Fake Reviews Detection based on Reviews and Comments
Team 03

Rajashekar Reddy Kommula
016007043
rajashekarreddy.kommula@sjsu.edu

Aanchal Agarwal
015923778
aanchal.agarwal@sjsu.edu

Ratika Bhuwalka
015721511
ratika.bhuwalka@sjsu.edu

Navika Babbar
015295033
navika.babbar@sjsu.edu

***Abstract*– For many different decision-making processes, including purchase and sale decisions, online comments and reviews have emerged as prominent assets. Since false reviews can influence customers' judgments owing to misleading descriptions and dishonest selling, their veracity is crucial for both buyers and sellers. Innocent clients may suffer financial losses as a result of this. Thus, detection of the fake reviews has garnered more interest nowadays. A majority of online sites for shopping have solely paid attention to handling negative reviews and comments. In contrast to only looking at reviews and comments, the strategy we suggest in this study for identifying outlier reviews involves looking at records related to the items under consideration. We initially use a dataset from Amazon China that has been crawled to assess the characteristics of this data, indicating that the reviewing records for each product are consistent with those for typical products. As part of the suggested methodology, we started by extracting the records of reviews for various commodities to a temporal feature vector. Using the review records of reviews and comments, we proceed by creating an isolation forest algorithm for the outlier product review identification. Using the dataset crawled from Amazon China, we will assess the performance of our suggested method and contrast it with a few other temporal outlier detection techniques already in use. We'll also look into the effects of the parameters used for record review.**

Keywords—fake words; reviewing records; products speculation; Isolation Forest Algorithm

I. INTRODUCTION

"For the purpose of identifying fake and fraudulent reviews, text mining algorithms have recently been generated. Such study focuses on examining one review at a time and pays little attention to the possible links between numerous reviews or reviewers. Han et al. examined reviewer behavior and outlier reviewer behavior using burst reviews. A subsequent step is to examine the review patterns of phony reviewers in order to identify fraudulent reviews based on their actions.

A brief review of some related feature learning work in networks is presented in Section 2. The review record and outlier behavior are empirically analyzed in Section 3. Using a real dataset, we describe a method for detecting fake product reviews based on an isolation forest-based approach in Section 4 and evaluate the method against several baseline methods in Section 5 to demonstrate its advantages. Section 6 serves as the paper's conclusion, and it also identifies several intriguing lines of inquiry for further study.

II. RELATED WORK

A. Spamming Detection

Several studies have been conducted on web spam and email spam in recent years. An example is the provision of a survey on the detection of web spam. A study is also conducted on email spam detection. Spam on blogs and networks is also intensively studied. Using fake reviews as an example, Fei et al. provided possible spam patterns according to the behavior of fake reviews.

1. Time Series Outlier Detection

An actively pursued method to outlier detection is time series analysis. In these methods, the similarity function is determined by measuring the similarity among the two sequences, and by using clustering the outlier is located. The data samples can be clustered based on their time series feature vectors, with the largest outlier score going to the sample that is farthest from all other clusters.

An unsupervised method of detecting outliers using parametric models consists of not specifying anomalous instances and creating a summary model of the base data. In addition, HMMs are easy to interpret but to pattern the complex data they do not scale well. Outlier detection strategies based on HMMs were brought up.

2. Outlier Detection for Streamed Data

Outlier detection for streaming data is a different group of the method, where the situation is more complicated than when outlier identification is used normally.

In order to make predictions with streamed data, prediction models that evolve as the data is processed are necessary, updating the parameters or model elements as the data is processed. The detection of outlier products can be achieved using online clustering, for example. Dynamic Bayesian networks were also proposed as a way to model time-varying data samples. It is possible to determine the state of a system by the addition of new state variables.

For detecting outliers in streams a method is proposed in this paper, combining ideas from outlier detection in streams and outlier detection in time series.

III. TREND ANALYSIS OF PRODUCT REVIEWS

Using Amazon data crawled from shopping reviews, we perform analysis in this section. We can clearly differentiate how reviews and comments for different products differ from each other from the analysis.

