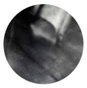
Opinion Spam and Analysis

Conference Paper · February 2008

DOI: 10.1145/1341531.1341560



Bing Liu
University of Illinois at Chicago

Opinion Spam and Analysis

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

Evaluative texts on the Web have become a valuable source of opinions on products, services, events, individuals, etc. Recently, many researchers have studied such opinion sources as product reviews, forum posts, and blogs. However, existing research has been focused on classification and summarization of opinions using natural language processing and data mining techniques. An important issue that has been neglected so far is opinion spam or trustworthiness of online opinions. In this paper, we study this issue in the context of product reviews, which are opinion rich and are widely used by consumers and product manufacturers. In the past two years, several startup companies also appeared which aggregate opinions from product reviews. It is thus high time to study spam in reviews. To the best of our knowledge, there is still no published study on this topic, although Web spam and email spam have been investigated extensively. We will see that opinion spam is quite different from Web spam and email spam, and thus requires different detection techniques. Based on the analysis of 5.8 million reviews and 2.14 million reviewers from amazon.com, we show that opinion spam in reviews is widespread. This paper analyzes such spam activities and presents some novel techniques to detect them.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering*. H.2.8: [**Database Management**]: Database Applications – *Data mining*

General Terms

Experimentation, Human Factors

Keywords

Opinion spam, review spam, fake reviews, review analysis

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 and express their views and interact with others via blogs and forums. Such content contributed by Web users is collectively called the *user-generated content* (as opposed to the content provided by Web site owners). It is now well recognized that the user generated content contains valuable information that can be exploited for many applications. In this paper, we focus on customer reviews of products. In particular, we investigate *opinion spam* in reviews. Reviews contain rich user opinions on products and services. 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 product problems and/or to find marketing intelligence information about their competitors [7].

In the past few years, there was a growing interest in mining opinions in reviews from both academia and industry. However, the existing work has been mainly focused on extracting and summarizing opinions from reviews using natural language processing and data mining techniques [7, 12, 19, 20, 22]. Little is known about the characteristics of reviews and behaviors of reviewers. There is also no reported study on the trustworthiness of opinions in reviews. Due to the fact that there is no quality control, anyone can write anything on the Web. This results in many low quality reviews, and worse still *review spam*.

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 [3, 5, 10, 12, 16, 24, 25]. Web page spam refers to the use of "illegitimate means" to boost the rank positions of some target pages in search engines [10, 18]. In the context of reviews, the problem is similar, but also quite different.

It is now very common for people to read opinions on the Web for many purposes. For example, if one wants to buy a product and sees that the reviews of the product are mostly positive, one is very likely to buy the product. If the reviews are mostly negative, one is very likely to choose another product. Positive opinions can result in significant financial gains and/or fames for organizations and individuals. This gives good incentives for *review/opinion spam*. There are generally three types of spam reviews:

Type 1 (untruthful opinions): Those that deliberately mislead readers or opinion mining systems by giving undeserving positive reviews to some target objects in order to promote the objects (which we call *hyper spam*) and/or by giving unjust or malicious negative reviews to some other objects in order to damage their reputation (which we call *defaming spam*).

Untruthful reviews are also commonly known as *fake reviews* or *bogus reviews*. They have become an intense discussion topic in blogs and forums. A recent study by Burson-Marsteller (http://www.burson-marsteller.com/Newsroom/Lists/BMNews/DispForm.aspx?ID=3645) found that an increasing number of customers are wary of fake or biased reviews at product review

WSDM'08, February 11-12, 2008, Palo Alto, California, USA.

sites and forums. Articles on such reviews also appeared in leading news media such as CNN (http://money.cnn.com/2006/05/10/news/companies/bogus_reviews/) and New York Times (http://travel.nytimes.com/2006/02/07/business/07guides.html). These show that review spam has become a major problem.

Type 2 (reviews on brands only): Those that do not comment on the products in reviews specifically for the products but *only* the brands, the manufacturers or the sellers of the products. Although they may be useful, we consider them as spam because they are not targeted at the specific products and are often biased.

Type 3 (non-reviews): Those that are non-reviews, which have two main sub-types: (1) advertisements and (2) other irrelevant reviews containing no opinions (e.g., questions, answers, and random texts).

Based on these types of spam, this paper reports a study of review spam detection. 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.

The main contribution of this paper is as follow: It makes the first attempt to investigate opinion spam in reviews and proposes some novel techniques to study spam detection (except some general discussions on the topic in [15]). In general, spam detection can be regarded as a classification problem with two classes, *spam* and *non-spam*. However, due to the specific nature of different types of spam, we have to deal with them differently. For spam reviews of type 2 and type 3, we can detect them based on traditional classification learning using manually labeled spam and non-spam reviews because these two types of spam reviews are recognizable manually. The main task is to find a set of effective features for model building.

However, for the first type of spam, manual labeling by simply reading the reviews is very hard, if not impossible, because a spammer can carefully craft a spam review to promote a target product or to damage the reputation of another product that is just like any other innocent review. We then propose a novel way to study this problem. We first discuss what kinds of reviews are harmful. For example, a spam review that praises a product that every reviewer likes (gives a high rating) is not very damaging. However, a spam review that criticizes a product that most people like can be very harmful. We then want to build a model to analyze only these likely harmful reviews. However, the problem is that there is no labeled training example. Fortunately, we found a large number of duplicate and near-duplicate reviews which are almost certainly spam reviews. Using them to build spam detection models can predict those likely harmful reviews to a great extent. What is even more interesting is that we also found a group of reviewers who might have written many spam reviews.

2. RELATED WORK

Analysis of on-line opinions became a popular research topic recently. As we mentioned in the previous section, current studies are mainly focused on mining opinions in reviews and/or classify reviews as positive or negative based on the sentiments of the reviewers [7, 12, 15, 29, 19, 22]. This paper focuses on studying opinion spam activities in reviews.

Since our objective is to detect spam activities in reviews, we discuss some existing work on spam research. Perhaps, the most extensively studied topic on spam is Web spam. The objective of Web spam is to make search engines to rank the target pages high in order to attract people to visit these pages. Web spam can be categorized into two main types: *content spam* and *link spam*. Link spam is spam on hyperlinks, which does not exist in reviews as there is usually no link among them. Content spam tries to add irrelevant or remotely relevant words in target pages to fool search engines to rank the target pages high. Many researchers have studied this problem [e.g., 3, 5, 9, 10, 11, 12, 16, 23, 24, 25, 26]. Review spam is quite different. Adding irrelevant words is of little help. Instead, spammers write undeserving positive reviews to promote their target objects and/or malicious negative reviews to damage the reputation of some other target objects.

