Association for Information Systems
AIS Electronic Library (AISeL)

ICIS 2010 Proceedings

International Conference on Information Systems (ICIS)

2010

An Empirical Study of Online Consumer Review Spam: A Design Science Approach

Raymond Y.K. Lau
City University of Hong Kong, raylau@cityu.edu.hk

Stephen S.Y. Liao
City University of Hong Kong, issliao@cityu.edu.hk

Kaiquan Xu
City University of Hong Kong, kaiquan.xu@student.cityu.edu.hk

An Empirical Study of Online Consumer Review Spam: A Design Science Approach

Completed Research Paper

Raymond Y. K. Lau
Department of Information Systems
City University of Hong Kong
Tat Chee Avenue, Kowloon,
Hong Kong SAR
raylau@cityu.edu.hk

Stephen S. Y. Liao
Department of Information Systems
City University of Hong Kong
Tat Chee Avenue, Kowloon,
Hong Kong SAR
issliao@cityu.edu.hk

Kaiquan Xu
Department of Information Systems
City University of Hong Kong
Tat Chee Avenue, Kowloon,
Hong Kong SAR
kaiquan.xu@student.cityu.edu.hk

Abstract

Because of the sheer volume of consumer reviews posted to the Internet, a manual approach for the detection and analysis of fake reviews is not practical. However, automated detection of fake reviews is a very challenging research problem given the fact that fake reviews could just look like legitimate reviews. Guided by the design science research methodology, one of the main contributions of our research work is the development of a novel methodology and an instantiation which can effectively detect untruthful consumer reviews. The results of our experiment confirm that the proposed methodology outperforms other well-known baseline methods for detecting untruthful reviews collected from amazon.com. Above all, the designed artifacts enable us to conduct an econometric analysis to examine the impact of fake reviews on product sales. To the best of our knowledge, this is the first empirical study conducted to analyze the economic impact of fake consumer reviews.

Keywords: Online reviews, review spam, spam detection, language modeling, kullback-leibler divergence, econometric analysis, electronic commerce.