A. General Trend for Product Review

Amazon-China is the dataset we used in this study. Increasing numbers of reviews have been recorded. The number of reviews recorded in 2006 was relatively small. The reviews and comments count increased as time went by.

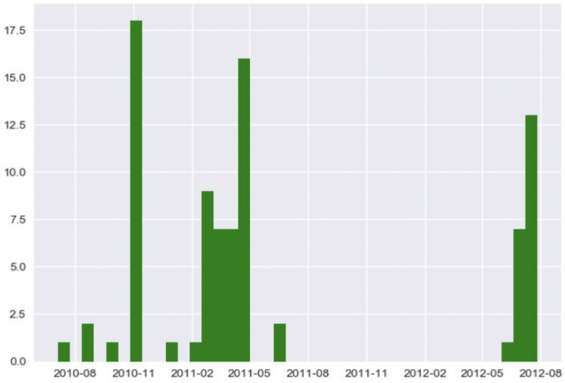
Table 1 summarizes the review parameters, including 166,624 products and 5,055 users between March 2006 and August 2012. 1205,125 reviews are available in total.

Information	Value
Products	166624
User	5055
Time period	March 2006 to August 2012
Number of reviews	1205125
Frequency	507.2 reviews per day

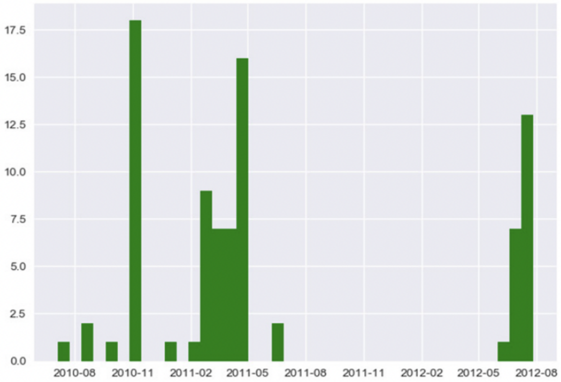
Table 1. Fake Review Dataset

B. The Trend in Product Reviews

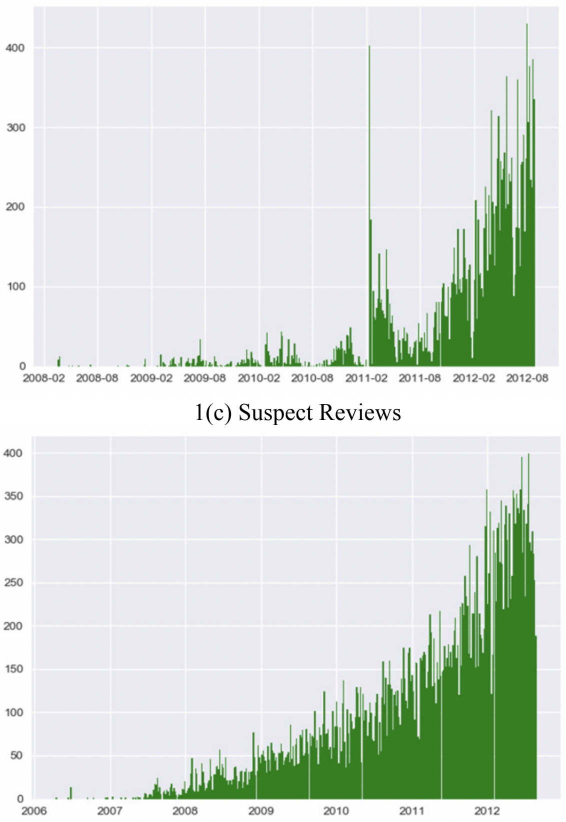
We examine the temporal structure of reviews based on a selection of two products, as depicted in Figure 1. Figure 1(a) makes it more obvious that the trend is not constant. Figure 1(c) depicts the suspicious reviews, whereas Figure 1(d) summarizes the normal reviews.



