Detecting the Spam Review Using Tri-training

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—**Some supervised learning methods were developed to detect spam review and some of them are considerably effective. Some researchers also find that the review spammer consistently produce spam reviews. We observe that the spamming store also consistently produce spam reviews. This provides us two other views to identify review spam: we can identify if the reviewer is spammer and if the store is spamming one. We introduce a three-view semi-supervised method, tri-training, to exploit the large amount of unlabeled data. The experiment results demonstrate that three-view tri-training algorithm can achieve better results than two-view co-training and single-view algorithm.**

Keywords—**deceptive reviews, semi-supervised learning,**

supervised learning, tri-training

I. Introduction

In most of the E-commerce sites, review section is provided for costumers so that they can write reviews of products at these sites and express their views. In the past few years, people have a lot of interest in mining opinions existing in reviews due to many popular applications (i.e. Amazon.com). Though these reviews support important information to us, they have no quality control in some E-commerce sites, so anyone can write fake reviews and mislead potential customers in making their choices and in buying low-quality products, but manufacturers with good reputation can be defamed by malicious reviews. In order to protect customers, manufacturers and the whole e-commerce environment, it is necessary to take required measures to detect and remove fake reviews.

Some researchers start to study this problem [1] [2]. Some supervised learning methods utilize the features of the review content to detect spam reviews and some of them are considerably effective. Other methods try to find the review spammer to detect spam review. In C2C E-commerce sites, stores are the main source of deceptive behaviours including spamming reviews, spamming reputation, spamming deals and spamming ratings. If a spamming store consistently pays someone to conduct these deceptive behaviours, researchers can detect spamming stores to find spam review.

Three problems, detecting spam review, detecting spam reviewer and detecting spamming stores, have their own respective methods and shortcomings. Less information about reviews and simulation of genuine customers are the biggest problem in review spam detection. Review spammer detection methods assume that all the reviews posted by review spammers are fake reviews. But some review spammers could post genuine reviews. In the spamming stores, some reviews posted by some genuine customers could support useful information. If we don't find integrated approaches to solve those problems, each of the problems will be considered separately, in spite of their mutual relevance.

Consequently, the three problems are closely related because solving one may help to solve the other two. For instance, detecting a positive review as a spam can help us find relevant review spammers and spamming stores. In the same way, detecting a spamming can help us find review spams and review spammers.

Based on above observations, we find three views to identify review spams. We thus introduce a three-view semi-supervised method, tri-training, to detect spam review.

II. Related Work

A. Review Spam Detection

A pilot study has been reported in [4]. Jindal N and others get the training data set by identifying the duplicate and near duplicate reviews. Then they construct the machine learning models to classify the reviews. The reviews are classified as spam reviews and non-spam reviews [4] by using duplicate reviews as labelled 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 it is difficult to label the data set.

Integrating work from psychology and computational linguistics [5] [6], Ott et al. [3] develop three approaches to detect deceptive opinion spam, and ultimately develop a classifier that is nearly 90% accurate on their dataset.

Feng et al. [7] observe that the features of CFG (context free grammar) of review text are useful on improving the performance of detection spam review.

B. Review Spammer Detection

If a user is a reviewer who makes the fake reviews, his other 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 behaviour-driven method. Detecting fake product reviewer is easier than detecting the content of review because a review only involves a user and a product, so it has less amount of information.

Jindal N and others explore the possibility of finding the suspicious behaviors of reviewers by identifying unusual review patterns [15]. For example, if a reviewer posts all positive reviews on all products of a company but other reviewers are generally negative ones, this reviewer's behavior is clearly suspicious.

In [8], authors make some researches on detecting the fake product reviewer. Lim E-P and other authors relied on patterns of review content and ratings to define four different fraud behaviour models to detect spam reviewers.

C. Spamming Store Detection

We focus on suspicious behaviours of stores to detect spamming stores. 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. We propose some 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.

D. Semi-supervised Learning

Li et al. [9] use a semi-supervised algorithm, co-training, to detect spam review, and achieve the most accurate result among semi-supervised algorithms. But, this method doesn't consider the features of deep syntax and psychological linguistics of review text, which are proved to be effective on detecting deceptive opinions.

Zhou et al. [10] propose a new co-training style semi-supervised learning algorithm, named tri-training algorithm. Tri-training neither requires the instance space to be described with sufficient and redundant views nor does it put any constraints on the supervised learning algorithm, its applicability is broader than that of previous co-training style algorithms. Unlike co-training classifiers, tri-training algorithm uses three classifiers standing three views to improve the performance. These classifiers are then refined using unlabelled examples in the tri-training process.

III. Detection Method

A. Supervised Learning

In [3], Ott et al. achieve higher performance with SVM than with Naïve Bayes, so we use SVM as a baseline. At the same time, we create the model using psychological linguistics [11], unigrams, Bigrams [12] and deep syntax [7] features of review text.

