to all pronouns can be expressed by the following equation:

$$\text{Self\_reference}(r_i) = \frac{\text{First Person\_pronouns}(r_i)}{\text{Total Person\_pronouns}(r_i)}, \quad (2)$$

where FirstPerson_pronouns ($r[i]$) is the number of first-person pronouns in the target review $r[i]$ and Total-Person_pronouns ($r[i]$) denotes the total number of personal pronouns in a review $r[i]$. We used a max-scale to normalize the self-reference (Normal Self-Reference Score, NPE_Score):

$$\text{NSR\_Score}(r_i) = \frac{\text{Self\_reference}(r_i)}{\text{Max}(\text{Self\_reference}(r_i))}. \quad (3)$$

(3) Review Length. Users tend to write reviews with appropriate lengths to express their true feelings about a product. From the psychological perspective, people are unwilling to spend too much time on non-self-publishing behaviors. The length of fake reviews is generally lower than that of true reviews. For example, spammers sometimes simply use a few simple words to express different opinions, such as "good" or "bad." Therefore, compared with normal reviews, the length of a fake review is shorter than the average length of a normal review:

$$\text{RL\_Score}(r_i) = \frac{\text{Length}(r_i)}{\text{Avg}[R(s)]}. \quad (4)$$

Here, Length ($r[i]$) indicates the word count in reviews $r[i]$ and Avg $[R(s)]$ refers to the average length for all the reviews of a product. To effectively measure the length of the review, this article counts the number of terms excluding stop words.

(4) Ratio of Opinion Words. Spammers usually tend to express extreme sentiment tendencies to promote or demote certain products while receiving posting assignment. For example, they tend to provide extreme sentiment and use words with one kind of polarity (e.g., neg.) much more than the opposite (e.g., pos.) in their reviews based on polarity lexicons [28]. Conversely, the sentiments of genuine reviews include both praise and disapproval. So, we can measure the proportion of the positive or negative words to all the sentiment words in an entire review to determine whether it is fake or not. The formula can be expressed as

$$\text{Opword\_Ratio}(r_i) = \frac{\text{Positive Op\_wordNum}(r_i)}{\text{Total Opword}(r_i)}. \quad (5)$$

Here, total_opword (r_i) represents the total sentiment words in a review r_i and PoNetiveOpinion_WordNum (r_i) denotes the positive or negative opinion words in a review r_i.

(5) Transition Words of the Sentiment Expression. Generally, the sentiment polarity of a fake review is either positive or negative, and transition words are rarely used. However, for genuine customers, their positive reviews may have more expectations and express dissatisfaction with some aspect. The existence of transition words may be a hint to determine whether a review is deceptive or not.

(6) Exclamatory Tone. Some spammers sometimes adopt exaggerated descriptions to emphasize a certain aspect of a product. They use corresponding interjections and exclamation marks when they write reviews. Therefore, by examining whether there are interjections or exclamation marks in a review and then quantifying the degree of the exclamation tone, a fake review can be determined.

3.2.2. Metadata Features

(1) Appending Review Time. Normally, genuine users post their reviews after using a product for a while, and they fully express their experiences and feelings. In contrast, spammers may choose to post reviews again in the same day or in a short period of time to obtain review cashback. So, investigating the appending review time is an important clue to detect whether a review is suspicious or not:

$$\text{Append Time}(r_i) = \frac{b\,\text{time}(r_i) - e\,\text{time}(r_i)}{\Delta t}, \quad (6)$$

where b_time (r_i) refers to the initial release time of the review r_i, e_time (r_i) is the appending release time of r_i, and Δt is the time span.

(2) Appending Pictures. Some genuine users will not only write their own experiences but also publish their corresponding photos after purchasing a product. Thus, there is additional picture information under their review. Spammers cannot post photos because they are not true buyers. Even if they click the purchase button on an online store, the seller will not actually deliver the product to them. From this perspective, we believe that buyers who post additional pictures are genuine buyers.

