

Detecting Spamming Stores by Analyzing Their Suspicious Behaviors

Ji Chengzhang*, Dae-Ki Kang**

*Weifang University of Science & Technology, Weifang 262700, China

**Dongseo University, 47 Churye Ro, Sasang-Gu, Busan 617-716, Republic of Korea

jcz8888@163.com, dkkang@dongseo.ac.kr

Abstract— The purpose of this paper is to detect the stores with spamming behaviors. We identify suspicious behaviors of these stores to detect spamming stores. These suspicious behaviors are from the two following observations. First, spamming stores may target quantity of sale and product reviews to influence consumers' decisions. Second, they tend to deviate from the other stores in quantity of the sale and reviews. From those observations, we propose a novel scoring methods to find spamming stores, and they are applied on Aliexpress dataset. Our experiment results show that our proposed methods are effective in finding spamming stores.

Keywords— Spamming behavior, spamming store, detection method

I. INTRODUCTION

A. Motivation

With the rapid development of e-commerce, more and more consumers prefer shopping online. Also, consumers are increasingly likely to post their reviews of their products. The product reviews provide some valuable information for manufacturers and potential consumers. Because of the existence of this interest, there are some untrue or false reviews which mislead consumers. So it is particularly important to detect fake reviews.

One of the main methods of detecting review spam is to treat the reviews as the target of detection [1]. This approach describes a review using features of review, reviewer and product, and trains a classifier to detect fake reviews. However, these features may not provide direct evidence against the fake reviews. Another method of detecting review spam focuses on detecting review spammers based on their spamming behaviors, instead of detecting fake reviews directly [2]. In fact, this method has some shortages: 1) some spam reviewers also post some normal reviews. 2) it can't find and cut off the genuine source of fake reviews, and new spam reviewers will be hired to post fake reviews even if some spam reviewers are removed. We think about this problem from another perspective which is to detect the source of spamming behaviors. In C2C E-commerce sites, the aim of spamming behaviors are to make the spamming stores to be more interesting, so spamming stores are the most important source of these spamming behaviors. If we find the stores that

hire people to post spam reviews, we can remove the stores' reviews or degrade the stores' rating so as to lower the influence of spam reviews.

In C2C shopping sites (e.g. AliExpress.com and ebay.com), there are a large number of stores. It is very possible that these stores hire some review spammers to post fake positive reviews in their stores to improve quantity of sales. If we can detect these spamming stores, we will find an important source of fake reviews and decrease the fake reviews.

Some stores also hire some "Internet water army" to buy their own products even along with positive reviews and high ratings to make their products look very popular¹. For instance, stores selling on Alibaba's main platforms, Taobao, Tmall and AliExpress have been increasing their sales totals by paying people to buy goods that are returned soon afterwards². This deceptive maneuver increases the rankings of the stores on the sites – therefore making it easier to make genuine sales in the future because shoppers flock to the high-grossing stores (called sheep flock effect). The Wall Street Journal reported on the practice of faking orders by vendors on Alibaba's platforms³.

In this paper, we focus on finding spamming stores on C2C shopping websites. Review spam and sale spam are designed to give unfair view of some products to affect consumers' perception of the products. In [3], 10 to 15% of reviews basically echo earlier reviews and may potentially be affected by fake review. Detecting spamming stores is a challenging task as nobody knows exactly the amount of spamming stores in existence. Due to the openness of shopping sites, spammers can pose as different customers buying products and writing fake reviews making them harder to eradicate completely.

B. Spamming Store Detection

In this paper, we address the problem of detecting spamming stores or finding sources of fake reviews. Unlike the methods of fake review detection, our proposed detection method is store centric. Like the methods of review spammer

¹ <http://www.ftchinese.com/story/001058005>

² <http://www.forbes.com/sites/gordonchang/2014/11/30/chinas-cyber-monday-fake-sales-does-alibaba-have-a-sustainable-business/>

³ <http://blogs.wsj.com/chinarealtime/2015/03/03/cat-and-mouse-game-alibaba-exec-on-fake-transactions>

detection in [2], our method is also store behavior driven. A store centric method is preferred over the review centric method as collecting behavioral evidence of stores is easier than that of fake reviews. A review usually includes the limited information about only one reviewer and one product. However, a store can involve more information about deals, reviewers and reviews, and the likelihood of finding evidence against spammers will be much higher if we consider stores than reviews. At the same time, the goal of a store centric approach is to find the source of spam reviews. Although reviewer centric approach can find and remove the review spammers who are direct source of spammed reviews, spamming stores will hire other reviewers to post spammed reviews until they are forbidden.

