# Review Spam Detection

Nitin Jindal and Bing Liu
Department of Computer Science
University of Illinois at Chicago
851 South Morgan Street
Chicago, IL 60607-7053
nitin.jindal@gmail.com, liub@cs.uic.edu

## ABSTRACT
It is now a common practice for e-commerce Web sites to enable their customers to write reviews of products that they have purchased. Such reviews provide valuable sources of information on these products. They are used by potential customers to find opinions of existing users before deciding to purchase a product. They are also used by product manufacturers to identify problems of their products and to find competitive intelligence information about their competitors. Unfortunately, this importance of reviews also gives good incentive for *spam*, which contains false positive or malicious negative opinions. In this paper, we make an attempt to study review spam and spam detection. To the best of our knowledge, there is still no reported study on this problem.

## Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering*.

**General Terms:** Experimentation, Algorithms.

**Keywords:** Product reviews, review spam, opinion spam

## 1. INTRODUCTION
The Web has dramatically changed the way that people express themselves and interact with others. They can now post reviews of products at merchant sites (e.g., amazon.com) and express their views in blogs and forums. It is now well recognized that such *user generated contents* on the Web provide valuable information that can be exploited for many applications. In this paper, we focus on customer reviews of products, which contain information of consumer opinions on the products, and are useful to both potential customers and product manufacturers [4, 8].

Recently, there was a growing interest in mining opinions from reviews. However, the existing work is mainly on extracting positive and negative opinions using natural language processing techniques [e.g., 2, 4, 5, 8, 10]. There is no reported study on the trustworthiness of reviews, which is crucial for all opinion based applications. Due to the fact that there is no quality control, anyone can write anything on the Web, which results in many low quality reviews, and worse still *spam reviews* [8].

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 of 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/opinion spam* [8].

There are generally two types of spam reviews. The first type consists of those that deliberately mislead readers or automated opinion mining systems by giving undeserving positive opinions to some target products in order to promote them and/or by giving unjust or malicious negative reviews to some other products in order to damage their reputation. The second type consists of non-reviews (e.g., ads) which contain no opinions on the product.

Review spam is related to but also different from Web or email spam. The objective of Web spam is to attract people to some target pages by manipulating the content of the pages and/or their link structures so that they will be ranked high by search engines. Spam emails are mainly ads. Spam reviews are very different as they give false opinions, which are much harder to detect even manually. Thus, most existing methods for detecting web spam and email spam [3, 7, 9, 11] are unsuitable for review spam.

In this work, we study review spam. Our investigation is based on 5.8 million reviews and 2.14 million reviewers (members who wrote at least one review) crawled from amazon.com. We discovered that spam activities are widespread. For example, we found a large number of duplicate and near-duplicate reviews written by the same reviewers on different products or by different reviewers (possibly different userids of the same persons) on the same products or different products.

Our objective of this work is to highlight review spam in order to shred some light on the trustworthiness of on-line reviews and to detect possible spam activities. We propose to perform spam detection based on duplicate finding and classification. For classification, we regard spam detection as a 2-class classification problem, *spam* and *non-spam*. Logistic regression is applied to learn a predictive model. Our experiment results demonstrated the effectiveness of the model.

## 2. REVIEW SPAM AND DETECTION
We perform spam detection using two methods:

1. Duplicates detection: There are a large number of duplicate reviews and many of them are clearly spam. For example, different userids posted duplicate or near duplicate reviews on the same product or different products. Duplicate detection is done using the shingle method [1] with similarity score > 0.9.
2. Spam classification: For the rest of spam reviews, we detect them based on 2-class classification (*spam* and *non-spam*).


*WWW 2007,* May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

We build a machine learning model to classify each review, i.e., to assign a probability likelihood of each review being a spam.

To build a classification model, we need labeled training examples of spam reviews and non-spam reviews. Recognizing whether a review is a spam review or not is extremely difficult by manually reading the reviews because one can carefully craft a spam review which is just like any other innocent review and the number of spam reviews is also small. We tried to read a large number of reviews and were unable to identify reliable spam reviews except finding a few obvious advertisements, which are irrelevant to the products being reviewed and contain no opinions. Thus, other ways have to be used to find training examples.

