# Detection of Spammer by using Rating Behaviors

Chan Myae Aye
*Universities of Computer Studies, Yangon.*
cmaye84@gmail.com

## Abstract

*Opinion reviews on products are used by potential customers before deciding to purchase a product. Large volumes of reviews are posted to the Internet and detection of fake reviews is now a challenging research problem. Opinion spam or trustworthiness of online opinions is an important issue and this issue has been neglected so far. To the best of our knowledge, there is not many published study on this topic, although Web spam and email spam have been investigated extensively. This paper is presented spammer detection techniques that define the spammer score on rating behaviors of the reviewer. The experiment showed that the presented technique has comparatively effective spammer detection than other techniques based on rating behaviors.*

Key words: Spam detection, rating behaviors, scoring method

## 1. Introduction

It is now quite common for people to read reviews on the Web for many purposes. For example, if one wants to buy a product, one typically goes to a merchant site (e.g., amazon.com) to read some reviews of existing users on the product. If the reviews are mostly positive, one is very likely to buy the product. If the reviews are mostly negative, one will most likely buy a different product. Positive opinions can result in significant financial gains and/or fames for organizations and individuals. This gives good incentives for review or opinion spam [12]. In Web 2.0 application, the user contributed comments offer the promise of a rich source of contextual information about Social Web content. Due to the fact that the quality is not control, anyone can write anything on the Web. This results in many low quality reviews, and worse still *review spam*.

Typically, the reviews consist of an overall product score (often in the form of a star-rating) and some free-form review text to allow the reviewer to describe their experience with the product or service in question. Web user can post products reviews at merchant sites to express their views and interact with other users via blogs and forums. It is now well recognized that the user generated content contains valuable information that can be exploited for many applications [3, 13].

The existing work has been mainly focused on extracting and summarizing opinions from reviews using natural language processing and data mining techniques [1, 4, 11, 13 and 16]. Review spam is similar to Web page spam. In the context of Web search, due to the economic and/or publicity value of the rank position of a page returned by a search engine, Web page spam is widespread. [5, 6, 7, 17, 19, 20] Web page spam refers to the use of "illegitimate means" to boost the rank positions of some target pages in search engines [2, 20]. In the context of reviews, the problem is similar, but also quite different.

Due to the openness of product review sites, spammers can pose as different contributing spammed reviews making them harder to eradicate completely. Spam reviews usually look perfectly normal until one compares them with other reviews of the same products to identify review comments not consistent with the latter. The efforts of additional comparisons by the users make the detection task tedious and non-trivial [10]. One approach taken by review site such as Amazon.com is to allow users to label or

vote the reviews as helpful or not. Unfortunately, this still demands user efforts and is subject to abuse by spammers. Spam review is sometime with high helpfulness score so that helpfulness score is not a good indicator of spam.

Most review spam detection system focus on review similarity and also focus on review behaviors. They present a technique to detect review spam on the basis of spammer behavior that target products to promote or unjust them. Language modeling technique and some similarity computation techniques on review text are also proposed to detect spam and this can have time consuming because of deeply analysis on opinion and text understanding. One should focus on detecting spammers based on their spamming behaviors instead of detecting spam reviews only. In fact, the more spamming behaviors the system can detect for a reviewer, the more likely the reviewer is a spammer. Subsequently, spam review can be removed to accelerate the interests of other review users.

The rest of the paper is organized as follows. Section 2 covers some related works. Section 3 presents spammer detection techniques with rating based spam score function. Experimental evaluation is described in Section 4 and Section 5 is devoted to conclusions.

## 2. Related Work

Analysis on online opinion becomes a popular research topic recently. Most research trends focus on opinion mining and opinion extraction. A preliminary study of opinion spam was reported in [13].

A spam activities analyzing and spam detection methods are presented in [15]. Three types of spam review such as untruthful opinion, review on brand only and non-review (e.g question and answer and random texts) are also discussed. From this analysis, spammer detection can use some features to detect behaviors of the spammer. For example, the analysis said that reviews with negative deviation (in rating) on the same brands give the highest lift curve (that is spam), only reviews are very likely to be candidates of spam. So that, the presented techniques use this analysis and then builds the detection technique that is more effective than the methods proposed in [10].

