International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:1, 2016

A Framework for Review Spam Detection Research

Mohammadali Tavakoli, Atefeh Heydari, Zuriati Ismail, Naomie Salim

***Abstract*—**With the increasing number of people reviewing products online in recent years, opinion sharing websites has become the most important source of customers' opinions. Unfortunately, spammers generate and post fake reviews in order to promote or demote brands and mislead potential customers. These are notably destructive not only for potential customers, but also for business holders and manufacturers. However, research in this area is not adequate, and many critical problems related to spam detection have not been solved to date. To provide green researchers in the domain with a great aid, in this paper, we have attempted to create a high-quality framework to make a clear vision on review spam-detection methods. In addition, this report contains a comprehensive collection of detection metrics used in proposed spam-detection approaches. These metrics are extremely applicable for developing novel detection methods.

***Keywords*—**Fake reviews, Feature collection, Opinion spam, Spam detection.

I. Introduction

THE importance of online reviews for products and services is significant for today's businesses. Opinion-sharing websites enable customers to post their opinions regarding purchased and utilized products. These posted reviews provide useful information for potential customers. In fact, it is quite helpful for a potential customer to read reviews of a product before making a purchase decision. Furthermore, business holders and manufacturers use product reviews not only to understand their customers' needs but also to determine the weaknesses of their products such that they can customize and reshape the product to increase customer satisfaction and consequently increase sales. Such reviews also help to understand their competitors' situations in the market and to plan for success. Therefore, such reviews are precious sources of information for individuals and organizations. Hence, the trustworthiness of the reviews is essential for the opinions to be validly used for aforementioned purposes.

In recent years, opinion-sharing websites are turning into a competitive arena for businesses. Unfortunately, there is an enormous drawback with most of the opinion sharing websites. The sites enable anyone from anywhere in the world to post reviews on products without any limits. This ease of posting allows manufacturers and organizations to hire spammers to post spurious positive reviews to promote or support their merchandise and sometimes unfair negative reviews to damage competitors' reputations and denigrate their reliability. Most of these harmful fake reviews are not detectable by readers, due to their manipulated structure and their placement in the midst of truthful reviews. Therefore, opinion mining techniques are being employed to distinguish fake reviews from real opinions. One of the most significant issues in opinion mining is opinion-spam detection, which requires the researchers' attention more so than other issues because the trustworthiness of a review makes it valuable for various purposes. The hired reviewers who write unfavorable reviews of a product to distract readers are called review spammers. The opinions or reviews written by spammers are called spam reviews.

In this research, we have proposed an effective framework to be used for spam detection research. First, we have discussed the categorization and fundamental explanations of some of the elements and factors involved with spam detection. Various types of spam, spammers, spamming and features are considered in this step to provide a clear vision of them. Finally, we have purposed a collection of all of the possible features that can be used for detecting spam reviews, as well as individual and groups of spammers. This comprehensive collection of features could be an asset for researchers working at the same domain and could be used as a worthy resource for spam detection techniques. In particular, these would be extremely applicable for supervised methods that need many effective features to achieve the most accurate result. The framework in the existing study is discussed in three sections. In section II, we categorized and explained the various types of spam reviews, spam reviewers, and spamming. In Section III, we assessed the different types of features or data that are used for spam detection. Finally, in Section IV, we proposed a useful and advantageous collection of classified tables that include all of the features that have been extracted from the previous spam detection studies.

II. Spam Detection Definitions

In this section, opinion-spam detection and some factors that are necessary for distinguishing spam from non-spam reviews are discussed and described in detail. Understanding the concepts of a truthful review, a spam review, the types of spam reviews and spamming spam reviewers and finally individual and group spam detection will facilitate the process of developing a spam-review detection technique.

A. Opinion Spam Detection

Spam detection has been used in many fields. Email spam and web spam are two of the most popular types of spam that are commonly assessed (more than other types of spam) by

Mohammadali Tavakoli, Atefeh Heydari, Naomie Salim are with the Universiti Teknologi Malaysia, Johor Bahru, 81300 Malaysia (corresponding author phone: 09212378119; e-mail: tmohammadali2@utm.my, hatefeh@live.utm.my, naomie@utm.my).