1(a) Suspect Reviews



1(b) Normal Reviews



1(c) Suspect Reviews

1(d) Normal Reviews

Figure 1. Pattern of reviews of two typical selected products

Our method uses products as its basic data unit. Outlier products are derived from a study of review patterns across all products. The number of products in a given time window may be rather significant due to the fact that we employ an unsupervised technique to identify false reviews, making it more probable that we may discover fake reviews in combination with other fake reviews. Our unit description of a product only includes reviews within one day in accordance with this characteristic. Using product Pi as an example, the review pattern is defined as follows:

$$V_{p_i} = [p_{i_1}, p_{i_2}, p_{i_3}, \cdots, p_{i_E}] \tag{1}$$

where i represents the date of the record, i_1 represents the first review found in the dataset (March 2006), i_E represents the last review found in the dataset (August 2012), and V_{P_i} represents the feature vector representing the reviews for product P_i. P_{ij} represents the number of reviews given in a certain date i_1 and each element in the vector is a date-review indicator.

C. Temporal Feature Extraction based on Reviews and Comments

The reviews of products can be aggregated into an N-dimensional vector, and we view the records as statically. The temporal feature's general form can be described as follows:

$$Z_{p_i} = \{z(1)_{P_i}, z(2)_{P_i}, \cdots, z(t)_{P_i}, z(t+1)_{P_i}, \cdots, z(N)_{P_i}\} \tag{2}$$

where $z(t)_{P_i}$ □ R^N ($t \geq 1$), the overall number of time slots is N. to be handled and specifies the number of reviews in time slot t for product P_i.

With a period of five years between 2006 and 2012, we have seven time slots total since the data is for the years 2006 to 2012. A common dimension should also be defined for all products. It follows that z(t) will be the following for a time slot of M days and N total time slots:

$$z(t)_{P_i} = \sum_{m=1}^{M} p_{i_{t*M+m'}} \tag{3}$$

where t is the t_{th} time slot of the feature and subscript t*M+m' indicates the date time of the specific review. We can then describe all products using a matrix:

$$Z = [Z_{P_1}, Z_{P_2}, \cdots, Z_{P_P}]_{P\times N} \tag{4}$$

where the total number of products is given by P. The data can be processed by the isolation forest algorithm using a set of nodes.

D. Isolation Forest Algorithm for Outlier Detection

Using the bootstrap sampling from dataset Z as a basis, we first develop the initial outlier detection model for Z the temporal feature vector. iTrees arranged in an ensemble E are comprised of a number of L iTrees, namely, that is built using data from ith time slot.

$$E = \{E_1, E_2, E_3, \cdots, E_L\} \tag{5}$$

Multiple isolation trees, called iTrees, are combined into an iForest in the algorithm. We know that iTrees are created by selecting features of a product temporal review randomly and then calculating feature values based on those features. Depending on the temporal review value chosen at each node of the isolation tree, the instances set is divided into two parts. An outlier review product is usually one with a record or value of reviews that is very different from the average product record. Averaging the depth of products in a forest can serve as an anomalous score for the products, alleviating the effects of random characteristics imported into isolation forests. It is more likely that an outlier product will have a lower score than the normal products, since the lower the score, the further away it is from the normal products. A further illustration of the algorithm can be found in figure 7.

The isolation forest is built based on the records of product reviews using an isolation forest algorithm. In the meantime, isolation forest algorithm can be applied to obtain the outlier score.

Outlier products are identified by their anomaly score. By applying formula(6) to product P_i, we can calculate the anomaly score.

$$S(z_{P_i},N)=2^{-\frac{E(h(z_{P_i}))}{c(N)}} \tag{6}$$

where

$$E(h(x))=\frac{1}{L}\sum_{i=1}^{L}h_i(x) \tag{7}$$

As shown in formula(6), the sampling size in Algorithm 1 is N, hi(x) indicates the length of the ith iTree, E(h(x)) indicates the average of h(x) under a given N, and c(N) represents the average of h(x) under just that N. $S(z_{pi},N)$ represents the outlier score for P_i. Outliers are considered high anomaly scores, while normal samples are considered low anomaly scores. Product review patterns that are high in anomaly indicate that the products are being reviewed at a different time than normal products. An anomaly score of high means that there is a high probability that the product contains fake reviews and comments. By using Algorithm and formula(6), we can determine whether there are abnormal reviews for products.