3.2.3. Reviewer Behavior Features. Some of the specific behaviors of reviewers also have hinted certain suspicious spammers. So, we examined spam reviewers from the following three behavioral characteristics:

(1) User Activity. Generally, once spammers accept a new task, they may use a newly registered user ID to generate continuously multiple fake reviews. After they finish the task, they most properly stop using their fake accounts. The purpose of them is to attract the attention of the consumers and improve the product praise rating and their economic benefits. Ordinary normal users usually post at most one reviews for a product. Therefore, the number of reviews posted by the same user can be used as an indicator to measure User Activity Score (UA_Score), and as a clue to determine whether the user is a spammer or not. The formula can be expressed as

$$\text{UA\_Score} = \frac{r\_\text{num}}{a\ \tan(r\_\text{num}) * (2/\text{pi})}, \tag{7}$$

where $r_$num is the number of reviews posted by the same user. The denominator represents the normalization of the UA_Score.

(2) Super Members. In Taobao website, registered users are divided into ordinary members and super members. Super members are usually the members with naughty values of more than 1000 points. To reach a naughty value of 1,000 or more, users inevitably have to make purchases on Taobao.com. Therefore, we believe that if users are identified as super members, it is more likely that they are genuine users and that their published reviews are genuine.

(3) Review Posting. To lure consumers to make wrong purchase decisions, spammers must post fake reviews. However, many consumers choose not to post reviews after buying products. According to Taobao's existing evaluation system, if a buyer does not post a review, the seller will give a favorable review, and the system will make that favorable review the default one after the transaction is successful. Therefore, from this perspective, users who do not post reviews are considered as normal users.

3.3. Collaborative Training Model for the Identification of Deceptive Reviews in Suspected Intervals. The original collaborative training algorithm was proposed by Blum and Mitchell in 1998 [29]. However, the algorithm needed to meet the conditions that the dataset has two redundant but not completely correlated views. Goldman and Zhou proposed an improved collaborative training algorithm that does not require sufficient redundant views [30]. They used different decision tree algorithms to train two different classifiers from the same attribute set. Each classifier can divide an example space into several equivalent classes. In the collaborative training process, each classifier estimates the label confidence through statistical techniques, marks the example label with the most confidence, and submits it to another classifier as a labeled training set for an update. This process is repeated until a certain stop condition is reached.

Based on the time series graph, suspicious intervals can be captured due to the bursty of the number of reviews. These intervals are considered to be highly suspected and fake, and they are investigated based on Goldman's co-training algorithm. The specific steps are as follows:

Step 1: accurate selection of fake and true reviews

The reviews in the suspected interval are not all fake reviews, as genuine users can make real purchases under the influence of fake reviews and deceptive sales. Therefore, suspected intervals usually contain some normal reviews. Thus, the reviews with appending texts or photos are selected as true seed reviews, and duplicate reviews are identified as fake seed reviews using high semantic similarity or extremely sentiment polarity. The two types of reviews together constitute the initial labeled review set L, and the rest of the reviews in the suspected interval constitute the unlabeled review set U.

Step 2: fake review identification based on the co-training two-view algorithm

Based on the purchase behavior of a reviewer, user activity, super members, review posting, and metadata features of reviews, such as appending review time and appending pictures, the random forest algorithm is used to train a classifier model of the labeled review set L, which is used to predict the reviews in U and then accurately select the top N positive and negative examples (labeled as $T1$). Meanwhile, based on the linguistic clue features, SVM is also used to train another classifier model from the same labeled review set L and to label the reviews in U. The top N positive and negative examples with high confidence are selected and labeled as $T2$. The final category of the reviews in $T1$ and $T2$ is determined using voting methods. That is, if the prediction results of the two sets are both consistent, the category is marked, and these reviews are taken from U to L.