Zuriati Ismail is with the Universiti Teknologi MARA Johor, KM12 Jalan Muar, 85009 Segamat, Johor, Malaysia (e-mail: zuria986@johor.uitm.edu.my) and a PhD candidate in Computer Science, Universiti Teknologi Malaysia, Johor Bahru.

researchers. However, opinion spam (spam review) is different from those two types. Email spam refers to unwanted advertisements that are sent by email; this spam is infrequent in opinion postings. In web spam, we have two major types: content spam and link spam [1], [2]. Spam with hyperlinks is called link spam, which is rarely observed in reviews. Although links to advertisements are frequent in social media, they are very easy to detect. Content spam adds some irrelevant words into some web pages to trick search engines into considering them to be pertinent to some search queries. This type of spamming is also inapplicable to opinion reviews.

One of the most essential issues in opinion mining is opinion-spam detection. This requires more attention than other types of spam because the trustworthiness of reviews makes them profitable for different proposes. The difference between opinion spam and other forms of spam makes opinion-spam detection more challenging. An ordinary reader is able to detect almost all other types of spamming activities easily. However, it is very hard, if not impracticable, to identify fake reviews by manually reading the reviews. This intensifies the need to screen spam reviews that are to be used for planning or appraising detection algorithms compared with any other types of spam. The ultimate goal of opinion-spam detection in the reviews is to capture every spam review, spam reviewer, and spam reviewer group. [3].

B. Truthful Review

A truthful review is a review written in an opinion sharing website by a customer who has really purchased the product and is honestly writing about his/her experience. This review may consist of the customer's opinion about the quality, after sale service, shipping duration, and other features of the product.

C. Spam Review

A review that is not a real and trustworthy reflection of the experience of a reviewer (and written about the product) is known as a spam review, a fake review, a bogus review or a fraudulent review. Sometimes, these reviews have positive or negative contents about the product, and rest of the time they have no negative or positive opinions, e.g., they simply provide advertisements for other products.

Many types of spam reviews are deleterious for business owners, organizations, and political cases. Unfortunately, spamming is going to become more complicated and grow increasingly out of control. Spam should be detected to protect innocent opinions and to remove fake reviews, defamation, and deception from social media.

Hired spammers have various purposes in creating and posting spam reviews which depend on the requirements of the manufacturers. There are four types of spam reviews defined in [4]:

1) Non-Opinion

These notes are devoid of any opinion about the target product. In fact these are totally irrelevant to the product. Non-opinion reviews consist of four core subtypes as discussed below [5]-[9]:

- Advertisement: There are three main types of advertisements: (1) Advertising the target product; the review explains the usage or features of the product. There is no presentation of any customer's opinion regarding the product. (2) Advertising other products; a review from this category is advertising other products. The structure of this review is similar to the aforementioned case, but promotes other or sometimes competing products. (3) Advertising different sellers; this type of review is advertising for other sellers, stores or websites for the purchase of the same product e.g., "This camera is very nice and can be purchased with a promotion from www…..com."
- Question or answer: This type of review is written by a reviewer to ask (or answer) questions about the product.
- Comments: A reviewer posts this type of reviews as a comment on some other reviews written by others on a similar product.
- Random texts: The review in this case contains some irrelevant notes about different things unrelated to the product.

2) Reviews of a Brand

These reviews also do not present the customer's opinion of the product. The reviews are not targeted at a specific product, but only focus on the brand, sellers and manufacturers of that product or on some merchandise of a specific organization. For instance, a review for a particular Dell laptop says "I dislike Dell. I will not purchase any of their merchandises or products".

3) Fake Reviews

Fake reviews are written by spammers (positively) to promote a product or (negatively) to damage the reputation of a product. This type of spam review is the most important one because people can easily recognize the previously mentioned types by reading them, but reviews of this type cannot be recognized as being false by an ordinary reader. Consequently, most of the studies have focused on detecting this type of spam review.

4) Harmful Reviews

The products can be placed in three categories by using the quality of the products as an indicator: (a) good quality products, (b) bad quality products, and (c) average quality product. Based on this, six types of fake reviews are possible. These are listed in Table I.

TABLE I
PRODUCT QUALITY VS. SPAM REVIEWS [10]

Product quality	Positive spam	Negative Spam
Good	1	2
Bad	3	4
Average	5	6

Stores and manufacturers usually create reviews in regions 1, 3, and 5 for their products to promote them. Conversely,

reviews in regions 2, 4, 6 are written by competitors of the product to damage its reputation. Obviously, positive reviews of good quality products are not detrimental, and neither are negative reviews of bad quality products. However, reviews from regions 2, 3, 5, and 6 can be very harmful for the products [10]. Thus, most of the techniques for review spam detection should focus on detecting these harmful reviews.