The store centric method is also scalable as one can always need new spamming behaviors. The main step of the spamming behavior detection is to model the spamming behavior on the basis of different spamming behavior patterns. By measuring the extent to which the store has a certain spamming behavior, each store is assigned a numeric score of spamming behavior. In this paper, we depend on patterns of review and transaction to get four spamming behavior models, i.e., (a) single user multiple deals (targeting deal or TD) behavior; (b) single user multiple reviews (targeting review or TR) behavior; (c) deal deviation (DD) behavior; and (d) review deviation (RD) behavior. To get an overall spamming score of each store, we combine the spam scores of the store's different spamming behaviors using linear weighted combination [2]. We empirically get the weights of the different spamming scores.

Like the model in [2], our proposed behavior method avoids high computational costs of deep text analysis and opinion extraction. This paper also takes a scoring method to detect spamming stores by examining their behaviors and derived features. Some text analysis methods can do well in opinion extraction of review text, and we can extend our spamming behavior models by considering review text analysis. Because of considering higher computational costs of content analysis, we will leave this work to our future research. In [2], the method of review spammer detection is effective and outperforms other baseline method based on helpfulness votes alone. We believe that the performance of our model will be improved by considering results of the review spammer detection, but this will also be left to our future research.

This paper is organized as follows. The related work is presented in Section 2. Our proposed spamming store detection method, the data representation and experiment in this research are described in Section 3. Section 4 presents user evaluation. Section 5 concludes this paper.

II. RELATED WORK

In this section, we survey the relevant work in spamming store detection and compare them with our approach presented in this paper.

A. Review Opinion and Sentiment Mining

In [4], Minqing Hu and Bing Liu summarize all the customer reviews of a product and try to mine the review opinion from the customer reviews. In [5], Ana-Maria Popescu and Oren Etzioni introduce an unsupervised information extraction system to extract important product features. These work focus mainly on text mining and sentiment mining of review content. We can use those extracted opinions to identify outlier of opinions in customer reviews, but those opinion outlier could be truth opinions and they do not represent spam opinions. So the first task of review opinion and sentiment mining is to detect fake review.

B. Review Spam Detection

A pilot study was reported in [6]. Jindal and others get the training data set by identifying the duplicate and near duplicate reviews. Then construct the machine learning models to classify the reviews. The reviews are classified as spam reviews and non-spam reviews [6] by using duplicate reviews as labeled data. Duplicate detection is done using the Shingle method, and the review is classified as duplicate review when the similarity score is great than 0.9. They used the logistic regression to build learning model. Using only review contents is very hard to detect fake reviews manually, so the data set is hard to label.

Based on above foundation, fake reviews are classified in three categories, that is, untruthful reviews, reviews on brands only and non-reviews (e.g., advertisements). Considering the review content has limited features, authors collect 24 features of text of review, reviewer, products [1]. In [7], Liu proposes some features deriving from reviews content that correspond to subjectiveness, informativeness and readability aspects of review.

If a review is the only review a reviewer has written, we call it a singleton review, and call the reviewer singleton reviewer. Wu and other authors attempted to detect fake singleton reviews by identifying singleton reviewers [8]. They present an evaluation of distortion strategy in the assessment of shill detection mechanisms on a dataset of hotel reviews collected from TripAdvisor using the features of singleton reviewers. This is distortion based method. In [9], the fake singleton review detection problem is mapped to an abnormally correlated pattern detection problem. Authors propose a hierarchical algorithm to robustly detect the time windows where such fake review attacks are likely to have happened. This algorithm is based on multi-scale multidimensional time series anomaly detection.

C. Review Spammer Detection

If a user is reviewer who make the fake reviews, his reviews are more likely to be fake reviews. Reviewer centric method is a method based on detecting the reviewer who writes the fake reviews, and it is a behavior-driven method. Detecting fake product reviewer is easier than detecting the content of review because a review only involves a user and a product and have less amount of information.

Ratings are usually supported when customers post their reviews. [8, 11] These work survey some research on unfair or

deceptive ratings [8,11]. However, these work did not address review spammer detection directly.