Step 3: when an unlabeled review set U is not empty, step 2 is iteratively performed to complete the labeling tasks to a large number of unlabeled reviews. Finally, all the reviews are distinguished to know whether they are true or fake.

4. Experimental Results and Analysis

The purpose of this section is to evaluate whether the proposed model can be effectively used for the rapid identification of fake reviews. In the following section, a dataset was first described, including the collection and labeling methods. Then, the experimental settings and evaluation criteria were discussed.

4.1. Dataset Collection and Description. The dataset of our work is based on the publicly collected Taobao review. The used dataset is also available as categorized in various genres of products. For this analysis, we used the Meidi rice cooker MB-WFS3018Q from the small appliances dataset (a total of 40 sellers and 10074 reviews).

From the data, as shown in Figure 1, we found that 91% of the reviews were for 5 stores (9216 reviews). Consequently, in the subsequent experiments, we used these five seller reviews to analyze our model.

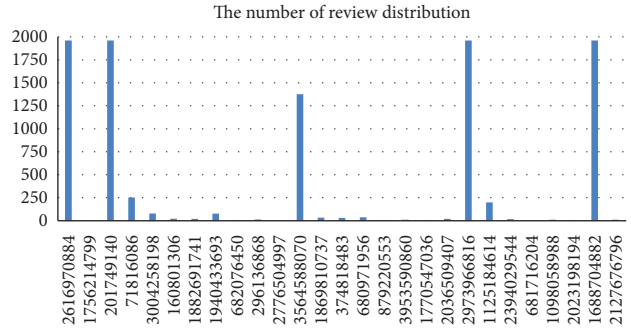

Figure 1: Distribution of the reviews per store.

Since there are no standard answers for the fake reviews in the entire dataset, we needed to manually label the dataset. Based on the comprehensive feature analysis of the previous fake reviews 1, we developed corresponding specifications for labeling the tasks and invited undergraduate and graduate students to mark the reviews in the dataset as true or deceptive. To more quickly and accurately complete the labeling task, each student was assigned 50 reviews, and each review was labeled by three people at the same time. The final labeling result was performed using the Simple MAJORITY Voting Ensemble. MAJORITY is a voting method based on the principle of the subordination of the minority to the majority. That is, at least two people in the manual annotation results unanimously judged the review as a fake review. Then, the review was eventually determined to be a fake review.

4.2. Results and Analysis

4.2.1. Experiment #1. Research Question addressed (RQ1: what is the most appropriate suspicious interval time window?)

To quickly identify fake reviews, it is necessary to detect suspicious intervals, which usually have a peak or a curve with a rapid growth trend in the time series graph based on a burst of reviews. However, different time windows slid on the time series would show fluctuations in the results. On one hand, if the value of a time window is too short, the sudden increase in the number of reviews and product sales would not be obvious. On the other hand, if the set value is too long, suspicious behaviors can be hidden, resulting in a large amount of following suspicious reviews. Generally, spammers start to generate fake reviews in a certain period of time. Therefore, appropriately selecting suspicious interval time windows is the first critical task. According to actual situations and experience, we used ten days, half a month (15 days), and one month (30 days) as time intervals and, respectively, carried out the detection of suspicious intervals. The experimental results were demonstrated as shown in Figure 2.

As seen from the above figure, the number of reviews showed a sharp increase in a certain period of time. We believe that these intervals have greater suspicion of cheating, so the reviews across them should be further investigated. In selecting the appropriate time windows, we observed that the number of reviews showed a consistent trend. However, the 14-day interval not only clearly displayed the increase trend but also covered more fake reviews than the other interval chart. In our proposed model, after the capturing of the suspicious intervals, the model performed the subsequent fast fake review identification step. The range of time windows cannot be too broad. Otherwise, the number of reviews to be quickly identified would be too large. Furthermore, compared with the 10-day trend figure, 14-day trend showed a burst interval with an obvious growth trend and meanwhile covered more fake reviews than the 7-day interval. Therefore, considering the above two aspects, we chose the 14-day range as a time window for the subsequent opinion spam recognition.