D.Spam Reviewer

The person who posts spam reviews is called a spam reviewer [11]. Spam reviews may be written by different types of people, e.g., employees of a company, friends, competitors, business holders and even real customers. Some businesses provide a repayment or discount to some of their consumers to write positive reviews for their products, and sometimes to write negative reviews of their competitors' products.

1) Individual Spammer

An individual spam reviewer is a spammer who works individually. Actually, he/she posts spam reviews using one user-id. Generally, a spam reviewer may work alone, or intentionally or unintentionally work as a member of a group.

2) Group of Spammers

Mukherjee in [12], [13] defined two different types of group spammers: (1) Sometimes a group of spammers work together to support or harm some products. In this mode, each of the individual spammers who work in the group may or may not know the others. (2) One person (spammer) has several user-ids. Therefore, he/ she posts spam reviews using these different user-ids, with results that look similar to that of a group. The authors of [13] argued that group spamming is much more destructive than individual spamming. The group can use various spamming methods which make the spamming incredibly difficult to detect.

III. Detection Metrics and Data Assessment

In this section, three types of data and features for identifying spam and non-spam are listed and described. Assessing them is essential for most of the spam detection techniques, specifically for the supervised methods. The features of each of these types (except the content of the review) can be classified as publicly available data and website private data [14]. The data that can be obtained from the opinion sharing website and that is open to being accessed by visitors is called publicly available data. Such data include the date of posting of a review, the content, or text of the review, the rating given by a reviewer to the target product and the price of the product. The data that are not accessible by visitors and that are under the possession of the associated website are known as website private data. Such data include the MAC address, the geographical location, and the IP address of the user.

A. Content of Review

The content of a review is the text of the review excluding features, such as the reviewer's name or id, date of posting, and rating. The contents of the reviews were considered as the foremost (and in some cases the only) aspects used in the spam detection technique. Moreover, linguistic features such as words and POS n-grams that are operational in identifying malicious behaviours (e.g., deceptions and lies) can be extracted from the contents of the reviews. However, a spammer can post a fake review so subtly that no one can identify this fake review by reading its content. Furthermore, for instance, when a person has a good experience with a product, he/she can easily write of his/her excellent experience, but do so for a poor product as a spammer (he/she posts a positive fake review that is based on his/ her real experience with another product) [1]. Therefore, linguistic features alone are not sufficient for developing an appropriate spam-detection method. These must be considered as part of a set of features needed for a proper spam detection method.

B. Meta-Data of Review

Some of the information about the review other than the review content is called meta-data. Examples include: the reviewer's id, the time of publishing of the review the star rating given to the target product by a reviewer the time duration of the writing of the review, the geo-location of the reviewer's computer, and the reviewer's computer MAC and IP addresses.

Multiple unusual behaviors of reviewers that are not detectable by using only the contents of the reviews can be extracted using the meta-data of the reviews. Some examples are: (1) Sometimes several user-ids from the same computer write multiple positive or negative reviews on a particular product, and for this reason these reviews are doubtful. This example demonstrates the role of the user id and user's computer MAC address for review-spam detection. However, the MAC address of the customers' computer is not publicly available to be used in the detection methods proposed by researchers. (2) When we assess the ratings of reviews for a product, we can see that a particular reviewer has posted illusory positive reviews for a particular brand and that he/she has posted multiple destructive negative reviews for competing brands. This example expresses the importance of the ratings given by a reviewer to a product, in the spam-detection method. Such rating anomaly behaviors are extremely helpful in spam-detection techniques. (3) Occasionally, when we are looking for spam in some hotel reviews, we may find many positive reviews about a hotel from places near the location of the hotel. Therefore, these are not most likely the opinions of real customers because the reviewers of a hotel would typically live in a geo-location that is far from the hotel. This example shows the usefulness of the geo-location of a reviewer for a spam-detection method. The geo-location is one of the website's private features.

C. Information Regarding the Target Product

A portion information about the product (or the entity under review) has been used in some proposed spam-detection techniques. Such information includes: the description of the product, price, date of launching of the product and volume of sales. For example, when a product has many positive reviews

but it has not sold very well, the reliability of the positive reviews is difficult to accept. Furthermore, considering the date of the launching of a product, when the number of reviews for the product is very high, the trustworthiness of the reviews is questionable.

With the abovementioned three types of data, one can explore many features that can be used in spam detection techniques.