Another related research is email spam [8, 14, 21], which is also quite different from review spam. Email spam usually refers to unsolicited commercial advertisements. Although exist, advertisements in reviews are not as frequent as in emails. They are also relatively easy to detect (see Section 4.2). Untruthful opinion spam is much harder to deal with.

Recent studies on spam also extended to recommender systems, where they are called *attacks* [17]. Although the objectives of attacks to recommender systems are similar to review spam, their basic ideas are quite different. In recommender systems, a spammer injects some attack profiles to the system in order to get some products more (or less) frequently recommended. A profile is a set of ratings (e.g., 1-5) for a series of products. The recommender system uses the profiles to predict product rating of a single user or a group of users. The spammer usually does not see other users' rating profiles. In the context of product reviews, there is no concept of profiles. Each review is only for a particular product, and is not used for any prediction. Also, the reviewer can see all reviews for every product. Rating is only part of a review and another main part is the review text. [27] studies the utility of reviews based on natural language features. Spam is a much broader concept involving all types of objectionable activities. Our work in [14] introduced the problem of review spam, and categorized different types of spam reviews. However, it did little study on detecting untruthful reviews/opinions.

3. OPINION DATA AND ANALYSIS

Before discussing how to detect opinion spam, let us first describe the data used in this study and show some behaviors of the data.

3.1 Review Data from Amazon.com

In this work, we use reviews from amazon.com. The reason for

Table 1. Various features of different categories of products

	Number of Reviewed			Total
Category	**Reviews**	**Products**	**Reviewers**	**Products**
All	5838032	1195133	2146048	6272502
Books	2493087	637120	1076746	1185467
Music	1327456	221432	503884	888327
DVD/VHS	633678	60292	250693	157245
mProducts	228422	36692	165608	901913

using this data set 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. It is thus reasonable to consider it as a representative ecommerce site. Our review data set was crawled from amazon.com in June 2006. We were able to extract 5.8 million reviews, 2.14 reviewers and 6.7 million products (the exact number of products offered by amazon.com could be much higher since it only displays a maximum of 9600 products for each sub-category).

Each amazon.com's review consists of 8 parts

> *<Product ID> <Reviewer ID> <Rating> <Date> <Review Title> <Review Body> <Number of Helpful Feedbacks> <Number of Feedbacks>*

We used 4 main categories of products in our study, i.e., *Books*, *Music*, *DVD* and *mProducts* (industry manufactured products like electronics, computers, etc). The numbers of reviews, reviewed products and reviewers in each category in our study are given in Table 1. These major categories were selected based on the number of reviewed products that they have. Categories like *Furniture & Décor* which has around 60000 products (the 4th largest) but only 2100 reviewed products, were not included.

3.2 Reviews, Reviewers and Products

Before studying the review spam, let us first have some basic information about reviews, reviewers, products, ratings and feedback on reviews. We first look at reviews, reviewers and products. Specifically, we show the following plots:

1. Number of reviews vs. number of reviewers
2. Number of reviews vs. number of products

Note that we do not show "number of reviewers vs. number of products" as it is almost the same as (2) above because all reviews for each product were written by distinctive reviewers (although there are some duplicate reviews for a product as we will see in Sections 4 and 5 when we analyze spam activities in reviews).

Not surprisingly, these relationships all follow the power law distribution. A power law relationship between two quantities x and y can be written as

$$y = ax^k ,$$

where a and k are constants. If we take the log on both sides, we obtain a straight line on a log-log plot.

Figure 1 shows the log-log plot of "number of reviews vs. number of reviewers". We can see that a large number of reviewers write only a few reviews, and a few reviewers write a large number of reviews. There are 2 reviewers with more than 15,000 reviews, and 68% of reviewers wrote only 1 review. Only 8% of reviewers wrote at least 5 reviews. Figure 2 shows the log-log plot of "number of reviews vs. number of products". Again, we can see that a large number of products get very few reviews and a small number of products get a large number of reviews. For example, 50% of products have only 1 review. Only 19% of the products have at least 5 reviews.

In fact, the relationship between the number of feedbacks (readers give to reviews to indicate whether they are helpful) and the number of reviews also closely follows the power law distribution. Figure 3 gives this plot. The graph is slightly lower for the first few points (compared to an ideal straight line), which has mainly reviews with fewer than 5 feedbacks. We can see that

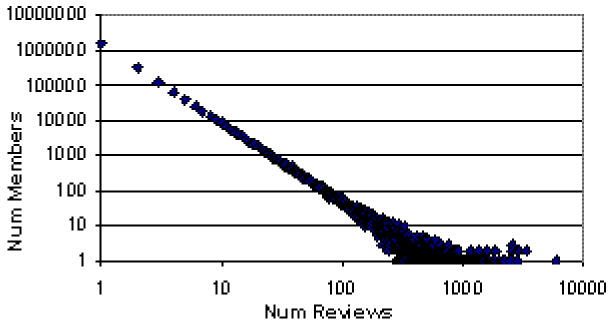


Figure 1. Log-log plot of number of reviews to number of members for amazon.

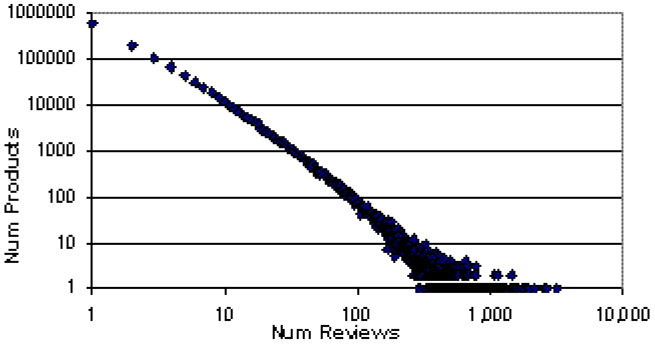


Figure 2. Log-log plot of number of reviews to number of products for amazon.

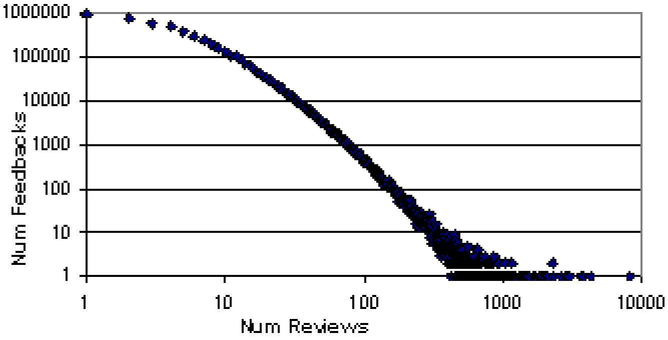


Figure 3. Log-log plot of number of reviews to number of feedbacks for amazon.