Jindal and others explore the possibility of finding the suspicious behaviors of reviewers by identifying unusual review patterns [10]. For example, if a reviewer post all positive reviews on all products of a company but other reviewers are generally negative about them, this reviewer's behavior is clearly suspicious.

In [2,5,8,10,12], authors make some research on detecting the fake product reviewer. Lim and other authors relied on patterns of review content and ratings to define four different fraud behavior models [2] as follows:

- targeting product
- targeting group
- general rating deviation
- early rating deviation

They proposed a scoring method to measure the degree demonstrating fraud behaviors by combining the every fraud scores of the reviewer's different fraud behaviors.

In [13], Knorr and Ng view review spammer detection as outlier detection. In some sense, the spammers behaving differently from other normal reviewers are "outliers". The spamming stores behaving differently from other normal stores are outliers too.

III. DETECTION METHOD

A store is expected to monitor targeted deals and reviews closely. We thus define two spamming behaviors involving targeted deals and reviews scores for each store suggesting the extent to which conduct the behaviors.

A. Notations

In the following, we firstly list the notations to be used in the subsequent sections.

- $U = \{u_i\}$: set of users
- $D = \{d_{ij}\}$: set of all deals by user j in store i
- $V = \{v_i\}$: set of all stores
- $R = \{r_k\}$: set of reviews r_k 's. The review r_k comes with a rating e_k
- $E = \{e_k\}$: set of ratings e_k 's ($e_k \in [0, 1]$)
- $u(r_k) = u(e_k)$: user of review r_k and rating e_k
- $d(r_k) = d(e_k)$: deal of review r_k and rating e_k
- $R_{ij} = \{r_k | u(v_k) = u_j \wedge v(v_k) = v_i\}$: set of reviews from user v_j in store v_i
- $E_{ij} = \{e_k | u(e_k) = u_j \wedge v(e_k) = v_i\}$: set of ratings from user u_j in store v_i
- $D_i^* = \bigcup_j D_{ij}$: set of all deals in store v_i

B. Deal Spamming

1) Multiple Deals Single Customer: As people having herd mentality, a customer usually buy product from a store that sold a large number of goods, instead of from the store that sold only a few goods after comparison of two stores. Some owners of store may utilize this to sell their products, and they may pay some people to buy products that are returned soon afterwards, which result in some false appearance that some commodities looks very popular. There is less possibility that

a customer buy products many times in a short time, e.g., buying clothing for many times from the same clothing store in one month. Such an assumption is reasonable. We get a score from this kind of suspicious behavior. In the dataset shown at section 5, 306 store-reviewer pairs involve multiple deals. This number is considered very small given the much larger numbers of reviewers and stores.

We easily observe a spamming behavior targeting a specific store observed by finding the number of deals from the same customer on the same store especially when those deals occur in the short time, e.g. Figure 1. There are several deals that are contributed by the same customers to the same stores ($|D_{ij}| \geq 2$).

The screenshot shows a feedback page for a seller with 711 reviews. It displays a list of transactions with details such as the buyer's name, the product, the price, and the date. Several transactions are highlighted with a red star icon, indicating a 5-star rating. The transactions are sorted by date, showing a sequence of purchases from the same customer over a short period.

Buyer	Transaction Details	Feedback
K*** B	2014 Hot Sale New Arrival Women Sne... US \$49.00 x 1 pair	★★★★★ 13 Nov 2014 10:12
D*** M	2014 Hot Sale shoes woman SK shoes... US \$49.00 x 1 pair	★★★★★ 13 Nov 2014 10:11
E*** S	2014 New Arrival shoes men Valid Me... US \$44.20 x 1 pair	★★★★★ 13 Nov 2014 10:11
Y*** Z	2014 New arrival Women Sneakers Lac... US \$58.00 x 1 pair	★★★★★ 13 Nov 2014 10:11
I*** R	2014 New Autumn Women Pumps US \$52.00 x 1 pair	★★★★★ 13 Nov 2014 10:11
L*** C	2014 New Oxford Shoes for Women Lea... US \$26.50 x 1 pair	★★★★★ 13 Nov 2014 10:11
K*** B	2014 Hot Sale New Arrival Women Sne... US \$49.00 x 1 pair	★★★★★ 26 Oct 2014 10:09
E*** Z	2014 New Oxford Shoes for Women Lea... US \$24.50 x 1 pair	★★★★★ 26 Oct 2014 09:59
K*** B	2014 Hot Sale New Arrival Women Sne... US \$49.00 x 1 pair	★★★★★ 24 Oct 2014 08:44