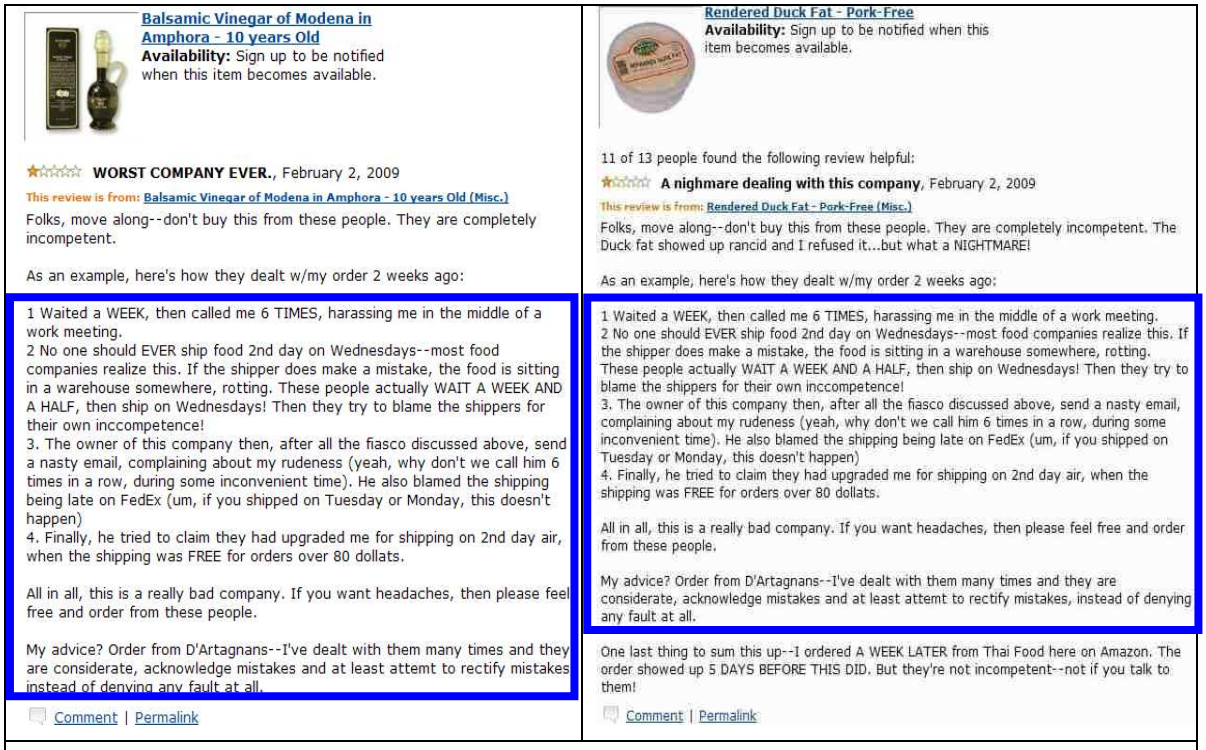
Introduction

In the era of Web 2.0 (Oreilly 2007; Raman 2009), user-contributed data is the norm and there has been an explosive growth of the number of user-generated data such as online consumer reviews posted to the c-Commerce Web sites such as amazon.com, cnet.com, and epinions.com. It has been suggested that user-generated product reviews can be treated as online sellers' free "sales assistants" who can help potential consumers choose products and services best meeting their specific purchasing needs (Wernerfelt 1994). Online consumer reviews are sometimes referred to as electronic word of mouth (eWOM) in the literature (Chen and Xie 2008; Godes and Mayzlin 2004; Mayzlin 2006). According to the result of the 2009 Nielsen global online consumer survey[1] that involved 25,000 respondents from 50 countries, 70% of the respondents said that they would refer to the consumer reviews posted to the Internet before making a purchase. Online consumer reviews or eWOM have been considered one of the most important information sources referred to by consumers or marketers (Dellarocas 2003; Godes and Mayzlin 2004; Mayzlin 2006). Nevertheless, the widespread sharing and utilization of online consumer reviews has

[1] http://en-us.nielsen.com/main/news/news_releases/2007/october/Word-of-Mouth_the_Most_Powerful_Selling_Tool__Nielsen_Global_Survey

also raised the concerns about the trustworthiness of these items (Cheung et al. 2009; Dellarocas 2003; Dellarocas 2006; Mayzlin 2006).

On 14 July 2009, the New York Times has reported a settlement case about shill reviews[2]; a U.S. based cosmetic surgery company has ordered its employees to pretend to be satisfied customers and posted glowing reviews to its own Web sites and other third party opinion sharing Web sites. A recent incident of fake reviews (i.e., spam) involves a senior marketing personnel of a computer retailer who has deliberately posted misleading product reviews to amazon.com in order to promote the sales of the company's backup devices[3]. The problem of fake consumer reviews has affected both individual consumers and firms. For individual consumers, they may purchase some products not really meeting their needs if they refer to shill reviews. For firms, the problem may even be more serious because the sales of their products or services could be reduced due to "bad-mouthing" reviews. In addition, market analysis and sales forecast that are conducted based on fake consumer reviews are unlikely to generate accurate business intelligence to guide future product design or marketing activities. In this paper, we collectively refer to all kind of fake reviews (created intentionally or unintentionally) as spam (Jindal and Liu 2008).



a. Product: B0000ZRGLQ **b. Product: B000GD3SW8**

Figure 1. An Example of Review Spam

Figure 1 shows an example of review spam which involves two near-duplicate reviews for different products sold at amazon.com. The first review (Figure 1a) is about the product "Balsamic Vinegar" (ASIN: B0000ZRGLQ), and the second one (Figure 1b) is about an entirely different product "Rendered Duck Fat" (ASIN: B0000GD3SW8). An ASIN represents Amazon's unique product number. These reviews are almost the same except that the second review has one more sentence in the first paragraph, and an extra paragraph at the end of the review. Despite the different product contexts, both reviews try to promote another seller called "D'Artagnans". As shown in Figure 1b, 10 out of 12 users still find the spam helpful; this observation is consistent with the previous studies that user-generated helpfulness vote may not be a good indicator of a review's true helpfulness nor its trustworthiness (Danescu-Niculescu-Mizil et al. 2009; Jindal and Liu 2008). Obviously, the existing review moderation process