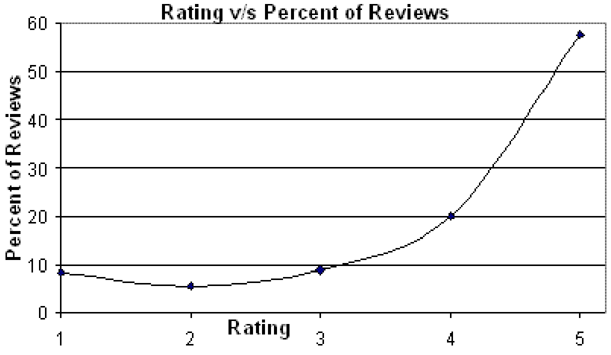


Figure 4. Rating vs. percent of reviews

a large number of reviews get a very small number of feedbacks and a small number of reviews get a large number of feedbacks.

3.3 Review Ratings and Feedbacks

The review rating and the review feedback are two of the most important items in reviews. This section briefly discusses these two items.

Review Rating: Amazon uses a 5-point rating scale with 1 being the worst and 5 being the best. A majority of reviews have very high ratings. Figure 4 shows the rating distribution. On amazon, 60% of the reviews have a rating of 5.0. Since most of the reviews have high ratings, most of the products and members also have a high average rating. Roughly 45% of products and 59% of members have an average rating of 5, which means that the rating of every review for these products and members is 5.

Review Feedbacks: Amazon allows readers to provide helpfulness feedback to each review. As we see above, the number of feedbacks on reviews follows a long tail distribution (Figure 3). On average, a review gets 7 feedbacks. The percentage of positive feedbacks of a review decreases rapidly from the first review of a product to the last. It falls from 80% for the 1st review to 70% for the 10th review. This shows that the first few reviews can be very influential in deciding the sale of a product.

Apart from rating and feedback, review body, review title and review length are also important items. Due to space limitations, we are unable to present their analyses. A detailed review centric and reviewer centric analysis is given in our technical report [13], which also includes analysis of various other interesting features, e.g., rating deviations, reviewer ranking, etc.

4. SPAM DETECTION

We now discuss how to detect the three types of spam reviews described in the introduction section. In general, spam detection can be regarded as a classification problem with two classes, *spam* and *non-spam*. Machine learning models can be built to classify each review as spam or non-spam, or to give a probability likelihood of each review being a spam. To build a classification model, we need labeled training examples of both spam reviews and non-spam reviews. That is where we have a problem.

For the three types of spam, we can only manually label training examples for spam reviews of type 2 and type 3 as they are recognizable based on the content of a review. However, recognizing whether a review is an untruthful opinion spam (type 1) is extremely difficult by manually reading the review because one can carefully craft a spam review which is just like any other innocent review. We tried to read a large number of reviews and were unable to reliably identify type 1 spam reviews manually. Thus, other ways have to be explored in order to find training examples for detecting possible type 1 spam reviews.

Interestingly, in our analysis, we found a large number of duplicate and near-duplicate reviews. Our manual inspection of such reviews shows that they definitely contain some type 2 and type 3 spam reviews. We are also sure that they contain type 1 spam reviews because of the following types of duplicates (the duplicates include near-duplicates):

1. Duplicates from different userids on the same product.
2. Duplicates from the same userid on different products.
3. Duplicates from different userids on different products.

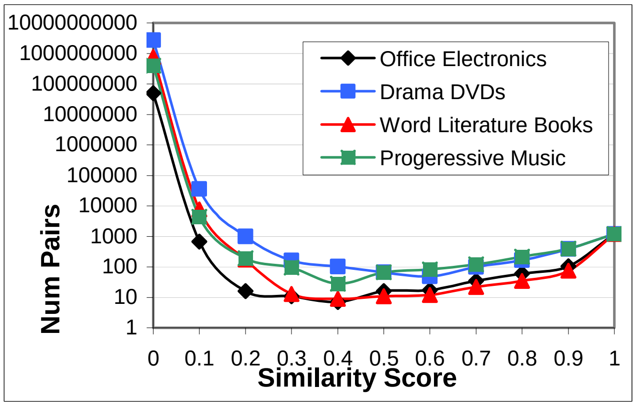


Figure 5. Similarity score and number of pairs of reviews for different sub-categories. Points on X axis are intervals. For example, 0.5 means between interval [0.5, 0.6).

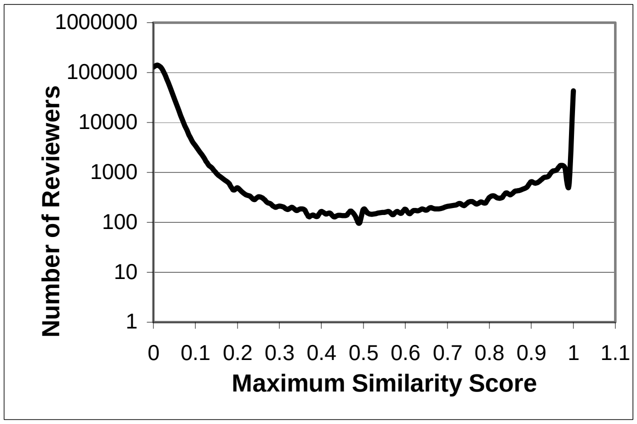


Figure 6. Maximum similarity score and number of members.

Most of such reviews (with type 2 and type 3 spam excluded) are almost certainly untruthful opinion spam (type 1). Note that duplicates from the same user on the same product may not be spam as we will see later.

Thus our review spam detection takes the following strategy. First, we detect duplicates and near-duplicates. We then detect spam reviews of type 2 and type 3 based on machine learning and manually labeled examples. Finally, we try to detect untruthful opinion spam (type 1), which exploits the above three types of duplicates and other relevant information.

4.1 Detection of Duplicate Reviews

Duplicate and near-duplicate (not exact copy) reviews can be detected using the shingle method in [4]. In this work, we use 2-gram based review content comparison. The similarity score of two reviews is the ratio of intersection of their 2-grams to the union of their 2-grams of the two reviews, which is usually called the *Jaccard distance* [6]. Review pairs with similarity score of at least 90% were chosen as duplicates.