IV. EXPERIMENT AND ANALYSIS

The dataset that is described in Section III has been used in some experiments. Our first objective is to demonstrate the effectiveness of our proposal by comparing it with a few baseline fake reviews speculation methods. We also compare the performance of our technique with various choices of temporal parameters.

A. Measurement Metrics

As a tool for assessing the performance of our method, we quantified the performance of the predicted outlier labels in comparison to the ground-truth outlier labels. Our method is measured by two metrics.

B. Comparative Analysis and Evaluation of Experiment Results

The next section describes how to detect outlier products from the dataset described in Section 3 by utilizing three baseline methods - ARIMA, LOF, and SVM. To detect whether a product is abnormal, we re-crawled it following the same approach. Low anomaly scores are considered abnormal for each product, making the product outlier commercially. Based on the results of accuracy, our method was able to detect fake reviews more effectively than the other methods, as shown in Figure 2. Detecting fake reviews is more likely to be successful when the method has higher accuracy rather than lower accuracy.

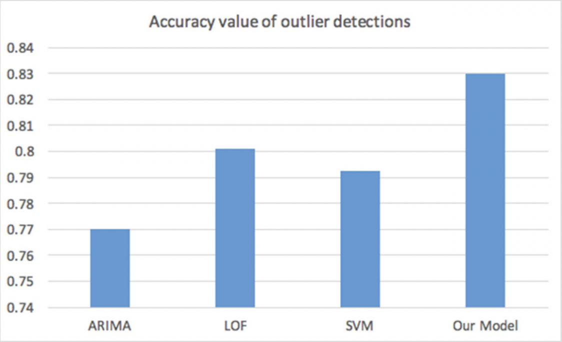


Figure 2. Comparison of accuracy

This figure shows how well our method compares with the other three baseline methods. Insignificant performance changes in time-series may explain ARIMA's accuracy of 0.77. Based on the results, LOF shows that outliers can occur locally as well as globally. All other methods do not perform as well as isolation forest-based methods.

The efficiency of our detection method was also compared with those of the three baseline methods as shown in Figure 3. As a result, the tree-based approach is both faster in the training phase and quicker in the evaluation phase when using the isolation forest method.

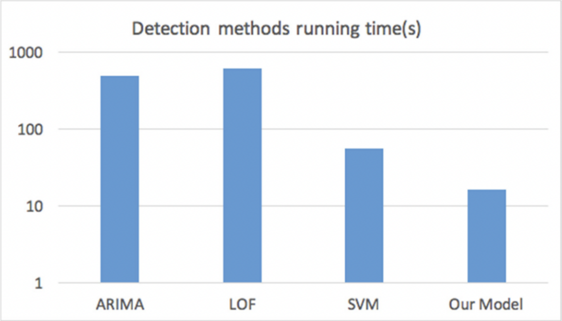


Figure 2. Comparison of efficiency

V. CONCLUSION

A novel approach is presented in this paper to detect fake reviews of products by analyzing review records of online shopping sites. By analyzing temporal trends of reviews and comments, this method detects the outlier products. Compared to some existing methods, ours offers an advantage. To prove the efficiency and effectiveness of our method, we have also compared it with several temporal outlier detection methods. The detection of fake reviews based on reviews records has many challenges. It is interesting to look into the likelihood of a product being involved in fake reviews and comments in the future, since the results of our study didn't indicate clearly when a product is most likely to be involved in fake reviews and comments.

VI. REFERENCES

[1]. Han S.; Prince J.; Zuo L.; Carass A. Automatic outlier detection using hidden Markov model for cerebellar lobule segmentation. Proceedings of International Conference on medical Applications in Molecular, Structural, and Functional Imaging, 2018.

[2]. Chirita, P.A.; Diederich, J.; Nejdl, W. MailRank: using ranking for spam detection. Proceedings of the 14th ACM International Conference on Information and Knowledge Management, 2005, pp. 373-380.

[3]. Smith, D.V.; Timms, G.P.; de Souza, P.A.; D'Este, C. A Bayesian framework for the automated online assessment of sensor data quality. Sensors, 2012, 12(1): 9476-9501.