B. Co-training Method

As it is a time-consuming task to manually label spam reviews, we only label a small set of review data. There are still a large number of unlabelled data, which may improve the performance of detection. To exploit the unlabelled data, we introduce a novel semi-supervised algorithm, tri-training, to detect spam review. For comparison, we also introduce co-training (see Algorithm 1) [13]. Li et al. [9] has verified that spam reviewers consistently write spam reviews and 85% of reviews that are spam. This finding supports us two views to identify the spam reviews: the first view is to directly detect whether a review is spam or not using the detection method described in the previous section; the second view is to detect whether the reviewer is a spammer or not using the detection method in [5]. In practice, the assumptions of conditional independent views may be not satisfied, therefore we use "agreement" strategy [14] to resolve this problem. This strategy needs to change $\cup$ to $\cap$ in the last step. We only select the *p* positive instances and *n* negative instances, when the two view classifiers agree most.

Algorithm 1 Co-Training Algorithm

```
Input: Two views of feature sets: review features $F_r$ and reviewer features $F_u$; the labelled training set $L$; the unlabelled training set $U$
Process: Loop for $k$ iterations:
    Step 1: Use $L$ to train a classifier $C_r$ from $L$ based on view $F_r$
    Step 2: Use $L$ to train a classifier $C_u$ from $L$ based on view $F_u$
    Step 3: Allow $C_r$ to choose $p$ positive and $n$ negative most confidently predicted reviews $R$ from $U$
    Step 4: Allow $C_u$ to choose most confidently predicted reviewers $P$ from $U$
    Step 5: Extract $p$ positive and $n$ negative the reviews $R'$ posted by $P$
    Step 6: Move reviews $R \cup R'$ from $U$ to $L$ with predicted labels
```

C. Tri-training Method

Co-training algorithm requires two sufficient and redundant views. In the strict sense, above two views can't completely satisfy this requirement. Therefore, we introduce a three-view semi-supervised method, tri-training, for review spam detection. Tri-training method starts with a set of labelled data, and increases the amount of annotated data by adding unlabelled data incrementally. In the context of review spam identification, each review has three types of features: features about reviews themselves, features about corresponding reviewers and features about corresponding stores. Tri-training method is shown in Algorithm 2.

Algorithm 2 Tri-Training Algorithm

```
Input: three set of features: review feature $F_r$; reviewer feature $F_u$; store feature $F_s$; labelled review set $L$; unlabelled review set $U$; number of iteration $I$;
Output: three classifiers;
Iteration process: (until all unlabelled reviews are labelled):
    Step 1: choose a subset $U'$ from $U$;
    Step 2: learn a classifier $C_r$ based on $F_r$;
    Step 3: use $C_r$ to label $U'$, marked by $U_r$;
    Step 4: learn a classifier $C_u$ based on $F_u$;
    Step 5: use $C_u$ to label $U'$, marked by $U_u$;
    Step 6: learn a classifier $C_s$ based on $F_s$;
```

Step 7: use C_s to label *U'*, marked by U_s;

Step 8: based on U_r, U_u, U_s, give each item of *U'* the final label using the weighted voting rule;

Step 9: move reviews of *U'* to *L* with predicted labels.

Weighted voting rule: considering the difference among three origin classifiers is considerably big, we employ weighted voting rule to finally label unlabelled data in Step 8. We assign weight as the probability that each classifier correctly classify the data of origin *L*, denoted by $P_i(L)$ as follows:

$$C(x) = a\,\text{rg}\max_{y \in label} \frac{\sum_{i=1}^{3} f(y, C_i(x)) \times P_i(x)}{\sum_{i=1}^{3} P_i(x)} \quad (1)$$

$$Where \quad f(y, C_i(x)) = \begin{cases} 0, & C_i(x) \neq y \\ 1, & C_i(x) = y \end{cases}$$

IV. Experiment

A. Setup

From our AliExpress dataset, we use human evaluation and majority voting strategy to label the data. Three human evaluators are trained to detect if a review is a spam. They independently label all evaluated reviews. Evaluators are not informed about the number of spam reviews to be labelled. Finally, we get 612 spam reviews of 2321 reviews.

For our supervised methods, we need to divide the data set into a training set and a test set. We conduct 10-fold cross-validation: the data set is randomly split into ten folds, where nine folds are selected for training and the tenth fold is selected for test. We apply our co-training and tri-training method on the same test data set as the supervised methods, for comparison.

The evaluation metrics are precision, recall and f-score. We try several thresholds for detecting spam reviewers and spamming stores, and select the best threshold based on the training data set.