Figure 5 plots the log of the number of review pairs with the similarity score for four different sub-categories: each belonging to one of the four major categories *books*, *music*, *DVDs* and *mProducts*. The sub-categories are *word literature* (305894 reviews), *progressive music* (65682 reviews), *drama* (177414 reviews), and *office electronic products* (22020 reviews). All the

Table 2. Three types of duplicate spam reviews on all products and on category *mProducts*

	Spam Review Type	**Num Reviews (*mProducts*)**
1	Different userids on the same product	3067 (104)
2	Same userid on different products	50869 (4270)
3	Different userids on different products	1383 (114)
	Total	55319 (4488)

sub-categories behave almost identically to each other. We also compared the reviews of other sub-categories. The behaviors are about the same. Due to space limitations, we are unable to show all of them. Note that it does not make much sense to use larger categories because they contain completely different products and their reviews are obviously very different.

From Figure 5, we observe that the number of pairs decreases as the similarity score increases. It rises after the similarity score of 0.5 and 0.6. The rise is mostly due to the cases that people copied their reviews on one product to another or to the same product (with minor changes).

Figure 6 plots the log of the number of reviewers with the maximum similarity score. The maximum similarity score is the maximum of similarity scores between different reviews of a reviewer. For 90% of the reviewers with more than one review, the maximum similarity score is less than 0.1 (10%), since they reviewed different products. The number of reviewers increases after the maximum similarity score of 0.6. 6% of the reviewers with more than one review have a maximum similarity score of 1, which is a sudden jump indicating that many reviewers copy reviews. In roughly half of the cases, a reviewer submitted the same review multiple times for a product. There were also some cases of different people (or the same people with multiple userids) writing similar reviews on the same or different products, though small in number.

Roughly 10% of the reviewers with more than 1 review wrote more than one review on at least one product. In 40% of these cases, the reviews were written on the same day with the same rating, body and title (exact duplicates). In 30% of the cases reviews were written on the same day but had some other attributes that are different. In 8% of the cases, a person wrote more than 2 reviews on a product.

Note that in many cases if a person has more than one review on a product, most of these reviews are exact duplicates. However, we do not regard them as spam as they could be due to clicking the submit button more than once. We checked the amazon.com site and found that this was indeed possible. Some others are also due to correction of mistakes in previous submissions.

For spam removal, we can delete all duplicate reviews which belong to any one of the three types described above, i.e., (1) duplicates from different userids on the same product, (2) duplicates from the same userid on different products, or (3) duplicates from different userids on different products. For other kinds of duplicates, we may want to keep only the last copy and remove the rest. Table 2 shows the numbers of reviews in the above three categories. The first number of the second column of each row is the number of such reviews in the whole review database. The second number within "()" is the number of such cases in the category *mProducts*. In the following study, we focus only on reviews in the category of *mProduct*, which has 228422 reviews. Reviews in other categories can be studied similarly.

4.2 Detecting Type 2 & Type 3 Spam Reviews

Duplicates only cover part of spam reviews. Many reviews that are not duplicated are also spam. We now detect type 2 and type 3 spam. As we mentioned earlier, these two types of reviews are recognizable manually. Thus, we employ the classification learning approach based on manually labeled examples.

We manually labeled 470 spam reviews of the two types. The breakdown is given in column 2 of Table 3 in Section 4.2.3. We did not label more as the proportion of such reviews is very small. Manually labeling them is extremely time-consuming. However, we are already able to achieve very good classification results (see Section 4.2.3). Below, we first introduce the classification algorithm that we use and also the features.

4.2.1 Model Building Using Logistic Regression

For model building, we used *logistic regression*. The reason for using logistic regression is that it produces a probability estimate of each review being a spam, which is desirable. In practice, the probabilistic output of logistic regression can be used in many ways in applications. For example, we can use the probability to weight each review. Since the probability reflects the likelihood that a review is a spam, those reviews with high probabilities can be weighted down to reduce their effects on opinion mining. No commitment is made on whether a review is a spam or not.

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 results. AUC is a standard measure used in machine learning for assessing the model quality.

Apart from using logistic regression, we also tried SVM, and naïve Bayesian classification, but they do not perform as well.

To build a model, we need to create the training data. We have defined a large set of features to characterize reviews. We describe them below.

4.2.2 Feature Identification and Construction

There are three main types of information related to a review:

(1) the content of the review,

(2) the reviewer who wrote the review, and

(3) the product being reviewed.

We thus have three types of features: (1) review centric features, (2) reviewer centric features, and (3) product centric features. As their names suggest, review centric features are characteristics of reviews, reviewer centric features are characteristics of reviewers and product centric features are information about the product. For some features, we divide products and reviews into three types based on their average ratings (rating scale: 1 to 5):

Good (rating ≥ 4), *bad* (rating ≤ 2.5) and *Average*, otherwise

4.2.2.1 Review Centric Features

We included the following features.

1. Number of feedbacks (F1), number of helpful feedbacks (F2) and percent of helpful feedbacks (F3) that the review gets. Intuitively, feedbacks are useful in judging the review quality.

2. Length of the review title (F4) and length of review body (F5). These features were chosen since longer reviews tend to get more helpful feedbacks and customer's attention. So, a spammer might want to use this to his/her advantage.
3. Position of the review in the reviews of a product sorted by date, in both ascending (F6) and descending (F7) order. This feature was chosen because we found that reviews which were written early tend to get more user attention, and thus can have bigger impact on the sale of a product. We also use binary feature to indicate if a review is the first review (F8) or the only review (F9).
4. Textual features:

   a. Percent of positive (F10) and negative (F11) opinion-bearing words in the review, e.g., "beautiful", "great", "bad" and "poor". Many researchers have compiled such lists for opinion or sentiment classification. We obtained some of the words from the authors of [12]. We then added a large number of other words of our own. Type 2 reviews would use these words excessively to praise or to criticize the brand or the manufacturer.

   b. Cosine similarity (F12) of the review and product features (which are obtained from the product description page at the amazon.com site). This feature is useful for detecting type 3 reviews, particularly advertisements.

   c. Percent of times brand name (F13) is mentioned in the review. This feature was used for reviews which praise or criticize the brand.

   d. Percent of numerals (F14), capitals (F15) and all capital (F16) words in the review. These features are useful for detecting non-reviews. Excessive use of numerals signifies too much technical detail. Capitals and all capitals signify poorly written and unrelated reviews.

5. Rating related features

   a. Rating (F17) of the review and its deviation (F18) from product rating. Feature indicating if the review is good, average or bad (F19).

   b. Binary features indicating whether a bad review was written just after the first good review of the product and vice versa (F20, F21). A spammer might have written such reviews to do damage control.