Figure 1. Multiple Deals Single Customer

Stores which are involved in store-reviewer pairs with larger number of deals are likely to be spamming stores, especially when their features (e.g. ratings, reviews) are not very different from one another. The following store spam score function $C_d(v_i)$ for store v_i based on deal spamming behavior on some targeted stores is defined:

$$C_d(v_i) = \frac{s_i}{\max_{v'_i \in U} s_{i'}} \quad (1)$$

where s_i represents the unnormalized score of spamming store v_i defined by:

$$s_i = \sum_{d_{ij} \in D_{ij}, |D_{ij}| > 1} |D_{ij}| \quad (2)$$

The denominator of (1) is the maximum s_i among all stores.

Based on the above function (1), stores with large proportions of deals involving multiple-deals customers are to be assigned high spam scores.

2) Deal Deviation: In general, a store's quantity of sales will fluctuate in a reasonable range. If it goes too much out of range, this store is likely to be conducting a spamming behavior. In order to detect such behavior, we view the average deals in a period of time as a reasonable deals, difference between the genuine sales and the average sales as a deviation and average deviation as a measure of spamming behaviors.

To model this kind of behavior that involves very large deviation of quantity of sale on products having the same

attribute in a span of time, the whole time period is divided into some fixed-size time windows and sets of big deviation are derived. We empirically choose a day interval as time window in our experiments. This interval can't be too large so that we can't capture the deal burstiness and can't be too small so that we correctly describe the deviation of quantity of deals and reviews. Such sets are suspicious of spamming product deal. We define the spam score of a store v_i in time window w as follows:

$$s_i^d = \text{Avg } ||D_{i*}(w)| - M_i| \quad (3)$$

Where M_i is the mean of $D_{i*}(w)$. We thus define the spam score of a store v_i based on general deviation using the average deviation of quantity of deals as follows:

$$C_d(v_i) = \frac{s_i^d}{\text{Max}_{v_i' \in U} s_i^{d'}} \quad (4)$$

The denominator of (4) is the maximum s_i^d among all stores.

C. Review Spamming

1) Multiple Reviews Single Customer: A customer could be hired to simply buy products, and other ones could be hired to post positive reviews after buying products for more payment. They possibly post multiple reviews on the same store (see Figure 2).

Buyer	Transaction Details	Feedback
K***y B. [B]	2014 Hot Sale New Arrival Women Sneakers US \$49.00 x 1 pair	★★★★★ 13 Nov 2014 10:12
D***I M. [B]	2014 Hot Sale shoes woman 3K shoes US \$49.00 x 1 pair	★★★★★ 13 Nov 2014 10:11
E***o S. [B]	2014 New Arrival shoes men Valid Me... US \$44.20 x 1 pair	★★★★★ 13 Nov 2014 10:11
Y***a Z. [B]	2014 New arrival Women Sneakers Lac... US \$58.00 x 1 pair	★★★★★ 13 Nov 2014 10:11
I***n R. [B]	2014 New Autumn Women Pumps US \$52.00 x 1 pair	★★★★★ 13 Nov 2014 10:11
L***a C. [B]	2014 New Oxford Shoes for Women Lea... US \$26.50 x 1 pair	★★★★★ 13 Nov 2014 10:11
K***y B. [B]	2014 Hot Sale New Arrival Women Sneakers US \$49.00 x 1 pair	★★★★★ 25 Oct 2014 10:09
E***a Z. [B]	2014 New Oxford Shoes for Women Lea... US \$24.50 x 1 pair	★★★★★ 25 Oct 2014 09:59
K***y B. [B]	2014 Hot Sale New Arrival Women Sneakers US \$49.00 x 1 pair	★★★★★ 24 Oct 2014 08:44