The scoring methods to measure the degree of spam for each reviewer is presented in [10] and applies on an Amazon review dataset. Scoring method for spammer on targeting product and product groups is proposed and they focus on scoring methods of target based spamming and deviation based spamming score methods. Then highly suspicious reviewers is selected for further scrutiny by user evaluators with the help of a web based spammer evaluation software specially developed for user evaluation experiments. The results show that the ranking and supervised methods are effective in discovering spammers and outperform the baseline method based on helpfulness votes alone. To assign an overall numeric spam score to each user, the spam scores of the user's different spamming behaviors are combined by using linear weighted combination. The weights on the different component spam scores can be empirically defined or learnt automatically. Only rating deviation and early rating deviation was not much effective. So that the presented techniques add more features on rating to detect spammer.

A language modeling approach for consumer review spam detection is presented in [9]. They showed that Kullback-Leibler (KL) Divergence and the probabilistic language modeling based computational model. This model is effective for the detection of untruthful reviews to estimate the similarity between any pairs of reviews in terms of the likelihood of a review “generating” the contents of another review. Moreover, the Support Vector Machine also called SVM -based method is also effective for the detection of non-reviews. The empirical study found that around 2% of the consumer reviews posted to a large e-Commerce Web site is spam. The computational model proposed spam detection based on review similarity and other behaviors of rating has not taken into account.

Review and rating similarity is also importance to detect spammer because most spammers are lazy and they copy and paste their

review instead of writing new one. This paper presented a technique to detect spammer using rating behavior. More deeply investigation is making in spamming behaviors for more efficient spamming score. Spamming behaviors are complicated and not easily captured. So many researches about review spam detection are required for improving web sites.

## 3. Rating Behaviors based Spammer Detection

In this work, spammer detection is targeted on the rating of the reviewer given to the products. The spamming behaviors of the reviewer are calculated by the scoring techniques. These techniques are considered on the detailed features of ratings. Table 1 describes about notations that are used in this work.

**Table 1. Notations and descriptions**

| Notation | Description |
|---|---|
| U | $\{u_i\}$: set of users |
| O | $\{o_j\}$: set of products |
| E | {e}: set of ratings e (e ∈ [1,5]). |
| o(e) | product of rating e |
| $\mathbf{E_{i*}}$ | $\{e \setminus u(e)= u_i \wedge o(e)=o_*\}$ : set of ratings on any products by user $u_i$ |
| $\mathbf{E_{*j}}$ | $\{e \setminus u(e)= u_* \wedge o(e)=o_j\}$ : set of ratings on product $o_j$ |
| p(e) | position of rating e |

Rating-based spam score is the spam score of user to the products by their rating behaviors. Reviewer gives same rating and review to the products to save their time. They just copy and paste their review to pack away the spamming effort. So that, rating similarity of reviewer on the products are one important behavior to consider spamming behaviors. Spammer promote or victimize a few products or product line and ratings they give to the product are quite difference from others. Spammer tends to give high rating in low quality product to promote that product and also give low rating in high quality product to damage that product reputation. The spammer's rating can be deviated from other product rating. Other rating related features such as high or low rating, high rating after first low rating, low rating after first high rating and negative deviation are considered to detect spammer. The spam score techniques based on these features are presented in the next subsections.

### 3.1 Rating Similarity Spam Score

As described above, rating similarity is an important factor to determine spammer. Spammer gives same rating to the product to conserve their spamming effort. Rating similarity scores of all ratings given by one reviewer is calculated to detect the spammer. The similarity function $\mathrm{sim}(e)$ is achieved by comparing ratings in a given set $E_{i*}$ and is defined as:

$$sim(e) = 1 - \underset{e \in E_{i*} \wedge p(e) < p(e')}{avg} (e - e') \quad (1)$$

Rating similarity score of user $u_i$ can be obtained by:

$$SR_S(u_i) = Avg_{e \in E_{i*}}\, sim(e) \quad (2)$$

Since spammer give almost same or not much different rating on the products, high similarity score of rating is more likely to be spammer.