4.2.2.2 Reviewer Centric Features

We included the following features:

1. Ratio of the number of reviews that the reviewer wrote which were the first reviews (F22) of the products to the total number of reviews that he/she wrote, and ratio of the number of cases in which he/she was the only reviewer (F23).
2. Rating related features: average rating given by reviewer (F24), standard deviation in rating (F25) and a feature indicating if the reviewer always gave only good, average or bad rating (F26). The first two features are obvious. The third is for reviewers who write only type 3 spam, they tend to give the same rating to all products they review to save time.
3. Binary features indicating whether the reviewer gave more than one type of rating, i.e. good, average and bad. There are four cases: a reviewer gave both good and bad ratings (F27), good rating and average rating (F28), bad rating and average rating (F29) and all three ratings (F30). These four features are for the cases where a reviewer praises products of some brand, but criticizes the products of a competitor brand.
4. Percent of times that a reviewer wrote a review with binary features F20 (F31) and F21 (F32).

4.2.2.3 Product Centric Features

The product related features are as follows:

1. Price (F33) of the product.
2. Sales rank (F34) of the product. Amazon assigns sales rank to "now selling products", which is updated every hour. The sales rank is calculated based on some combination of recent and historic sales of the product.

These features were helpful since spams could be concentrated on cheap/expensive or less selling products.

3. Average rating (F35) and standard deviation in ratings (F36) of the reviews on the product.

4.2.3 Results of Type 1 and Type 2 Spam Detection

We run logistic regression on the data using 470 spam reviews for positive class and rest of the reviews for negative class. The average AUC values based on 10-fold cross validation are given in Table 3.

Table 3. AUC values for different types of spam

Spam Type	**Num reviews**	**AUC**	**AUC – text features only**	**AUC – w/o feedbacks**
Types 2 & 3	470	98.7%	90%	98%
Type 2 only	221	98.5%	88%	98%
Type 3 only	249	99.0%	92%	98%

Some observations are:

1. The average AUC for all spam types is 98.7%. Using only text features also performs well, but poorer than using all features, stressing the value of meta-features. Without using feedback features, the same results can be achieved. This is important because feedbacks can be spammed too (see Section 5.2.3).
2. The algorithm performs equally well on types 2 and 3 spam.
3. One reason that the numbers are very high is that these reviews are not sophisticated, which means that the reviewers did not make much effort to hide such types of reviews.

5. ANALYSIS OF TYPE 1 SPAM REVIEWS

From the presentation of Section 4, we can conclude that type 2 and types 3 spam reviews are fairly easy to detect. Duplicates are easily found too. Detecting type 1 spam reviews is, however, very difficult as they are not easily recognizable manually. Thus, we do not have manually labeled training data for learning. In order to study type 1 spam reviews, let us analyze what kinds of reviews are harmful and are likely to be spammed.

Table 4. Spam reviews vs. product quality

	Positive spam review	Negative spam review
Good quality product	1	**2**
Bad quality product	**3**	4
Average quality product	**5**	**6**

Let us recall what type 1 spam reviews aim to achieve:

1. To promote some target objects, e.g., one's own products (*hype spam*).
2. To damage the reputation of some other target objects, e.g., products of one's competitors (*defaming spam*).

To achieve the above objectives, the spammer usually takes both or one of the actions: (1) write undeserving positive reviews for the target objects in order to promote them; (2) write malicious negative reviews for the target objects to damage their reputation.

Table 4 gives a simple view of type 1 spam. Spam reviews in regions 1, 3 and 5 are typically written by manufacturers of the product or persons with direct economic or other interests in the product. Their goal is to promote the product. Although opinions expressed in region 1 may be true, reviewers do not announce their conflict of interests. Note that good, bad and average products can be defined based on average ratings given to products.

Spam reviews in regions 2, 4, and 6 are likely to be written by competitors. Although opinions in reviews of region 4 may be true, reviewers do not announce their conflict of interests and have malicious intensions.

Clearly, spam reviews in region 1 and 4 are not so damaging, while spam reviews in regions 2, 3, 5 and 6 are very harmful. Thus, spam detection techniques should focus on identifying reviews in these regions.

5.1 Making Use of Duplicates

We would like to build a classification model to detect type 1 spam. Since we have no manually labeled training data, we have to look from other sources. A natural choice is the three types of duplicates discussed in Section 4, which are almost certainly spam reviews.

Note that in some cases, the same person writes the same review for different versions of the same product (hardcover and paper cover of the same book) may not be spam. Out of the total of 4488 reviews, about 30% of them are from reviewers on more than one product. We manually checked the products which had exactly the same reviews. We found that these products have at least one feature different, e.g., two televisions with different dimensions, two iPods with different colors, etc. We labeled them as the same or different products based on the significance of the features that are different. Only in 3% of the cases, the products were labeled as the same. Since the number of such products is very small and many duplicate reviews on such products are also quite suspicious, we thus consider all such duplicates as spam.

We thus propose to treat all duplicate spam reviews as positive examples, and the rest of the reviews as negative examples. Since the number of such duplicate spam reviews is quite large, it is reasonable to assume that they are a fairly good and random sample of many kinds of spam. We then use them to learn a model to identify non-duplicate reviews with similar characteristics, which are likely to be spam reviews.

5.1.1 Model Building Using Duplicates

To ensure that duplicates can be used in prediction, we need to be sure that the models built based on them are indeed predictive. We thus performed experiments using duplicates as positive training examples and the rest of the reviews as negative training examples to build logistic regression models. In model building, we only use reviews from the category *mProducts*. Thus our data set has 4488 duplicate spam reviews (Table 2) and 218514 other reviews. We performed 10-fold cross validation on the data. It gives us the average AUC value of 78% (Table 5) using all the features described in Section 4.2.2 (no feature overfit duplicates). This AUC value is quite high considering that many non-duplicate reviews may be spam and thus have similar probabilities as spam reviews. Table 5 also gives the average AUC values of different feature combinations. Review centric features are most helpful. Using only text features gives only 63% AUC, which shows that it is very difficult to identify spam reviews using text content alone. Combining all the features gives the best result. This demonstrates that duplicates are predictable.

Table 5. AUC values on duplicate spam reviews.

Features used	AUC
All features	78%
Only review features	75%
Only reviewer features	72.5%
Without feedback features	77%
Only text features	63%

Of course, building the logistic regression model using duplicates and non-duplicates is not for detecting duplicate spam because duplicates can be identified easily using content comparison (see Section 4.1). Our real purpose is to use the model to identify type 1 spam reviews that are not duplicated. The above experiment results show that the model is predictive of duplicate spam. To further confirm its predictability, we want to show that it is also predictive of reviews that are more likely to be spam and are not duplicated, i.e., *outlier reviews*. That is, we use the logistic regression model to check whether it can predict outlier reviews.