4.2.2. Experiment #2. Research Question addressed (RQ2: how is the identification performance using the time series and co-training scheme?)

In e-commerce platforms, the monitoring management is based on each store. We extracted 5 store reviews from suspected intervals based on 14-day intervals and marked the fake reviews as 1and the normal reviews as 0. We considered the hit rate (the ratio of the number of correctly identified fake reviews to the number of identified fake reviews), recall (the ratio of the correctly identified fake review to all fake reviews in dataset), precision (the ratio of the correctly identified fake review and real reviews to all reviews in dataset), and F1 measures as performance metrics in our experiment. In addition, to reduce the biases of the classifiers, we used 10-fold cross-validation with proper data splitting (training and testing). Table 1 illustrates the ten-fold cross-validation results.

Meanwhile, to verify the effectiveness of the model, Atefeh's algorithm (Atefeh Heydari et al. 2016) and Wen Zhang's algorithm (Wen et al. 2016) were employed as baselines for the comparison experiments. Atefeh's algorithm is an unsupervised learning method using the time series for fake review detection. Wen Zhang's algorithm is a semisupervised fake review detection method in which two features of lexical PCFG are used for co-training. Considering the consistency and number of reviews in the comparison experiment, we selected the reviews in the store number 201749140, and the results are shown in Table 2.

From the results, it can be seen that the proposed model outperformed the baseline models, as it showed higher identification accuracy in addition to better efficiency. Hence, a compromise balance of the identification performance was achieved. Specifically, in terms of the detection efficiency, although our model was inferior to Atefeh's algorithm, compared with Wen Zhang's work, the time performance was greatly improved. This is because Atefeh's algorithm does not require training, leading to an optimal time cost. Our model captured the suspicious interval using the time series, resulting in a narrowed detection range, and hence the time cost of the fake review identification could be significantly decreased. In the model proposed by Wen Zhang, extracting the syntactic structure of the text itself takes a lot of time, and the processing data scale is not