B. Results and Analysis

Table 1 shows the experiment results. When we use supervised method (SVM) that uses review content features as a single view, the obtained f-score is 0.59. Co-training (Agreement) regard review content and reviewer features as two views and select the unlabeled data with the most agreement for two view classifiers. The f-score from co-training is 0.65. Tri-training adds the third view: spamming store view. The f-score from tri-training is higher than that of co-training. Tri-training is suitable for review spam detection. It achieves more accurate result than other supervised and semi-supervised methods.

Table 1 Experiment Result

	Precision	**Recall**	**F-score**
SVM	0.61	0.57	0.59
co-training	0.63	0.68	0.65
tri-training	0.71	0.69	0.70

V. Conclusion and Future Research

In this paper, we make a survey on the review spam identification. We start our research based on the fact that the spammer consistently writes spams. This provides us another view to identify review spam. At the same time, we find that the spamming stores produce a large amount of spam reviews. This provides us the third view to identify review spam. Based on these observations, we introduce a three-view semi-supervised method to exploit the large number of unlabelled data. The experiment results show that the three-view tri-training algorithm can achieve more accurate results than the two-view and single-view algorithm.

In future work, we will introduce more on the multi-view learning methods in review spam detection.

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] B. Liu, “Sentiment analysis and opinion mining,” Synthesis Lectures on Human Language Technologies, 2012.

[2] N. Jindal and B. Liu, “Opinion spam and analysis,” in *Proc. the 2008 International Conference on Web Search and Data Mining*, 2008.

[3] M. Ott, Y. Choi, C. Cardie, “Finding deceptive opinion spam by any stretch of the imagination,” in *Proc. the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011.

[4] N. Jindal and B. Liu, “Review spam detection,” in *Proc. the 16th international conference on World Wide Web*, 2007.

[5] J.T. Hancock, L.E. Curry, S. Goorha and M. Woodworth, “On lying and being lied to: A linguistic analysis of deception in computer-mediated communication,” Discourse Processes, 2007.

[6] M.L. Newman, J.W. Pennebaker, D.S. Berry, J. M. Richards, “Lying words: Predicting deception from linguistic styles,” Personality and social psychology bulletin, 2003.

[7] S. Feng, R. Banerjee, Y. Choi, “Syntactic stylometry for deception detection,” in *Proc. the 50th Annual Meeting of the Association for Computational Linguistics*, 2012.

[8] E.P. Lim, V.A. Nguyen, N. Jindal, B. Liu, “Detecting product review spammers using rating behaviors,” in *Proc. the 19th ACM international conference on Information and knowledge management*, 2010.

[9] F. Li, M. Huang, Y. Yang, X. Zhu, “Learning to identify review spam,” *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 2011.

[10] Z.H. Zhou, M. Li, “Tri-training: Exploiting unlabeled data using three classifiers,” *Knowledge and Data Engineering*, 2005.

[11] K.H. Yoo, U. Gretzel, “Comparison of deceptive and truthful travel reviews,” *Information and communication technologies in tourism,* 2009.

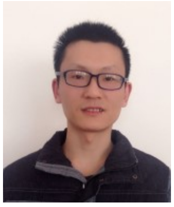
[12] R. Mihalcea, C. Strapparava, “The lie detector: Explorations in the automatic recognition of deceptive language,” in *Proc. the ACL-IJCNLP 2009 Conference Short Papers*, 2009.

[13] A. Blum, T. Mitchell, “Combining labeled and unlabeled data with co-training,” in *Proc. the eleventh annual conference on Computational learning theory*, 1998.

[14] M. Collins, Y. Singer, “Unsupervised models for named entity classification,” in *Proc. the joint SIGDAT conference on empirical methods in natural language processing and very large corpora*, 1999.

[15] N. Jindal, B. Liu, E.P. Lim, “Finding unusual review patterns using unexpected rules,” in *Proc. the 19th ACM international conference on Information and knowledge management*, 2010.

[16] D.J. Miller, H.S. Uyar, “A mixture of experts classifier with learning based on both labelled and unlabelled data,” *Advances in neural information processing systems*, 1997.



Ji Chengzhang He received the B.S. degree in computer science and technology from Linyi University, Linyi, Shandong Province, China in 2005, and received the M.S. degree in computer application technology from Ocean University of China, Qingdao, Shandong Province, China in 2010. He had been working as an Assistant or Lecturer in Weifang University of Science & Technology, China from 2005 to 2015. Currently he is a doctoral candidate in machine learning at Dongseo University, Korea. His research interests include machine learning, data mining, spam review detection.

Dae-Ki Kang is an Associate Professor at Dongseo University in Korea. He was a Senior Member of the Engineering Staff at the Associated Institute of Electronics & Telecommunications Research in South Korea. He earned a PhD in Computer Science from Iowa State University in 2006. He received a Master of Science degree in computer science from Sogang University in 1994 and a BE in Computer Science and Engineering from Hanyang University in 1992.