Outlier reviews are those whose ratings deviate from the average product rating a great deal. They are more likely to be spam reviews than average reviews because high rating deviation is a *necessary condition* for a *harmful* spam review (regions 2, 3, 5 and 6 in Table 4) but not *sufficient* because some reviewers may truly have different views from others. Thus, spam reviews are mostly those with outlier ratings (of course, the converse is not true) and fall in those harmful regions of Table 4. Note that a person may write a spam review with a good (bad) rating to a good (bad) product so that his/her review will fall in region 1 (4) to escape being detected as an outlier based on rating, but gives a bad (good) review in terms of its content. Such cases are not likely to be many because reviews are also read by human users who can easily identify such reviews as spam. Sentiment classification techniques [22] may be used to automatically assign a rating to a review solely based on its review content.

If our classification model built based on duplicates can predict outlier reviews to a great extent (high lift, see below), we will be able to announce with some level of confidence that the logistic regression model built using duplicate spam reviews can be used to predict spam reviews that are not duplicated. For the following predictions, the test data set, which is not used in training, consists of only those non-duplicated reviews.

5.1.2 Predicting Outlier Reviews

To show the effectiveness of the prediction, let us use lift curves to visualize the results. The lift curve is commonly used in data mining for marketing types of applications with highly skewed class distribution. That is, the minority class is usually only a very small percentage of the data, e.g., 1-2% or less. This is quite suitable for spam detection because spam reviews are minorities.

In lift analysis, the classification model first gives each test instance a probability estimate for being a positive class instance, e.g., a spam review in our case. All the test cases are then ranked according to their probability estimates. The data is then divided into *n* equal-sized bins (*n* = 10 is commonly used). The lift curve is drawn as follows: Assume we are interested in reviews of type *T*. The X-axis shows the X% of reviews in 10 bins. That is, X-axis has 10 bins representing 10%, 20%, …, 100% of the test data (see Figure 7). Y-axis shows the cumulated percentage of reviews of type *T* (positive instances) from the first bin (10% in X-axis) to the current bin. The base case is the random distribution, which is represented by a straight line of 45 degrees, which means that top X% of test reviews will also have X% of the reviews belonging to type *T* (the positive class) (see Figure 7). Below, we will use the example in Figure 7 to provide more explanations.

Let us come back to outliers. Outliers cover reviews in regions 2, 3, 5 and 6 of Table 4. In this study, only reviews of products with at least 5 reviews were considered to get reliable deviations from the average product ratings.

We also discuss behavior of reviews where their reviewers may be *biased* towards some product brands. If a reviewer wrote more than one review on a brand and all the reviews have either positive or negative deviations, then the reviewer is considered biased towards that brand. Negative deviation is considered as less than -1 from the mean rating and positive deviation as larger than +1 from the mean rating of the product, which is enough for a review to fall into a dangerous region. Spammers may not want their review ratings to deviate too much from the norm to make the reviews too suspicious.

Figure 7 plots 5 lift curves for the following types of reviews.

- Negative deviation (num = 21005)
- Positive deviation (num = 20007)
- Negative deviation on the same brand (num = 378)
- Positive deviation on the same brand (num = 755)
- Bad product with average reviews (num = 192)

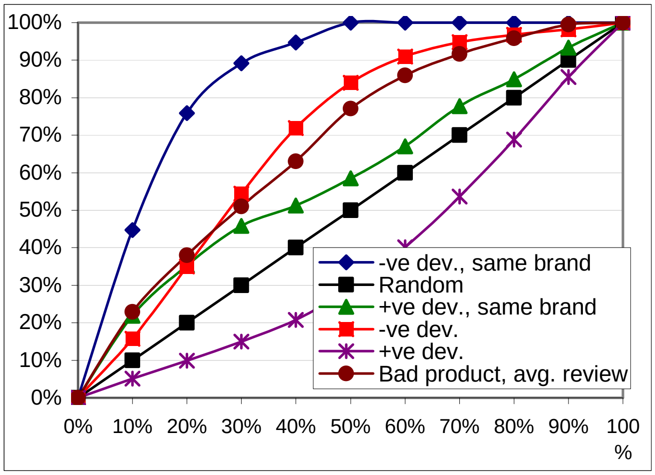


Figure 7: Lift curves for reviews with positive and negative deviations. "-ve (+ve) dev., same brand" means those reviews where member wrote multiple reviews on same brand and all reviews have negative (positive) deviation.

Let us look at the highest curve in Figure 7 which is for negative deviation on the same brand "-ve dev., same brand". If the prediction is completely random, we get the baseline (45 degree line marked *random*), i.e., outlier reviews are randomly distributed in all the bins (recall the percentage in Y-axis is the cumulative value). If the model is able to predict, then the curve should go up. For example, in the first bin (10%), the random distribution should also contain 10% outlier reviews. However, in this case the model is able to do much better. It, in fact, catches 44% of outlier reviews. By bin 2 (20% on X-axis), the model has already caught 75% of outliers. This prediction power is quite remarkable, which shows that the model can predict outlier reviews of this type quite accurately.

Let us make several important observations:

1. Reviews with negative (-ve) deviation (in rating) on the same brands give the highest lift curve. This means reviewers who are biased towards a brand and give reviews with negative deviations on products of that brand are very likely to be spammers. Those highly ranked reviews from such reviewers should be removed as spam reviews.
2. The lift curve of reviews with positive (+ve) deviations (in rating) on the same brands are much lower, which means that such reviews are much less likely to be spammed. Reviews with negative deviations of all types have similar behaviors.
3. Interestingly, reviews with positive (+ve) deviations of all types are very unlikely to be spammed as the lift curve for such reviews is actually below the random line (the baseline). The only case where reviews with positive deviations are above the random line is for bad products with average ratings (defined in section 4.2.2). These reviews lie in region 3 (Table 4), which shows that spammers try to promote bad products but not giving them too high ratings, which is intuitive.

   It is important to note that the lift curve of +ve deviation being below the random line does not mean that our model is completely useless or wrong because the model was not built using training data from +ve deviation. Instead, the model is built using duplicates. It thus only means that the model ranks +ve deviation reviews low, from which we infer that such reviews are less likely to be spam reviews.

Summary: This set of experiments shows that the model built using duplicated spam as positive data is also predictive of non-duplicated spam reviews to a good extent. If we accept this conclusion, then according to observations 1 and 2 above, we can conclude that reviews in regions 2 and 6 of Table 4 are targets of spam. Biased reviews which deviate from product ratings are more likely to be spam as compared to other reviews. That is, people who write multiple reviews on one brand which are all negative are very likely to be spammers. Reviews in regions 3 and 5 are not heavily spammed, although there are some activities.