[2] http://www.nytimes.com/2009/07/15/technology/internet/15lift.html

[3] http://pogue.blogs.nytimes.com/2009/01/27/carbonite-stacks-the-deck-on-amazon/

(manual or automatic) adopted at amazon.com is not effective to deal with the review spam problem because this spam has been left online for almost one year! We believe that a manual approach to detect review spam is not feasible because of the problem of information overload (Lau et. al. 2008; Lau and Lai 2008).

One may wonder why there is little research work and very few publications discussing methodologies for the automatic detection of fake reviews. One of the reasons is that it is difficult to find prominent features to separate the spam from the legitimate ones (i.e., ham). Jindal and Liu (2008) have defined three types of review spam, namely, untruthful reviews, non-reviews, and brands only reviews. The reviews shown in Figure 1 are a kind of untruthful reviews because they are intentionally composed to misled human readers. The detection of untruthful reviews is fundamentally different from the detection of email or Web spam which has been extensively studied by researchers in the field of computer science in recent years (Cormack et al. 2007a; Cormack et al. 2007b; Gyöngyi and Garcia-Molina 2005). For email or Web spam, the primary objective of the spammers is to persuade the readers to traverse to the target merchant sites by injecting spam keywords (e.g., "making money $$$") or URLs. Accordingly, there are relatively obvious spam "features" that could be utilized by an email or Web spam detection program to separate the spam from the ham. However, existing supervised machine learning techniques (Chang et al. 2008; Chen et al. 2009; Cormack et al. 2007b) which excel in learning prominent features to classify different classes of objects (e.g., spam and ham emails) may not be effective for untruthful review detection because prominent features may not be available to the supervised classifiers. For instance, it is difficult to identify prominent features from the reviews shown in Figure 1 and use them to distinguish between spam and ham. Given the fact that the detection of untruthful reviews is a more challenging research problem (e.g., they just look like other legitimate reviews) (Jindal and Liu 2008), the main focus of this paper is about the design and development of a novel methodology to detect "untruthful reviews".

Driven by the design science research methodology (Hevner et al. 2004; March and Storey 2008), one of the main contributions of our research work is the design and development of a novel un-supervised detection methodology to combat untruthful reviews. By using an un-supervised spam detection methodology, we can address to the "missing features" problem in untruthful review detection. In particular, a novel semantic language model is designed and applied to estimate the conceptual semantic similarity links among reviews, and hence to identify untruthful reviews. Our proposed semantic language modeling approach for untruthful review detection is different from the traditional plagiarized content detection method in that "substituted" terms can be taken into account when the similarity of review contents is estimated. Another main contribution of our work is the construction of an instantiation of our design such that rigorous experiments can be applied to evaluate the effectiveness of our proposed design. Above all, our design artifacts allow the econometric analysis of fake reviews to be conducted. To the best of our knowledge, this is the first empirical study about the economic impact of fake online consumer reviews.

The managerial implications of our research work are two fold. First, business managers or marketers can apply the proposed methodology to identify and analyze fake reviews related to their products and services. Accordingly, they can develop more effective product design strategies and marketing plans based on the sheer volume of genuine online consumer comments. Second, online merchants can apply our methodology to continuously monitor and moderate the sheer volume of user-contributed online reviews so that the quality of these reviews as well as the popularity of the merchants' Web sites can be enhanced. In addition, our designed artifacts can help individual consumers assess the true quality of products and facilitate their comparison shopping processes.

The rest of the paper is organized as follows. The next section highlights previous research work related to the detection of various review (opinion) spam, email spam, or Web spam in general. An illustration of our novel review spam detection methodology and the underlying computational models is then followed. The evaluation of our designed artifacts and the application of our design to empirically assess the trustworthiness of online consumer reviews are then described. Finally, we offer concluding remarks and describe future directions of our research work.