IV. Collection of Detection Metrics

To the best of our knowledge, there has been no gathering of the features employed in the various spam detection techniques before this. This task was very time consuming and required the careful scrutinizing of proposed review-spam detection techniques. Fortunately, we have successfully accomplished this task. Finally, with this information, we created two classes. The first class contains all of the features used in distinguishing spam and non-spam reviews. The second one includes all of the features used in identifying group and individual spammers. Thus, considering the following classified features facilitates the development of any review-spam detection technique (especially supervised methods).

A. Review Spam Detection Metrics

The following sections contain a large number of features used to distinguish spam from non-spam reviews that are extracted from related studies. There are fifty one rows of features in the different tables. We categorized them into three main groups: review features, reviewer features and product features. Then, we formed twelve subsets for the proliferation of features that branched off from each category. Thus, the review features category includes countable features, positional features (meta-data), textual features, sentiment features, and rating features. The reviewer features category includes countable features, rating features, behavioral features and positional features (meta-data). Finally, the product features category includes centric features and ranking features.

1) Review-Based Metrics

A feature that contains any information about the review can be included in this category. These features are generally used for review-spam detection techniques that only focus on analyzing reviews and ignored the reviewers' behaviors completely. Table II outlines the review features.

2) Author-Based Metrics

A feature that has any information about the person who has written the review can be included in this category. Reviewer features are generally used in spammer detection techniques. However, different combinations of the review features are used in different cases. Table III illustrates the reviewer features.

TABLE II
Review-Based Detection Metrics

	Review Features
Countable Features	1. Number of feedbacks [10]
	2. Number of helpful feedbacks [10]
	3. percent of helpful feedbacks [10]
	4. Number of reviews [15]
	5. Length of the review title [10]
	6. length of the review body [11], [16]
	7. Ratio of single reviews [15]
Positional Features (meta-data)	1. Position of the review in the reviews of a product sorted by date ascending [10]
	2. Position of the review in the reviews of a product sorted by date descending [11]
	3. Binary feature to indicate if a review is the first review [10]
	4. Binary feature to indicate if a review is only review [4]
	5. Date of publishing review [1]
	6. Time duration of writing review [1]
	7. Honesty of review [19]
Textual Features	1. Percent of positive [10]
	2. Percent of negative [10]
	3. Cosine similarity of the review and product features [10]
	4. Percent of times brand name [10]
	5. Percent of numerals [10]
	6. Percent of capital [11]
	7. Percent of all capital words in the review [11]
	8.Unigram, Bigram and Trigram [17], [18]
	9.Square of normalized length [17]
	10. First Person vs. Second Person [17]
	11. High Similarity Score [17]
	12. POS distribution [18]
	13. Subjective vs. Objective [17]
	14. Positive vs. Negative [17]
Sentiment Features	1. Subjective vs. Objective [17]
	2. Positive vs. Negative [17]
Rating Features	1. Rating of the review [10]
	2. Deviation of the review from product rating [10]
	3. Feature indicating if the review is good, average or bad [10]
	4. Binary features indicating whether a bad review was written just after the first good review of the product [10]
	5. Binary features indicating whether a good review was written just after the first bad review of the product [10]

TABLE III
Authors-Based Detection Metrics

	Reviewer Features
Countable Features	1. Ratio of the number of reviews that the reviewer wrote which were the first reviews of the products to the total number of reviews that he/she wrote. [10]
	2. Ratio of the number of cases in which he/she was the only reviewer. [10]
	3. Number of reviewer of the product. [10]
Positional Features (meta-data)	1. The geo-location of the reviewer's computer [1]
	2. The reviewer's computer MAC and IP addresses [1]
	3. User-id of reviewer [1], [16]
Rating Features	1. Average rating given by reviewer [10]
	2. Standard deviation in rating [10]
Behavioural Features	1. Authority score [17]
	2. Brand deviation score [17]
	3. A feature indicating if the reviewer always gave only good, average or bad rating [11]
	4. Reviewer gave both good and bad ratings [11]
	5. Reviewer gave Good rating and average rating [11]
	6. Bad rating and average rating [11]
	7. Good rating, bad rating and average rating [11]
	8. Percent of times that a reviewer wrote a review with binary features [10]
	9. Percent of times that a reviewer wrote a review with binary features [10]
	10. Reviewer's trustiness [19]
	11. The star rating that given to each review by reviewers [1]

3) Product-Based Detection Metrics

Description of the target product and its features can be listed in this category. Product features can be used in different types of detection techniques. Examples of product features are listed in Table IV.