5.2 Some Other Interesting Reviews

As we believe that our logistic model is predictive of spam in non-duplicated reviews, let us now use it to analyze some other interesting reviews: 1) reviews that are the only reviews of some products, 2) reviews from reviewers of different ranks, 3) reviews with different levels of feedbacks, 4) reviews of products with varied sales ranks. Some interesting conclusions are drawn.

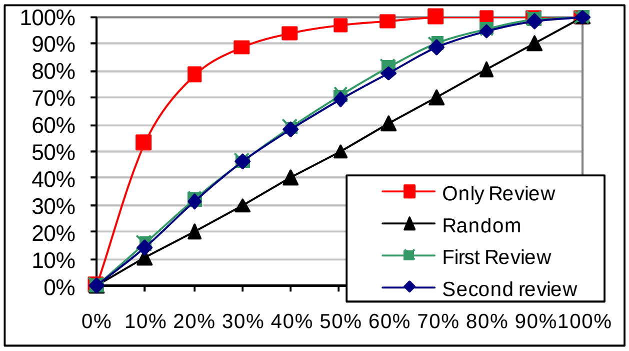


Figure 8: Lift curves for only reviews, first and second reviews of products.

5.2.1 *Only Reviews*

A large number of products have only one review as the number of reviews on a product follows the power law distribution. For example, for manufactured products (*mProduct*), 46% of the reviewed products have only one review. For a customer it is very difficult to tell if that review is trustworthy since there is no other review on the same product to compare with. In this section, we discuss whether such reviews can be trusted. We compare the lift curve for reviews of products that has only one review with the first and second reviews of products with more than one review. Ideally, they should behave the same. However, they do not.

Figure 8 plots the lift curves for the following types of reviews.

- Only reviews of the products (num = 7330)
- The first reviews of the products (num = 9855)
- The second reviews of the products (num = 9834)

We did not use any position features of reviews (F6, F7, F8, F9, F20 and F21) and number of reviews of product (F12, F18, F35, and F36) related features in model building to prevent overfitting.

Figure 8 shows that reviews that are the only reviews of products have a very high lift curve compared to the first and second reviews (which are also somewhat spammed). The model catches 52% of only reviews in the first bin (top 10%) and by bin 2 (top 20%) the model has caught 77% of such reviews. This shows that only reviews are very likely to be candidates of spam.

5.2.2 *Reviews from Top-Ranked Reviewers*

Amazon.com gives ranks to each member/reviewer based on the frequency that he/she gets helpful feedbacks on his/her reviews. Amazon only displays member rank of the reviewer next to his/her review on the product page for the top 1000 ranked members. The ranks obviously have an effect on the customer since he/she is more inclined to read and trust reviews written by a top-ranked reviewer. In this section, we discuss whether ranking of reviewers is effective in keeping spammers out.

Since our review features do not contain any reviewer rank related features, so we use all of them in model building. Figure 9 plots the lift curve for reviews written by reviewers who belong to the following reviewer ranks:

- Top 100 rank (num reviews = 609)
- Between 100 to 1000 ranks (num reviews = 1413)
- After more than 2 million rank (num reviews = 1771), which is the tail of members/reviewers ordered by ranks.

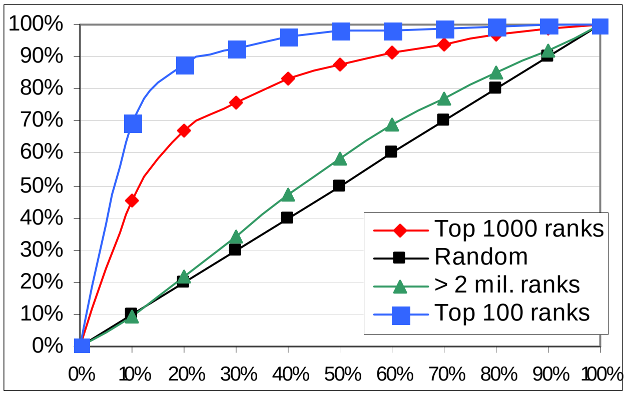


Figure 9: Lift curves for reviewers of different ranks.

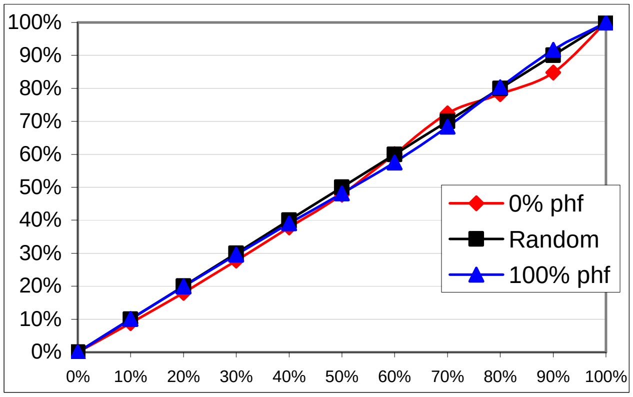


Figure 10: Lift curves for reviews with 0 and 100% positive feedbacks with minimum of 1 feedback.

Figure 9 shows that reviews written by top-ranked reviewers are given higher probabilities of being spam as compared to reviews by bottom ranked reviewers. The model catches 69% of reviews by top 100 ranked reviewers in the first bin (top 10%) itself and it reaches 87% by the second bin (top 20%). This is a very surprising result because it shows that top ranked reviewers are less trustworthy as compared to bottom ranked reviewers.

Further analysis shows top-ranked reviewers score higher than low ranked ones on following indicators of reviews being spam:

1. Top-ranked reviewers generally write a large number of reviews, much more than bottom ranked reviewers. People who wrote a large number of reviews are natural suspects. Many wrote hundreds and thousands of reviews, which is unlikely for an ordinary consumer. Note that the number of reviews for top-ranked reviewers is not large as shown above because we only studied reviews of manufactured products. Those top reviewers typically also reviewed a large number of products or items in other categories, e.g., *books* and *music*.
2. Top ranked reviewers also score high on some important indicators of spam reviews. For example, they often deviate a lot from the average product rating and write much more reviews that are the only reviews of the products. We have showed in Section 5.1.2 and 5.2.1 that such reviews are more likely to be candidates of spam.

5.2.3 Reviews with Different Levels of Feedbacks

Apart from reviewer rank, another feature that a customer looks at before reading a review is the percent of helpful feedbacks that the review gets. Amazon uses feedbacks to find spotlight reviews, which appear right next to product description separated from other reviews. We now discuss whether the percentage of positive feedbacks that a review gets is helpful in filtering out spam.