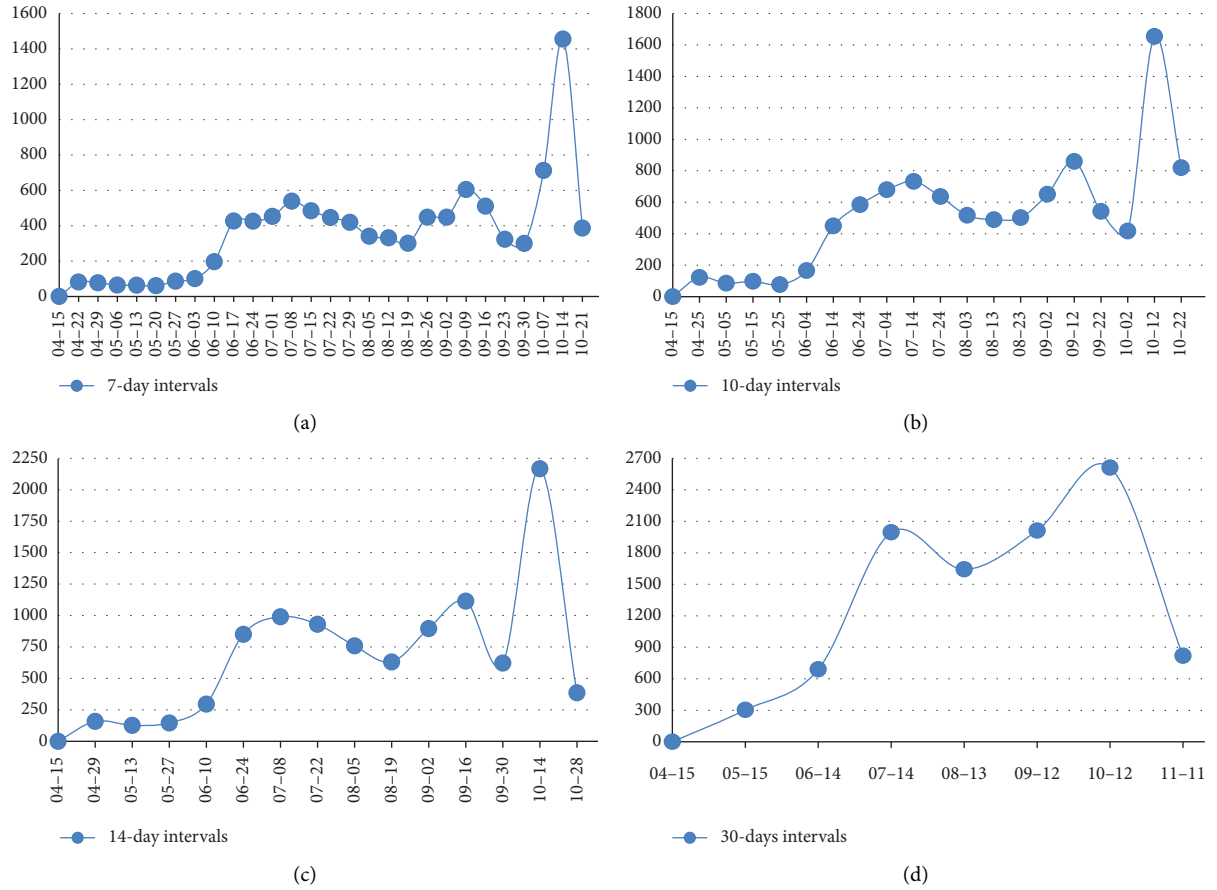

Figure 2: Trend chart of the number of reviews in different time intervals. (a) Trend graph of the review number at 7-day intervals. (b) Trend graph of the review number at 10-day intervals. (c) Trend graph of the review number at 14-day intervals. (d) Trend graph of the review number at 30-day intervals.

Table 1: Identification results of each store.

Store no.	The number of reviews	Hit rate	Recall	F1 measures	Precision
201749140	1960	0.90	0.77	0.82	0.81
3564588070	714	0.71	0.62	0.66	0.72
2616970884	574	0.68	0.62	0.65	0.74
2973966816	571	0.88	0.80	0.83	0.83
1688704882	322	0.82	0.70	0.76	0.75

Table 2: Performance comparison results of the fake reviews.

	Hit rate	Recall	F1 measures	Precision	Identification time (second)
Our model	0.90	0.77	0.82	0.81	725
Atefeh's work	0.70	0.68	0.69	0.65	372
Wen Zhang's work	0.78	0.92	0.84	0.80	8618

Table 3: Performance comparison results of the two types of features in a single view.

Features	Hit rate	Recall	F1 measures	Precision
Random forest classifier based on text features	0.82	0.76	0.79	0.82
Random forest classifier based on behavior features	0.57	0.67	0.61	0.52

Table 4: Effect results for 11 features.

Features	Hit rate	Recall	F1 measures	Precision
Duplicates reviews	0.58	0.40	0.48	0.49
Self-reference	0.83	0.18	0.30	0.51
The length of review	0.79	0.80	0.79	0.76
The ratio of opinion words	0.81	0.47	0.60	0.61
Transition words of the sentiment expression	0.72	0.15	0.25	0.48
Exclamation tone	0.85	0.18	0.30	0.51
Appending time	0.63	0.80	0.70	0.55
Appending pictures	0.58	0.90	0.70	0.57
User activity	0.59	0.85	0.69	0.57
Super member	0.57	0.42	0.48	0.48
Review posting	0.58	0.92	0.71	0.59