Related Research Work

Jindal and Liu (2007a, 2007b, 2008) classified three types of review spam, namely untruthful opinions, brands only spam, and non-review. A logistic regression model (a supervised classification method) was applied to detect the three types of fake reviews. Duplicated reviews were first identified, and then these reviews were used as the training examples for the logistic regression classifier. By using a dataset consisting of 218,514 legitimate reviews (ham) and 4,488 duplicated reviews (i.e., spam) from the manufacturing product category of Amazon, the logistic regression model achieved 78% area under curve (AUC) in untruthful review detection. Our work discussed in this

paper focuses more on the first type of opinion spam, that is, untruthful opinions because it is considered a much more difficult detection task (Jindal and Liu 2008). In addition, we propose an unsupervised detection method for untruthful reviews given that it is difficult to manually label training examples. A quantitative and qualitative research method was applied to investigate the online economies of reputation management and user practices in online product reviews at several e-Commerce Websites (David and Pinch 2006). Unfortunately, the details of the computational method and its evaluation were not described in their paper (David and Pinch 2006).

More recently, Xiao and Benbasat (2010) have reported a typology of product-related deceptive information practices that explains how online traders use a variety of ways to deceive consumers at e-Commerce sites. The work reported in this paper can be seen as an extension of the work of Xiao and Benbasat (2010) by developing a novel methodology to combat deception related to online reviews. In the context of trust in online environments, Gefen et al. (2008) distinguished between trust (i.e., trustor's willingness to depend on the trustee) and trustworthiness (i.e., the credibility, ability, and benevolence of the trustee). They further suggested that the text contents of online Web sites helped buyers infer signals of seller's trustworthiness in general. Our research is one of the few empirical studies on evaluating the trustworthiness of the online text contents.

Abbasi et al. (2008) proposed a stylometric approach to identify online traders based on the writing style traces embedded in the traders' online comments or reviews. In particular, the Karhunen-Loeve transform was applied to generate n-dimensional feature vectors called the Writeprint which represented a trader's writing pattern. Their method was evaluated based on 600K online comments contributed by 200 eBay traders (Abbasi et al. 2008). The primary objective of our research is to determine if the online comments are truthful or not; our empirical study was conducted based on 1,484K online reviews contributed by thousands of review writers.

Li et al. (2009) developed a hybrid content-based and link-based approach for patent classification. Essentially, the knowledge evolution process was considered to be performed based on the relationships among individual artifacts (Li et al. 2009). For example, a patent's innovation process could be represented by both the content-based features of cited patents and the structure of a patent citation network. Accordingly, a labeled citation graph kernel was proposed for patent classification. Experimental results confirmed that Support Vector Machines (SVMs) with a labeled citation graph kernel outperformed that with a content-based linear text kernel for patent classification tasks (Li et al. 2009). Arazy and Woo (2007) explored the relationships between terms to extract semantically rich collocations to represent documents, and their experimental results showed that a combination of token-based (i.e., considering terms as independent) and collocation-based document representation can improve retrieval effectiveness. Our proposed methodology for untruthful review detection is based on an unsupervised classification method which explores both the contents of reviews and the "conceptual similarity links" among reviews. A novel semantic language model is developed to estimate the conceptual similarity links among reviews. Similar to the approach of Arazy and Woo (2007), our proposed semantic language model can take into account the term relationships when the generation probability between two reviews is estimated.

A taxonomy of Web spam was developed to analyze the common techniques applied to Web page spamming (Gyöngyi and Garcia-Molina 2005). Web spam refers to Web page created for any deliberate action that is meant to trigger an unjustifiably favorable relevance or importance. According to the random samples crawled from the Web, it was estimated that around 10-15% of the contents on the Web are spam (Gyöngyi and Garcia-Molina 2005). For Web page spamming, both content-based spamming (Ntoulas et al. 2006) and link-based spamming (Zhou and Pei 2009) were examined. These Web spam techniques are mainly used to fool a search engine rather than humans in order to obtain a higher page rank for the target Web pages. It was pointed out that Web spam could be automatically generated by stitching together phrases drawn from a limited corpus based on a variant of the Rabin fingerprints generation method (Gyöngyi and Garcia-Molina 2005). Linguistic features were also examined for Web spam detection (Piskorski et al. 2008).