Reviews with at least 5 feedbacks were considered to get reliable estimate of positive feedbacks. Figure 10 plots the lift curves for reviews with 0% positive feedbacks (num reviews = 638) and 100% positive feedbacks (num reviews = 19237). We excluded feedback features (F1, F2 and F3) in model building for this analysis to prevent overfitting.

The lift curves follow the random distribution. This means that spam reviews can also get good feedbacks and non-spam reviews can also get bad feedbacks. This is a very important result. It shows that if usefulness of a review is defined based on the feedbacks that the review gets [27], it means that people can be readily fooled by a spam review. It also explains the difficulty in labeling spam reviews of type 1 manually.

Feedback spam: Finally we note that feedbacks can be spammed too. Feedback spam is a sub-problem of click fraud in search advertising [16], where a person or robot clicks on some online advertisements to give the impression of real customer clicks. For amazon.com, since a person has to register to actually submit a feedback and the same person cannot feedback on his/her own reviews, ruthless clicking on one review is minimized. However, for Web sites such as CNET where anonymous feedbacks are allowed, feedback spam may affect the model accuracy.

Fortunately, our models perform equally well without using feedback features, thus also showing that feedbacks are not good distinguishing factors for spam and non-spam. The results are given in Table 3 and Table 5,.

5.2.4 Reviews of Products with Varied Sales Ranks

This sub-section discusses if certain products are more likely to be targets of spam compared to others. We excluded both product price (F33) and product sales rank (F34) features for this analysis, to prevent any overfitting in model building using duplicates.

Product sales rank is an important feature for a product. Amazon displays the result of any query for products sorted by their sales ranks. So, we want to know whether there are spam reviews even for high ranked products.

Figure 11 plots the lift curves for reviews of corresponding products with high and low sales ranks.

- Products with sales rank <= 50 (num review = 3009)
- Products with sales rank >= 100K (num reviews = 3751)

We can see that reviews on products with high sales ranks generally have very low level of spam. Its lift curve is even below the random line. This is good news since it shows that spam activities are more limited to low selling products. This is quite intuitive since it is difficult to damage reputation of a high selling or popular product by writing a spam review.

6. CONCLUSIONS

This paper studied opinion spam in reviews, which to the best of our knowledge has not been studied in the literature. The paper first identified three types of spam. Detection of such spam is done first by detecting duplicate reviews. We then detect type 2 and type 3 spam reviews by using supervised learning with manually labeled training examples. Results showed that the logistic regression model is highly effective. However, to detect type 1 opinion spam, the story is quite different because it is very hard to manually label training examples for type 1 spam. We thus proposed to use duplicate spam reviews as positive training examples and other reviews as negative examples to build a model. We showed the effectiveness of the model. The current study, however, only represents an initial investigation. Much work remains to be done. In our future work, we will further improve the detection methods, and also look into spam in other kinds of media, e.g., forums and blogs.

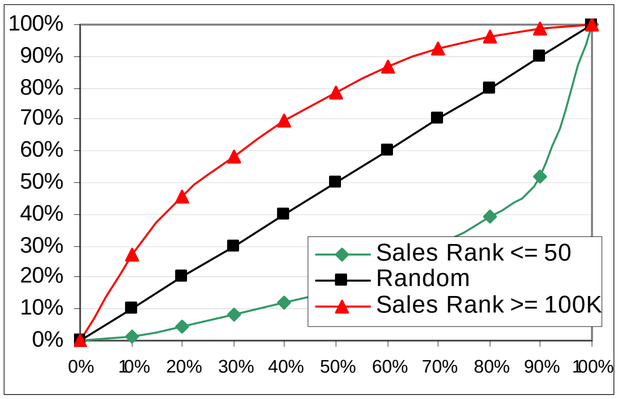


Figure 11: Lift curves for reviews corresponding to products with different sales ranks.

7. ACKNOWLEDGEMENTS

We would like to thank Microsoft Corporation for funding this research and Ling Bao for many helpful discussions. We are also grateful to amazon.com for allowing us to crawl all their reviews.

8. REFERENCES

[1]. E. Amitay, D. Carmel, A. Darlow, R. Lempel & A. Soffer. The connectivity sonar: detecting site functionality by structural patterns. *Hypertext'03*, 2003.

[2]. M. Andreolini, A. Bulgarelli, M. Colajanni & F. Mazzoni. Honeyspam: Honeypots fighting spam at the source. In *Proc. USENIX SRUTI 2005*, Cambridge, MA, July 2005.

[3]. R. Baeza-Yates, C. Castillo & V. Lopez. PageRank increase under different collusion topologies. *AIRWeb'05*, 2005.

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

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

[6]. S. Chakrabarti. *Mining the Web: discovering knowledge from hypertext data*. Morgan Kaufmann, 2003.

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

[8]. I. Fette, N. Sadeh-Koniecpol, A. Tomasic. Learning to Detect Phishing Emails. *WWW2007.*

[9]. D. Fetterly, M. Manasse & M. Najork. Detecting phrase-level duplication on the World Wide Web. *SIGIR'*2005.

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

[11]. M. R. Henzinger: Finding near-duplicate web pages: a large-scale evaluation of algorithms. *SIGIR*'06, 2006.

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

[13]. N. Jindal and B. Liu. Product Review Analysis. Technical Report, UIC, 2007.

[14]. N. Jindal and B. Liu. Analyzing and Detecting Review Spam. *ICDM2007*.

[15]. W. Li, N. Zhong, C. Liu. Combining Multiple Email Filters Based on Multivariate Statistical Analysis. *ISMIS 2006.*

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

[17]. A. Metwally, D. Agrawal, A. Abbadi. DETECTIVES: DETEcting Coalition hiT Inflation attacks in adVertising nEtworks Streams. *WWW2007.*

[18]. B. Mobasher, R. Burke & J. J Sandvig. Model-based collaborative filtering as a defense against profile injection attacks. *AAAI'2006*.

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

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

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

[22]. M. Sahami and S. Dumais and D. Heckerman and E. Horvitz. A Bayesian Approach to Filtering Junk {E}-Mail. *AAAI Technical Report* WS-98-05, 1998.

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

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

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

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

[27]. S. Ye, R. Song, J.-R. Wen, W.-Y. Ma. A Query-dependent duplicate detection approach for large scale search engines. *APWeb'04*, 2004.

[28]. Z. Zhang & B. Varadarajan, Utility scoring of product reviews, *CIKM'2006*.