diminished, so the performance in terms of the recognition efficiency is not satisfactory. In terms of accuracy, compared with the algorithm proposed by Atefeh, the hit rate of our model increased by 29%, the F1 measure increased by 18.8%, and the precision increased by 24.6%. The proposed model in this study is a semisupervised learning method of collaborative training in which different classification models are used to train datasets multiple times based on different features and then combine ensemble learning to vote. The algorithm proposed by Athefeh is an unsupervised identification method in which the features of fake reviews are not fully mined and the setting of a threshold value is uncertain. Furthermore, our proposed model takes into account various characteristics of Chinese product reviews and fully analyzes the characteristics of fake reviews written by spammers in Chinese. Also, our model put forward more comprehensive features from three aspects: language clues, review metadata, and reviewer behavior, which meet the language environment of Chinese and thus achieve better recognition performance.

4.2.3. Experiment #3. Research Question addressed (RQ2: what is the role of spam indicators (features) in detecting spam reviews?)

This paper proposed a dual-view collaborative training model in which all language cue features are trained by the random forest model and the other metadata and behavior features are trained by the SVM algorithm. In this section, we analyzed the roles of different features in review spam detection. To facilitate the comparison, the random forest classification algorithm was implemented in the experiment.

(1) The Influence of the Text Content and Behavior Factors on the Model. Our proposed model not only considered the text content factor but also the user's purchase behavior characteristics. We first analyzed the impact of the two types of features on the model performance. Thus, we separately conducted the detection of fake reviews experiment. That is, when testing the text features (language clue features), only the text features were imported to the random forest text classifier. Similarly, when verifying the behavioral features, only the behavioral features were used to train the random forest classifier. The performance comparison results are shown in Table 3.

From the data mentioned in Table 3, it can be seen that both the text and behavioral characteristics were effective in identifying fake reviews. However, compared with the metadata and user behavioral characteristics, the linguistic cue features of the reviews could impact more on the identification performance of the fake reviews and obtain a higher hit ratio, recall, and F1 measure. Opinion spammers inevitably show some suspicious clues when submitting fake reviews, such as the use of more emotional words, exclamatory tones, and first-person expressions. Meanwhile, they also control themselves to avoid supervision. For example, the appending review time should be as long as possible, and the same user name should be released as little as possible so that the linguistic cue characteristics of the reviews are more effective than the behavior and metadata features.

(2) The Utility of Each Feature. To further analyze the role of each linguistic clue, metadata, and the behavioral features of the reviewers in the proposed model and find the key factors affecting the identification accuracy, we continued to analyze the 11 features based on the random forest classification model. Table 4 shows the performance results of different features in fake review detection.

From the data in Table 4, we could draw the following conclusions:

(1) Among the linguistic clues, exclamation tones, the ratios of the opinion words, and self-reference were more effective in identifying fake reviews than the other features, and they showed a higher hit ratio. However, these single features are not sufficient for fake review identification, and the recall is generally low. Opinion spammers try their best to write fake reviews from multiple aspects. Obviously, a single language clue aspect cannot cover all cases, resulting in relatively low recall.

(2) The two features of the appending pictures and not posting reviews are far more useful in distinguishing genuine reviews from fake ones, and it can be seen that the corresponding recall is very high, reaching above 0.9. This is because these two features are binary features. In our proposed model, we considered the reviews with images as genuine reviews and the users who do not post reviews as ordinary

users and not spammers. In the experimental data, most of the crawled reviews contained text content without pictures, and only a few reviews were system default positive reviews (the users did not post reviews). Therefore, our detection model can treat most of the reviews as deceptive reviews, resulting in a high recall and high F measure. In fact, the reviews without additional pictures may also be true reviews, and the users who posted these reviews may be ordinary users. Thus, these features alone cannot be very effective in identifying fake reviews.