### 3.2 Rating Deviation Spam Score

Reasonable user is expected to give ratings that are similar to other rating of user on the same product. Mostly, the ratings of the spammer are different from these reasonable ratings. Reasonable user give high rating (good rating) to the good product but spammer give low rating (bad rating) to damage that product reputation. Sometimes, they give same rating as other to hide from spammer detection system but after they get believe from the detection system, they start their spamming effort and give deceptive rating. As spammer attempt to promote or demote products, their rating could be quite different from others. General deviation is therefore a possible rating behavior demonstrated by a spammer. The deviation of a rating *'e'* is defined as its difference from the

average rating on the same product:

$$dev(e) = e - Avg_{e \in E_{k*} \wedge e' \in E_{*k}} e' \quad (3)$$

The spam score of a user $u_i$ based on rating deviation is thus defined as the average deviation of his or her ratings:

$$SR_D(u_i) = Avg_{e_{ij} \in E_{i*}} |dev(e)| \quad (4)$$

### 3.3 Other Rating Behaviors Spam Score

Only rating deviation is not a good indicator to obtain effective detection, so the system is considered on other ratings behaviors that likely to be spammer. Some reviewer gives high rating and low rating to promote or demote product. Suspicious rating behaviors involve very good ratings and bad rating on products. Spammer writes good rating just after bad rating and also writes bad rating just after good rating to damage control. N. Jindal and B. Liu analyzed review spam [15] and they presented that reviews with negative deviation on the same product give highest spam score. So that, the presented technique emphasis on these features of rating and calculate their spamming score by:

$$orb(e) = \frac{1}{4}\,[r(e) + rgb(e) + rbg(e) + ndev(e))] \quad (5)$$

where $orb(e)$ is other rating behaviors spam score.

$$r(e) = \begin{cases} 1, if\ e\ is\ good\ rating\ or\ bad\ rating \\ 0, else \end{cases} \quad (6)$$

$$rgb(e) = \begin{cases} 1, if\ rating\ e\ is\ bad\ rating\ and\ come \\ \quad after\ the\ first\ bad\ rating \\ 0, else \end{cases} \quad (7)$$

$$rbg(e) = \begin{cases} 1, if\ rating\ e\ is\ good\ rating\ and\ come \\ \quad after\ the\ first\ good\ rating \\ 0, else \end{cases} \quad (8)$$

$$ndev(e) = \begin{cases} 1, if\ dev(e) is\ negative \\ 0, else \end{cases} \quad (9)$$

All the spam score of ratings given by each reviewer can be obtained by:

$$SR_B(u_i) = Avg_{e \in e_{i*}} orb(e) \quad (10)$$

### 3.4 Combined Rating Spam Score

Combined spam score is the spam score of user to the products by their rating spamming behaviors. It needs to combine all the spam scores on rating to achieve effective spam score by simply giving some weight of each score function. In this case, rating similarity score get more weight because the similar rating is more likely to be spammer than others. The required rating based spam score of user $u_i$ is defined by:

$$SR(u_i) = \frac{1}{2}\,SR_S(u_i) + \frac{1}{4}SR_D(u_i) + \frac{1}{4}\,SR_B(u_i) \quad (11)$$

## 4. Evaluation

The presented methods are evaluated by using data from amazon.com. The reason for using this data set (http://131.193.40.52/data) is that it is large and covers a very wide range of products. Amazon.com is considered one of the most successful e-commerce Web sites with a relatively long history. This dataset gives the information such as Product ID, Reviewer ID, Rating, Date, Review Title, Review Body, Number of Helpful Feedbacks and Number of feedbacks. The statistics of the dataset are presented in the following table.

**Table 2. Dataset statistics**

| | Number |
|---|---|
| **U** | 11,038 |
| **O** | 5,693 |
| $\|\mathbf{E_{i*}}\|$ | 48,894 |

Spam score techniques of (i) rating similarity (ii) rating deviation (iii) other rating behaviors (iv) rating similarity and deviation by taking average (v) combined rating spam scores are evaluated in this section. Since there was no standard dataset to test the accuracy, three human evaluators are used to check whether a reviewer is spammer or not by using spammer evaluation software that are developed and presented in figure 1. The evaluator can look at the feature of each reviewer's rating behaviors and can give feedback whether the reviewer is spammer or not.