TABLE IV
PRODUCT-BASED DETECTION METRICS

	Product Features
Centric Features	4. Price of the product [10]
	5. Reliability of the product [10]
	6. Number of reviewer for the product [10]
	7. Reliability of the store [19]
Ranking Features	8. Sales rank of the product [10]
	9. Average rating [10]
	10. Rate of product rank [15]

TABLE V
SPAMMER(S) DETECTION METRICS

Group Spam Detection Features
1. Time Window (TW) [13]
2. Group Deviation (GD) [12]
3. Group Content Similarity (GCS) [13]
4. Member Content Similarity (MCS) [13]
5. Early Time Frame (ETF) [12]
6. Ratio of Group Size (RGS) [13]
7. Group Size (GS) [12]
8. Support count (SC) [13]
9. High Similarity Score [17]
10. Time of publishing review [1]
11. Standard deviation in rating [10]
12. Reliability of the product or store [19]
13. Number of reviewer for the product [10]
14. Sales rank of the product [10],[11]

B. Individual and Group of Spammers Detection Metrics

Apart from detection metrics used in capturing fake reviews, there are variety of methods focusing on detecting spammers who work individually or in groups. A comprehensive review of the proposed approaches in this domain is provided in our previous work [3]. Table V shows all of the existing features used to detecting individuals and groups of spammers.

C. Conclusion

As fake reviews are increasingly harming businesses and potential customers, it is critical to detect and eliminate such reviews to expurgate the opinion sharing websites. This paper proposed a high-quality framework for review-spam detection research. First, the paper discussed and explained the critical aspects and concepts in this field. Next, the paper provided a collection of features that is applicable for the development of review-spam detection methods.

ACKNOWLEDGMENT

This work is supported by the Ministry of Education Malaysia and Soft Computing Research Group (SCRG) of Universiti Teknologi Malaysia (UTM). This Work is also supported in part by grant from Vote 4F373.

REFERENCES

[1] Liu, Bing. *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media, 2007. Liu (2011). Opinion mining and sentiment analysis. Web Data Mining, Springer: 459-526.
[2] Castillo, Carlos, and Brian D. Davison. "Adversarial web search." *Foundations and trends in Information Retrieval* 4, no. 5 (2011): 377-486.
[3] Heydari, Atefeh, Mohammad Ali Tavakoli, Naomie Salim, and Zahra Heydari. "Detection of review spam: A survey." *Expert Systems with Applications* 42, no. 7 (2015): 3634-3642.
[4] Jindal and Liu (2007a). Analyzing and Detecting Review Spam. Seventh IEEE International Conference on Data Mining.
[5] Newman *et al.* (2003). "Lying Words: Predicting Deception from Linguistic Styles." Personality and Social Psychology Bulletin 29: 5.
[6] Hancock *et al.* (2007). "On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer- Mediated Communication." Discourse Processes 45: 23.
[7] Pennebaker *et al.* (2007). "The Development and Psychometric Properties of LIWC." www.LIWC.Net.
[8] Zhou *et al.* (2008). "A Statistical Language Modelling Approach to Online Deception Detection." IEEE Transactions on Knowledge and Data Engineering - TKDE, 20: 8.
[9] Mihalcea and Strapparava (2009). "The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language." Conference: Meeting of the Association for Computational Linguistics - ACL: 4.
[10] Jindal and Liu (2008). "Opinion Spam and Analysis." Conference of web search and web data mining: 11.
[11] Jindal and Liu (2007b). "Review Spam Detection." World Wide Web Conference Series: 1189-1190.
[12] Mukherjee *et al.* (2011). Detecting Group Review Spam. in Proceedings of International Conference on World Wide Web (WWW-2011, poster paper).
[13] Mukherjee *et al.* (2012). Spotting Fake Reviewer Groups in Consumer Reviews. in Proceedings of International World Web Conference (WWW-2012).
[14] Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis Lectures on Human Language Technologies* 5, no. 1 (2012): 1-167.
[15] Xie *et al.* (2012). Review Spam Detection via Temporal Pattern Discovery. international conference on Knowledge discovery and data mining
[16] Zuriati Ismail, Atefeh Heydari, Mohammadali Tavakoli, Naomie Salim. "Incorporating Author's Activeness in Online Discussion in Thread Retrieval Model" *ARPN Journal of Engineering and Applied Sciences* 10 (2), 473-479
[17] Li *et al.* (2010). Learning to Identify Review Spam. Joint conference on AI
[18] Ott *et al.* (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination. 49th annual meeting of the association for the computational linguistics.
[19] Wang *et al.* (2011). Review Graph based Online Store Review Spammer Detection. IEEE International Conference on Data Mining - ICDM 6.