5. Conclusions and Future Work

By focusing on improving the detection efficiency of fake online reviews and reducing the dependence on training sets, this paper proposed a semisupervised approach for fast fake review identification. A time series model is first adopted to capture the suspicious intervals by investigating the burst patterns in the review process. Aiming at the reviews in these intervals, a co-training dual-view algorithm is used to distinguish fake reviews from real ones, where comprehensive features, such as language clues, metadata, and user purchasing behaviors, are, respectively, used to train a classifier model by the random forest or SVM algorithms.

The contributions of this paper can be summarized as follows. First, our proposed approach could achieve high detection efficiency, which is a great advantage for real-time spam filter systems. Compared with traditional identification methods, our model greatly narrowed down scope samples using suspicious interval detection. Second, we approached the review spam detection problem using a co-training dual-view algorithm to solve the shortcomings of the supervised learning technique, which needs large sets of labeled instances from both classes in addition to deceptive and truthful opinions. Meanwhile, good detection accuracy was achieved. The obtained experimental results are encouraging and indicate that only using few positive and negative opinions for training can make it possible to reach an F1 measure of 0.8 and get a compromise balance of the identification performance.

To some extent, although this model outperforms other traditional statistical methods, it still has some restrictions that need to be improved in the future. (1) Limited data are used in the experiment for inaccessibility of them. On the Taobao website, not all the reviews are public and the website only allows checking a small amount of top N review data. So, experiments can be conducted on other large datasets, such as Jingdong or Suning. (2) The feature selection is simply performed, especially the sentiment factor consideration. Thus, exploring and analyzing more linguistic features, such as modifiers, negations, emojis, and ironic words, may be beneficial in improving the detection performance.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

The work presented in this article was supported by the National Science Foundation of China under grant nos. 71762017, 71861014, 71861013, and 61802120; Key Science and Technology Project of Hunan Provincial Department of Education under grant nos. 20A081 and 19A077; and Changsha Municipal Natural Science Foundation under grant no. kq2014063.

References

[1] S. Mohamed, A. K. Hatem, and A. andAmira, "An analytic framework for enhancing the performance of big heterogeneous data analysis," *International Journal of Engineering Business Management*, vol. 13, pp. 1–11, 2020.

[2] M. Luca and G. Zervas, "Fake it till you make it: reputation, competition, and yelp review fraud," *Management Science*, vol. 62, no. 12, pp. 3412–3427, 2016.

[3] J. K. Rout, S. Singh, S. K. Jena, and S. Bakshi, "Deceptive review detection using labeled and unlabeled data," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 3187–3211, 2017.

[4] Z. Hai, P. Zhao, P. Cheng et al., "Deceptive review spam detection via exploiting task relatedness and unlabeled data," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1817–1826, Austin, TX, USA, November 2016.

[5] H. Atefeh, T. Mohammadali, and S. Naomie, "Detection of fake opinions using time series," *Expert Systems with Application*, vol. 58, pp. 83–92, 2016.

[6] W. Zhang, C. Bu, T. Yoshida, and S. Zhang, "Cospa: a co-training approach for spam review identification with support vector machine," *Information*, vol. 7, no. 1, pp. 12–26, 2016.

[7] L. Huayi, F. Geli, W. Shuai et al., "Bimodal distribution and Co-Bursting in Review spam detection," in *Proceedings of International World Wide Web Conference*, pp. 1063–1072, Geneva, Switzerland, August 2017.

[8] A. Heydari, M. a. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: a survey," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3634–3642, 2015.

[9] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the first ACM international conference on web search and data mining (WSDM)*, pp. 219–229, Palo Alto, CA, USA, November 2008.

[10] Y. Lin, T. Zhu, X. Wang et al., "Towards online review spam detection," in *Proceedings of the companion publication of the 23rd international conference on World Wide Web Companion*, pp. 341-342, Seoul, Korea, April 2014.

[11] K. D. Rupesh and A. K. Singh, "Identification of fake reviews using new set of lexical and syntactic features," in *Proceedings of the sixth international conference on Computer and Communication Technology*, pp. 115–119, Hue, Vietnam, December 2015.