Blog spam, a special case of Web page spam, was examined by using probabilistic unigram language models and Kullback-Leibler (KL) divergence to distinguish spam blogs (splog) from the normal blog posts (Mishne et al. 2005). Martinez-Romo and Araujo (2009) also employed unigram language models for Web spam detection. Our work reported in this paper deals with a much more challenging research problem because the language usages in legitimate consumer reviews and spam reviews could be quite similar. Lin et al. (2008) employed self-similarity matrices and the histogram intersection similarity measure to analyze the regularities of blog posts over time. The TREC Blog dataset (Macdonald and Ounis 2007) was used to evaluate the effectiveness of their proposed method. Macdonald et al. (2009) discussed the issue of spam for opinionated blog posts although they did not propose a technique to detect splogs automatically.

Research was conducted to examine the helpfulness of user-generated product reviews (Danescu-Niculescu-Mizil et al. 2009, Ghose and Ipeirotis 2007, Kim et al. 2006, Liu et al. 2008). Although the prediction of review helpfulness is related to the identification of review spam, the goals of these two kinds of tasks and the underlying techniques are quite different. Ghose and Ipeirotis (2007) developed two linear models to predict the sales ranks of products and the helpfulness of reviews based on the subjectivity analysis of the review contents. Kim et al. (2006) applied SVM regression model to examine the correlation between the structural, lexical, syntactic, semantic, and meta features of reviews and their helpfulness rating. Unfortunately, these user-generated helpfulness votes could be spam by themselves, and so the accuracy of the supervised classification methods may be affected.

Recent research work has applied the least-square SVM, a supervised learning approach, to conduct co-classification based on both bookmarks and user data to detect spam in social networks (Chen et al. 2009). Chang et al. (2008) proposed a hybrid partitioned logistic regression and naive Bayes classifier to detect email spam. In the context of spam detection for short SMS messages, Cormack et al. (2007a) compared several supervised learning models, including SVM, logistic regression, Dynamic Markov Compression (DMC), and so on. The DMC and the Prediction by Partial Matching (PPM) compression models were evaluated for email spam filtering and they were found more effective than that of the other supervised classifiers (Bratko et al. 2006). However, for review spam detection, it is extremely difficult to find representative training examples to train a classifier no matter it is a compression model or a classical supervised classifier. Zheleva et al. (2008) examined a user-based reputation management system which makes use of the feedback of trustworthy users to detect and remove email spam. Given the fact that it is difficult for human readers to distinguish fake reviews from the legitimate ones (Jindal and Liu 2008), it may not be feasible to rely on user-based reputation management system to identify fake reviews.

Methodology

Research Methodology

Our research work is driven by the "Design Science" research methodology (Hevner et al. 2004, March and Storey 2008). The design science research methodology focuses on the discovery of novel knowledge of a problem domain by the construction and application of "designed artifacts". For our research, the designed artifacts include a methodology for automatic review spam detection, the computational model for the detection of fake reviews, and an instantiation of our design (i.e., a prototype system). The design science research methodology also emphasizes on rigorous evaluation of the designed artifacts. Accordingly, our prototype system is evaluated by applying the TREC (Cormack 2007) like benchmark procedure and the standard effectiveness measures for spam detection systems. Our evaluation dataset is constructed by extracting realistic reviews from e-Commerce Web sites such as amazon.com. Furthermore, our design is developed based on sound theories developed in the fields of statistics such as Kullback-Leibler divergence (Kullback and Leibler 1951) and information theory such as probabilistic language modeling (Liu and Croft 2004; Ponte and Croft 1998). Above all, our designed artifacts can be applied to identify fake reviews so as to empirically assess the trustworthiness of consumer reviews at a typical e-Commerce Web site. Based on the detected fake reviews, an econometric analysis is performed to assess the impact of review spam on actual product sales. By means of this empirical study, we have shown that our designed artifacts can really help address a serious business problem (i.e., the detection of fake reviews) which has not been effectively resolved by existing methods (e.g., the review moderation procedure adopted at typical e-Commerce Web sites). The main research questions of our study are:

Can an effective methodology be developed to automatically detect untruthful consumer reviews?