Figure 2. Multiple Reviews Single Customer

Firstly, such reviews involved similar products from the same store (e.g., clothing with different brands, TVs with different sizes) are likely to be identical or look similar so as to conserve spamming efforts. Secondly, there are some customers who post multiple reviews for the same products in different time window for the same reason. For instance, some reviewers make minor changes to their reviews from the same store. So, we need to distinguish this type of reviews from the more suspicious ones. We introduce the similarity function between two reviews r_k and $r_{k'}$, $\text{sim}(r_k, r_{k'})$, by:

$$\text{sim}(r_k, r_{k'}) = \text{cosine}(r_k, r_{k'}) \quad (5)$$

where we use a bag of bi-grams to denote respective review and the cosine similarity of the bi-gram TFIDF vectors of v_k and $v_{k'}$ is represented by $\text{cosine}(v_k, v_{k'})$. When $\text{cosine}(v_k, v_{k'}) = 1$, the two reviews are considered identical. TFIDF equals to the product of the frequency and inverse document frequency

of a bi-gram. Given a set of reviews R_{ij} ($|R_{ij}| > 1$), we derive a similarity score $\text{sim}(R_{ij})$ as follows:

$$\text{sim}(R_{ij}) = \text{Avg}_{\substack{r_k, r_{k'} \in R_{ij}, \\ k < k'}} \text{sim}(r_k, r_{k'}) \quad (6)$$

Then, the store spam score function $C_r(v_i)$ for store v_i based on review spamming behavior is defined as follows:

$$C_r(v_i) = \frac{s_i^r}{\text{Max}_{v_i' \in U} s_i^{r'}} \quad (7)$$

Where the denominator is the maximum s_i^r among all stores, and s_i^r represents the unnormalized score of spamming store v_i , we define it as:

$$s_i^r = \sum_{r_{ij} \in R_{ij}, |R_{ij}| > 1} |R_{ij}| \cdot \text{sim}(R_{ij}) \quad (8)$$

2) Review Deviation: for the reason similar to above, a store's quantity of reviews will fluctuate in a reasonable range. In order to detect such behavior, we view the average quantity of review in a period of time as a reasonable one, difference between the genuine quantity of reviews and the average one as a deviation and average deviation as a measure of spamming behaviors.

To model this spamming behavior, like modeling the deal deviation, the whole time period is divided into some fixed-size time windows and sets of big deviation are derived. Such sets are suspicious of spamming product review. We define the spam score of a store v_i in time window w as follows:

$$s_i^{d'} = \text{Avg } ||R_{i*}(w)| - M_i'| \quad (9)$$

Where M_i' is the mean of $R_{i*}(w)$. We thus define the spam score of a store v_i based on general deviation using the average deviation of quantity of reviews as follows:

$$C_r^d(v_i) = \frac{s_i^{d'}}{\text{Max}_{v_i' \in U} s_i^{d'}} \quad (10)$$

The denominator of (10) is the maximum $s_i^{d'}$ among all stores.

D. Combined Behavior

The spamming store detection methods are based on: TD, DD, TR and RD. At the same time, we introduce a new combined method (ALL) that combines all the behaviors:

$$c(v_i) = \frac{1}{3} C_d(v_i) + \frac{1}{6} C_d^d(v_i) + \frac{1}{3} C_r(v_i) + \frac{1}{6} C_r^d(v_i) \quad (11)$$

We empirically give the same weight to deal spamming and review spamming, smaller one to deviation behaviors as deviations have generally weaker evidences.

IV. USER EVALUATION

A. Store Dataset

To evaluate the above score function, we crawl the data from AliExpress website to get AliExpress Dataset. In this study, each store has one or more products, each store also has some profile attributes, e.g. store name, store rating and type of store. Each product of store has information of deal, review and reviewer. Each product may also have product attributes depending on the product type. Taking clothing for example,

the relevant attributes include size, color, style, and price. In AliExpress website, you can leave feedback for suppliers within 30 days of order completion. The feedback includes a rating for seller and a review.

As AliExpress just show specific transactions and reviews in six months, we randomly crawl some information about some clothing stores in a period of time.

B. Preprocessing

Several steps of data preprocessing are performed on the above AliExpress dataset before experiment is done.

- Processing of anonymous customers: in Figure 1, a lot of customers can leave their feedback as anonymous ones, but the customers with same ratings, nationality and part of name are very likely to be the same one. It is reasonable that we view those customers as identical one.
- Combination of deals: deals where two or more products can be bought by the same customer at the same time are reasonable ones, but they should be considered as one normal one instead of spamming Multiple Deals Single Customer (see Section III-B-1).
- Removal of non-main products: there are usually big differences between features of main and non-main products in certain store, so it is very difficult to compare the deals and reviews of main and non-main products. We remove the non-main products with relevant information.
- Removal of non-English and negative reviews: we focus on the positive reviews and English reviews.
- Removal of inactive stores: To focus on active stores, our dataset includes only stores with more than 3 reviews and more than 10 deals.