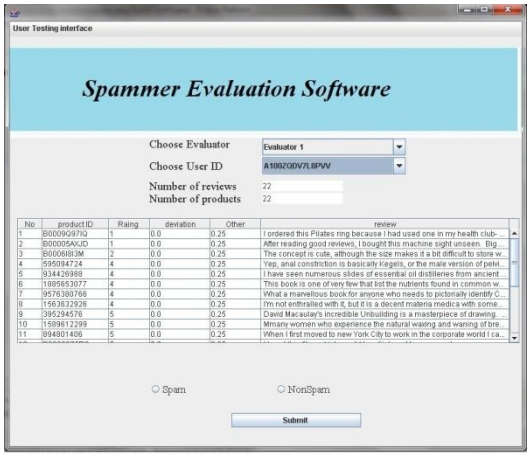


**Figure 1. User Evaluation Software**

There are several challenges in conducting the user evaluation experiments. There are many reviewers and it is impossible for the human evaluators to judge everyone. Only small subsets of reviewers are evaluated to handle this issue. These comprise both reviewers highly suspicious of spamming by each technique as well as those un-suspicious ones.

## 4.1 Evaluation Setup

For each spammer detection method, select 10 top ranked reviewers and 10 bottom ranked reviewers. And then merge all the selected spammers into a pool which consists of 62 reviewers. These reviewers are then sorted by their combined spamming behavior scores. 25 top ranked reviewers and 25 bottom ranked reviewers are then selected for user evaluation. This number of reviewers is quite reasonable for a human evaluator to examine. The system further randomly order the reviewers so that there is no relationship between reviewers' order and their spammer scores

For each reviewer, select his/her ratings to be highlighted for human evaluator's attention. These selected ratings must be seen by the human evaluator before the latter makes judgment on whether the reviewer is a spammer. The ratings are selected based on their involvement in the spamming behaviors identified. Specifically, identical (or similar) ratings with other ratings, deviated rating from the average ratings of the reviewed products, other rating behaviors such as good or bad rating, good rating after bad rating, bad rating after good rating and negative deviation of the rating are selected.

Three human evaluators are recruited to examine the selected reviewers and ratings using the spammer evaluation software. For each reviewer, the labeled decision is either spammer" and non-spammer" and the evaluators are not informed about the number of spammers to be labeled.

## 4.2 Results

Given the results of the three evaluators, final label to each reviewer is assigned by using majority voting. A reviewer is assigned a final spam or non-spam label if the label is agreed by two or more evaluators. The number of top 10 reviewers with the final spammer labels, and the number of bottom 10 reviewers with non-spammer labels for different methods are shown in Table 3. Rating deviation spam score $SR_D(u_i)$ is a base line method and the presented techniques are significantly better than base line method.

Text Retrieval Conferences(TREC) spam track (Tormack and Lynam. 2005; Cormack 2007) is used to measure the performance of the presented techniques. They were widely used to evaluate other kinds Web spam. With reference to a confusion matrix depicted in Table 4, the various effectiveness measures can be defined by:

**Table 3. Results of Top 10 and Bottom 10 Ranked Reviewers**

| | Spam score methods | | | |
|---|---|---|---|---|
| | **Similarity** | **Deviation (Base line)** | **Similarity and Deviation** | **Combined Similarity, Deviation and Other behaviors** |
| #Spammer in top 10 | 7 | 6 | 8 | 9 |
| #nonspammer in bottom 10 | 10 | 6 | 7 | 8 |

$$hm = \frac{b}{b+d} \quad (12)$$

$$sm = \frac{c}{a+c} \quad (13)$$

$$lam = logit^{-1}\left(\frac{logit(hm)+logit(sm)}{2}\right) \quad (14)$$

$$tp = \frac{a}{a+c} \quad (15)$$

**Table 4. Confusion matrix for the definition of the Effectiveness Measure**

| | | Gold Standard – Human Classification | |
|---|---|---|---|
| | | Spam | Ham |
| System's Classification | Spam | a | b |
| | Ham | c | d |

where a, b, c, and d refer to the number of reviews falling into each category. The ham misclassification rate (hm) is the fraction of all ham misclassified as spam; the spam misclassification rate (sm) is the fraction of all spam misclassified as ham. It is desirable to have a single measure which combines both of the above measures. Therefore, the TREC Spam track also made use of the logistic average misclassification rate (lam) to measure the effectiveness of spam detection systems, where $logit^{-1}(x) = \frac{e^x}{1+e^x}$ and $logit(x) = \ln\left(\frac{x}{1-x}\right)$. Since hm, sm, and lam are the measures for failure rather than effectiveness, the lower scores imply a better detection performance. The true positive rate (tp) is the fraction all spam identified by the system. On the other hand, the common effectiveness measure accuracy=$\frac{a+d}{a+b+c+d}$ may be measured for spam detection System. The accuracy figure is report in table [5]. These values are calculated from results of the user evaluation test on 50 reviewers. The presented techniques have more accuracy than base line method.