We propose to treat duplicate reviews as the positive training examples (spam), and the rest of the reviews as the negative training examples. We then use them to learn a model to discover non-duplicate reviews with similar characteristics, which are very likely to be spam reviews. Since the number of such duplicate spam reviews is large, it is reasonable to assume that they may be a fairly good sample of many types of spam reviews if not all.

To confirm that using only review contents is very hard to detect spam manually, we ran the text classification technique, naïve Bayes, to classify spam (duplicates) and non-spam reviews. The results were very poor with precision and recall for spam reviews being around 4%–5%. Thus, our spam detection technique has to depend on meta-features about (1) reviews and (2) reviewers.

**Feature construction:** Review centric features are characteristics of reviews. Reviewer centric features are characteristics of reviewers. We have altogether 24 features. We did not use any feature that would overfit the positive examples or are directly related to our manual spam labeling later. See [6] for details.

**Model building:** We used *logistic regression*. The reason for using logistic regression is that it produces a probability estimate that each review is a spam review, which is exactly what we need. It is almost certain that in the non-spam training or test data there are spam reviews which were not duplicated. This means that the labeled non-spam data has many errors.

## 3. RESULTS

We used the statistical package *R* (http://www.r-project.org/) to perform logistic regression. The AUC (Area under ROC Curve) is employed to evaluate the classification model. Our experiments are done using only the reviews for *manufactured products* due to data size. It does not make sense to combine it with other reviews for other category of products because they are too different.

Three types of duplicate reviews are used as the positive data (4488 cases), which are most likely to be spam: (1) duplicates from different userids on the same product, (2) duplicates from the same userid on different products; and (3) duplicates from different userids on different products. Negative data consists of the rest of the reviews (218524).

We performed 10-fold cross validation on the data. It gives us the average AUC value of 78%, which is quite high considering that many non-spam test reviews are actually spam and thus have similar probabilities as spam reviews in the spam set.

Recall our purpose of building the model is to detect spam that are not duplicates (because duplicates can be detected easily). We now show that this method works for non-duplicate reviews too. We want to see whether many highly ranked non-duplicate reviews are actually spam by manual inspection. To do this, we first ranked the negative test reviews (non-duplicates) based on their probabilities. We then manually checked the top ranked reviews to see if they were spam. The manual inspection was done by both authors based on consensus. We found that most of the top ranked reviews were very long. The average length of the top 100 reviews was 1800 words, much more than the average length of a normal review (123). After inspection, we found that 52% of top 100 reviews were long ads of related products. These reviews can clearly be regarded as spam reviews.

Row 2 of Table 1 shows the results. To access the performance on reviews with word length similar to the average length, we ran logistic regression again without using "length of review" as a feature. Few average length reviews were ads; so we relaxed the definition of spam reviews by including two more cases: 1) review praising the brand, but not the product; 2) reviews un-related to the product. Row 3 of Table 1 shows the results for negative test reviews (which are non-duplicates) belonging to this relaxed definition of spam.

The numbers are small because not many sophisticated spammers will make such fairly obvious mistakes which make it difficult to label them manually. But, it does show that the algorithm works.

**Table 1**. Spam reviews in top ranked negative reviews.

| Top ranked reviews | 1-15 | 15-30 | 30-45 | 45-60 | 60-75 | 75-90 |
|---|---|---|---|---|---|---|
| Spam Reviews | 14 | 13 | 10 | 3 | 7 | 5 |
| Spam Reviews without length feature | 10 | 9 | 4 | 7 | 5 | 3 |

We also performed spam detection from reviews with outlier ratings with promising results. However, due to space limitations, we are unable to give the results. See [6] for details.

## 4. CONCLUSION

This paper proposed to use duplicate detection and classification to detect review spam. Our preliminary experiments showed promising results. Our future work will focus on improving the accuracy and detecting more sophisticated spam reviews.

## 5. REFERENCES


[1]. Broder, A. Z. On the resemblance and containment of documents. In *Proceedings of Compression and Complexity of Sequences 1997*, IEEE Computer Society, 1997.

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

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

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

[5]. Jindal, N., & Liu, B. Identifying comparative sentences in text documents. *SIGIR'2006*.

[6]. Jindal, N. & Liu, B. Review Analysis. Tech. Report, 2007.

[7]. Li, K., & Zhong, Z. Fast statistical spam filter by approximate classifications. *SIGMETRICS 2006*, 2006.

[8]. Liu, B. *Web Data Mining*. Springer, 2007.

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

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

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