Finally, we get the preprocessed dataset with statistics (see Table 1).

TABLE 1. DATASET STATISTICS

	Number
V: set of stores	33
R: set of reviews	6432
D: set of transactions	22345

C. Evaluation Methodology

Following [2], to evaluate the performance of the above solution methods, we need know the truth labels of evaluated stores. But those labels do not exist, so we decide to use human evaluation on different scoring methods representing the extent to spamming behavior. Three undergraduate students are recruited as human evaluators, and then they need to be trained to ensure that they provide high quality result. The result of evaluation is analysed to derive the truth labels.

Given that labels are given by human evaluation, subsequently, we can examine if the proposed methods can accurately detect spamming store and the extent to that. We rank the stores by decreasing order of behavior score. The highly ranked stores are more likely to be spamming ones.

When a store's spam score is more than a threshold, this store is considered as spamming store.

Like the evaluation of the relatively small sample in information retrieval field, NDCG method [2, 8] is used to evaluate the performance of a method's ranking by spamming behavior.

D. Experimental Setup

The experiment includes the following steps:

- For each store, we choose its relevant information to show the three evaluators. The information is selected based on their involvement in the spamming behaviors identified. Specifically, we select: (a) deals by the evaluated store having same buyer. (b) the quantity of deals by the evaluated store having more deals deviated from the average of deals of the evaluated store in time window w . (c) Reviews that have high ratings with other reviews of the evaluated store or that are exact (or near-exact) duplicates of other reviews of the evaluated store. (d) the quantity of reviews by the evaluated store having more reviews deviated from the average of reviews of the evaluated store in the time window w . For stores with less than 20 reviews or 20 deals under the above four categories, we randomly select other reviews and deals to fill the shortfall.
- Three human evaluators are chosen and trained to examine the selected stores and their selected reviews and deals. In order to more accurately ensure that evaluation is effective, each evaluator independently labels all evaluated stores. The spamming store label decisions of each evaluator are recorded and suspicious behaviors of each store are marked by the evaluator. Each store is labeled either "spamming store" or "non-spamming store" by each evaluator. Evaluators are not informed about the number of spamming stores to be labeled.
- After getting labels, we can calculate the Normalized Discounted Cumulative Gain (NDCG) [14] of each method. The discounted cumulative gain (DCG) of a ranked list of stores is defined as:

$$DCG = \sum_{p=1}^c \frac{f(i_p)}{\log_2(1+p)} \quad (12)$$

where $f(i_p)$ equals to the number of votes v_{i_p} received ($f(i_p) \in [0,3]$). c is the number of evaluated stores. NDCG is simply DCG normalized by the DCG of the ideal rank order of the stores where stores which get more votes rank before the ones which get less votes (e.g. the store having three votes rank before other ones).

$$NDCG = \frac{DCG}{DCG_{ideal}} \quad (13)$$

Where DCG_{ideal} is the DCG of the ideal rank order of the stores. NDCG determines how well a method's rank the order of stores following the ideal rank order by giving more weight (determined by $\log_2(1+p)$ in DCG) to the stores with higher importance (determined by $f(i_p)$ in DCG).

E. Results

Table 2 shows the number of spamming stores and non-spamming stores labeled by the evaluators. The number in the diagonal cells is the one labeled by single evaluator, and the overlapping number is the one labeled by pair of evaluators. All the three evaluators agree that 3 stores are labeled as “spamming store” and 26 stores are labeled as “non-spamming store”. The evaluators are thus largely consistent in their judgements of spamming and non-spamming stores.