**Table 5. Comparative performance of the spammer detection techniques**

| | Spam score methods | | | |
|---|---|---|---|---|
| | **Similarity** | **Deviation (Base line)** | **Similarity and Deviation** | **Combined Similarity, Deviation and Other behaviors** |
| tp% | 0.70 | 0.60 | 0.80 | 0.90 |
| hm% | 0.00 | 0.40 | 0.30 | 0.20 |
| sm% | 0.30 | 0.40 | 0.20 | 0.10 |
| lam% | 0.39 | 0.40 | 0.09 | 0.14 |
| Accuracy | 0.85 | 0.60 | 0.75 | 0.85 |

## 5. Conclusion and Future Work

This paper presents review spammer detection techniques based on rating behaviors such as rating similarity, rating deviation, other rating behaviors and combination of these. The presented methods are better than baseline method but it is still needed to improve than other methods. Moreover review behaviors are also important for spammer detection. So that, review behaviors based scoring methods such as review similarity, proportion of review words length from other reviews, proportion of positive and negative words length are needed to be embedded to achieve more effective spammer detection system.

## References


[1] A-M. Popescu and O. Etzioni. Extracting Product Features and Opinions from Reviews. *EMNLP'2005*.

[2] A. Ntoulas, M. Najork, M. Manasse & D. Fetterly. Detecting Spam Web Pages through Content Analysis. WWW'2006.

[3] B. Liu. Web Data Mining: Exploring hyperlinks, contents and usage data. Springer, 2007.

[4] B. Pang, L. Lee & S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. *EMNLP'2002*.

[5] B. Wu and B. D. Davison. Identifying link farm spam pages. *WWW'06*, 2006.

[6] B. Wu, V. Goel & B. D. Davison. Topical TrustRank: using topicality to combat Web spam. *WWW'2006*.

[7] C. Castillo, D. Donato, L. Becchetti, P. Boldi, S. Leonardi, M. Santini, S. Vigna. A reference collection for web spam, *SIGIR Forum'06,* 2006.

[8] C. Danescu-Niculescu-Mizil, G. Kossinets, J. Kleinberg, and L. Lee. How opinions are received by online communities: a case study on amazon.com helpfulness votes. In WWW, 2009.

[9] C.L. Lai, K.Q. Xu, Raymond Y.K. Lau ,Y. Li and L. Jing, A language modeling approach for consumer review spam detection, IEEE International Conference on E-Business Engineering , 2010.

[10] L . Ee-Peng, V. Nguyen, N. Jindal, B. Liu, H. W. Lauw. Detecting Product Review Spammers using Rating Behaviors, CIKM'10, Toronto, Ontario, Canada, October 26–30, 2010.

[11] K. Dave, S. Lawrence & D. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *WWW'2003*.

[12] B. Liu, Web Data Mining. Springer, 2007.

[13] M. Hu & B. Liu. Mining and summarizing customer reviews. KDD'2004.

[14] N. Jindal and B. Liu. Review spam detection. In WWW (poster), 2007.

[15] N. Jindal and B. Liu. Opinion Spam and Analysis. WSDM'08, Palo Alto, California, USA, February 11-12, 2008.

[16] P.Turney. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *ACL'2002*.

[17] R. Baeza-Yates, C. Castillo & V. Lopez. PageRank increase under different collusion topologies. *AIRWeb'05*,

[18] K. Soo-Min, P. Pantel, T. Chklovski, M. Pennacchiotti. Automatically Assessing Review Helpfulness. Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pages 423–430, Sydney, July 2006.

[19] Y. Wang, M. Ma, Y. Niu, H. Chen. Spam Double-Funnel: Connecting Web Spammers with Advertisers. *WWW2007.*

[20] Z. Gyongyi & H. Garcia-Molina. *Web Spam Taxonomy*. Technical Report, Stanford University, 2004.