Can an instantiation of the design be constructed so as to evaluate its effectiveness?

Can the instantiation of the design be applied to solve real-world business problems such as the detection of untruthful reviews and the assessment of the trustworthiness of online consumer reviews at a typical e-Commerce Web site?

What is the economic impact of fake online consumer reviews on product sales?

A Methodology for Automatic Review Spam Detection and Analysis

Our proposed review spam detection methodology supports the automatic detection of three types of spam identified by Jindal and Liu (2008). However, the main focus of this paper is the detection of "untruthful reviews" given that it is a more challenging research problem (Jindal and Liu 2008). An overview of our proposed methodology for

review spam detection is depicted in Figure 2. First, a user (e.g., a business manager or a consumer) selects the scope (e.g., all the archived reviews, all reviews of a product category, or all reviews of a specific product) for review spam detection. This user requirement is then translated into a spam detection query to be processed by the query processor (Task 1 in Figure 2). If the reviews for a product is not yet available in the system's local database or have not been updated for a pre-defined period of time, the Web services or APIs provided by external e-Commerce sites (e.g., amazon.com[4], cnet.com[5], shopping.com[6], etc.) and Internet Search Engines (e.g., Google[7]) will be invoked to retrieve the consumer reviews and the related product information (Task 2 in Figure 2). For our current prototype system (an instantiation of the design), the user can either use the Amazon's ASIN (a unique product identification number) or a product name to compose their query. If comprehensive APIs are not available for directly downloading review contents (e.g., the APIs of cnet.com only supports the extraction of product or product category information), the crawlers of our prototype system will be invoked to retrieve the consumer reviews from the specific Web sites (Task 3 in Figure 2). The crawler programs and the APIs will also be invoked periodically to update the system's local database of consumer reviews.

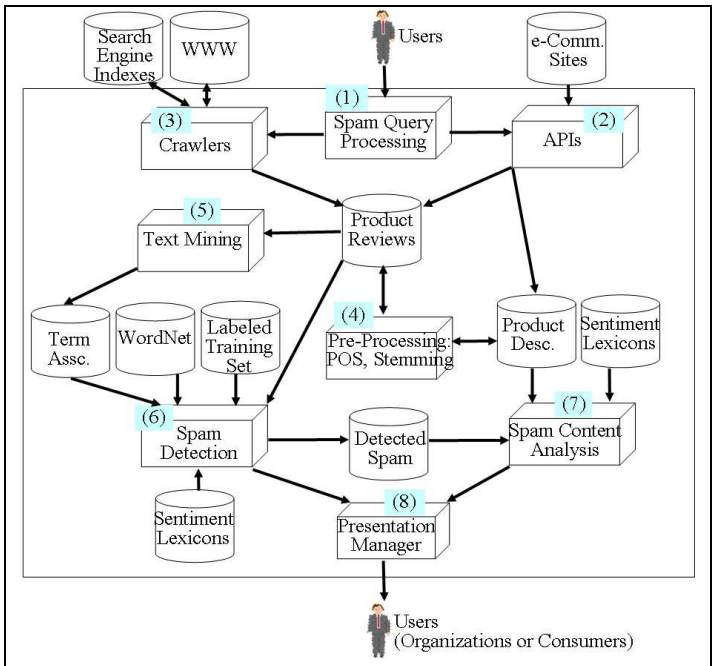


Figure 2. A Methodology for Review Spam Detection

Traditional document pre-processing procedures (Salton et al. 1975; Salton and McGill 1983) such as stop word removal, Part-of-Speech (POS) tagging, and stemming (Porter 1980) are then invoked to process the consumer reviews and product descriptions retrieved from the Web; these reviews and product descriptions are stored in the system's local database (Task 4 in Figure 2). We develop our POS tagger based on the WordNet lexicon (Miller et al. 1990) and the publicly available WordNet API[8]. A context-sensitive text mining module (Lau 2003; Lau et al. 2008) is invoked periodically to extract term association information from the collection of consumer reviews; these term associations will be used by our novel semantic language model to detect fake reviews (Task 5 in Figure 2). Review spam detection (Task 6 in Figure 2) is carried out based on an unsupervised probabilistic language model (for untruthful review detection), and a supervised classifier (for non-review detection). The output from this stage is