TABLE 2. EVALUATION RESULTS

	Evaluator 1	Evaluator 2	Evaluator 3
#Spamming store			
Evaluator 1	6	4	4
Evaluator 2	-	5	4
Evaluator 3	-	-	5
#Non-spamming store			
Evaluator 1	27	26	26
Evaluator 2	-	28	27
Evaluator 3	-	-	28

A store is assigned final “spamming store” label if this label is agreed by majority of evaluators. Based on above results and majority voting, 6 stores finally are labeled as “spamming store” and 27 ones are labeled as “non-spamming store”. Table 3 show the number of top 5 stores with the final spamming store labels, and the number of bottom 5 stores with non-spamming store labels for TD, TR, DD, RD and ALL approaches. The results show that the ALL and TD approaches correctly rank the spamming stores at the top rank, ALL and TR approaches correctly rank the non-spamming stores at the bottom rank.

TABLE 3. RESULTS OF TOP 5 AND BOTTOM 5 RANKED REVIEWERS

	TD	TR	DD	RD	All
# spamming stores in top 5	4	3	2	3	4
coverage rate in top 5	80%	60%	40%	60%	80%
# non-spamming stores in bottom 5	4	5	4	3	5
coverage rate in bottom 5	80%	100%	80%	60%	100%

NDCG is used to examine the accuracy of TD, TR, DD, RD and ALL methods on finding spamming stores. We calculate some sets of NDCG value when top c positions in the rank list produced by each method change from 1 to 33. As shown in table 4, ALL, TD, TR outperform other methods on rank orders.

TABLE 4. NDCG RESULTS

c	TD	TR	DD	RD	All
5	0.83	0.80	0.66	0.70	0.84
10	0.84	0.84	0.59	0.66	0.85
15	0.81	0.80	0.65	0.57	0.83
20	0.93	0.92	0.76	0.77	0.90
25	0.90	0.91	0.82	0.81	0.92

30	0.92	0.90	0.84	0.83	0.91
33	0.92	0.93	0.83	0.84	0.94

In summary, the results show that our method is effective as indicated by the inter-evaluator consistency in judging spamming stores and non-spamming stores, and by the higher NDCG values for the our methods.

V. CONCLUSION AND FUTURE RESEARCH

We propose a behavior scoring method to detect spamming stores who produce fake reviews, deals or/and rankings on some products of these stores. We rank stores by the scores denoting their spamming behaviors. To evaluate methods proposed, we conduct human evaluation on an AliExpress dataset containing data of clothing stores. Result shows that our proposed methods are generally effective.

In our future work, we will integrate spamming store detection into spam review detection and vice versa.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2013R1A1A2013401).

REFERENCES

- [1] N. Jindal and B. Liu, “Opinion spam and analysis,” in *Proc. WSDM*, 2008.
- [2] E.P. Lim, V.A. Nguyen, and N. Jindal, “Detecting product review spammers using rating behaviors,” in *Proc. the 19th ACM international conference on Information and knowledge management*, pp. 939-948, ACM, 2010.
- [3] E. Gilbert and K. Karahalios, “Understanding deja reviewers,” in *Proc. CSCW*, 2010.
- [4] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *Proc. KDD*, 2004.
- [5] A. Popescu and O. Etzioni, “Extracting product features and opinions from reviews,” in *Proc. EMNLP*, 2005.
- [6] N. Jindal and B. Liu, “Review spam detection,” in *Proc. WWW (poster)*, 2007.
- [7] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou, “Low-quality product review detection in opinion summarization,” in *Proc. EMNLP-CoNLL*, 2007.
- [8] G. Wu, D. Greene, B. Smyth, and P. Cunningham, “Distortion as a validation criterion in the identification of suspicious reviews,” University College Dublin, Tech. Rep. UCD-CSI-2010-04, 2010.
- [9] S. Xie, G. Wang, S. Lin, and P.S. Yu, “Review spam detection via temporal pattern discovery,” in *Proc. KDD*, 2012.
- [10] N. Jindal, B. Liu, and E.-P. Lim, “Finding unusual review patterns using unexpected rules,” in *Proc. CIKM*, 2010.
- [11] C. Dellarocas, “Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior,” in *Proc. ACM EC*, 2000.
- [12] A. Mukherjee, B. Liu, and N. Glance, “Spotting fake reviewer groups in consumer reviews,” in *Proc. the 21st international conference on World Wide Web*, pp. 191-200, ACM, 2012.
- [13] E. Knorr and R. Ng, “A unified notion of outliers: Properties and computation,” in *Proc. KDD*.
- [14] K. Järvelin and J. Kekäläinen, “IR evaluation methods for retrieving highly relevant documents,” in *Proc. SIGIR*, 2000.