[12] A. Hafifiz, M. A. Siagian, and M. Aritsugi, "Combining word and character N grams for detecting deceptive opinions," in *Proceedings of the IEEE 41st Annual Computer Software and Applications Conference*, pp. 828–833, Turin, Italy, July 2017.

[13] S. Banerjee, A. Chua, and J. Kim, "Using supervised learning to classify authentic and fake online reviews," in *Proceedings of the 9th international conference on Ubiquitous Information Management and Communication*, pp. 938–942, Bali, Indonsia, January 2015.

[14] L. Li, B. Qin, W. Ren, and T. Liu, "Document representation and feature combination for deceptive spam review detection," *Neurocomputing*, vol. 254, pp. 33–41, 2017.

[15] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, "DRI-RCNN: an approach to deceptive review identification using recurrent convolutional neural network," *Information Processing & Management*, vol. 54, no. 4, pp. 576–592, 2018.

[16] Y. Liu and B. Pang, "A unified framework for detecting author spamicity by modeling review deviation," *Expert Systems with Applications*, vol. 112, pp. 148–155, 2018.

[17] S.-j. Ji, Q. Zhang, J. Li et al., "A burst-based unsupervised method for detecting review spammer groups," *Information Sciences*, vol. 536, pp. 454–469, 2020.

[18] L.-y. Dong, S.-j. Ji, C.-j. Zhang et al., "An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews," *Expert Systems with Applications*, vol. 114, pp. 210–223, 2018.

[19] M. Zhong, L. Tan, and X. Qu, "Identification of opinion spammers using reviewer reputation and clustering analysis," *International Journal of Computer Communication & Control*, vol. 14, no. 6, pp. 759–772, 2019.

[20] S. David, Z. Z. Xiu, H. Y. Xing et al., "Detection of opinion spam based on anomalous rating deviation," *Expert System Application*, vol. 42, pp. 8650–8657, 2015.

[21] Y. Shaalan, X. Zhang, J. Chan, and M. Salehi, "Detecting singleton spams in reviews via learning deep anomalous temporal aspect-sentiment patterns," *Data Mining and Knowledge Discovery*, vol. 35, no. 2, pp. 450–504, 2021.

[22] J. K. Rout, A. Dalmia, K.-K. R. Choo, S. Bakshi, and S. K. Jena, "Revisiting semi-supervised learning for online deceptive review detection," *IEEE Access*, vol. 5, pp. 1319–1327, 2017.

[23] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proceedings of the 49th annual meeting of the Association for Computational Linguistics: human language technologies*, pp. 309–319, Portland, OR, USA, June 2011.

[24] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," in *Proceedings of the international joint conference on artificial intelligence (IJCAI)*, pp. 2488–2493, Barcelona, Spain, July 2011.

[25] H. F. Donato, M. Manuel, R. Paolo, and R. Guzmán Cabrera, "Detecting positive and negative deceptive opinions using PU-learning," *Information Process Management*, vol. 51, pp. 433–443, 2015.

[26] Y. F. Ren, D. H. Ji, H. B. Zhang et al., "Deceptive reviews detection based on positive and unlabeled learning," *Journal of Computer Research Development*, vol. 52, pp. 639–648, 2015.

[27] G. Salton, A. Wong, and C. S. Yang, "A vector space model for information retrieval"" *Communications of ACM*, vol. 18, no. 11, pp. 513–523, 1975.

[28] R. Basili, D. Croce, and G. Castellucci, "Dynamic polarity lexicon acquisition for advanced social media analytics," *International Journal of Engineering Business Management*, vol. 9, pp. 1–18, 2017.

[29] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT)*, pp. 92–100, Madison, WI, USA, July 1998.

[30] S. Goldman and Y. Zhou, "Enhancing supervised learning with unlabeled data," in *Proceedings of the 17th international conference on Machine Learning (ICML)*, pp. 327–334, Bellevue, WA, USA, July 2000.