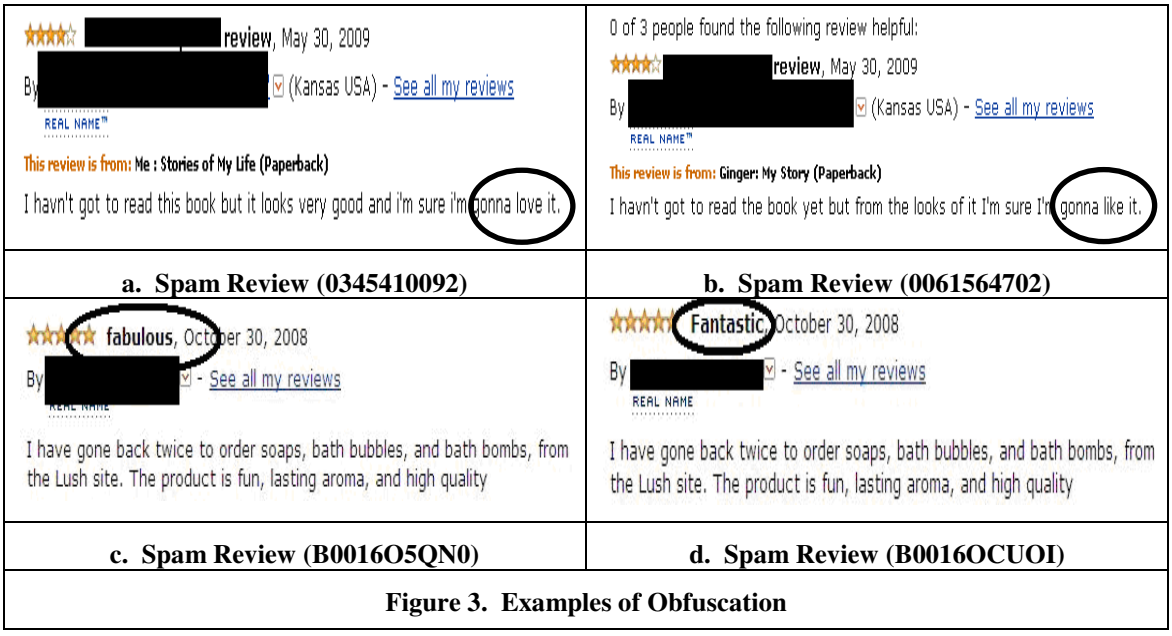
[4] http://ecs.amazonaws.com/onca/xml?Service=AWSECommerceService
[5] http://api.cnet.com/
[6] http://developer.shopping.com/docs/API_Use_Cases
[7] http://code.google.com/apis/ajaxsearch/
[8] http://wordnet.princeton.edu/

the collection of detected fake reviews. To facilitate business managers or marketers to develop effective product design strategies or marketing campaigns, the contents of the detected fake reviews (e.g., specific product features and their sentiments) will be analyzed by the spam content analysis component (Task 7 in Figure 2). The spam analysis component needs to consult some sentiment lexicons and the archived product descriptions to identify the spam related to specific product features and sentiments. In particular, this component is developed based on an ontology-based (Lau et al. 2009a) sentiment analysis technique which has been successfully applied to opinion mining before (Lau et al. 2009b). Because of limited space, we will only focus on the spam detection module, in particular, the untruthful review detection module in this paper. Finally, the detected fake reviews and the results of the feature driven spam analysis will be presented to the users via the presentation manager (Task 8 in Figure 2). Alternatively, users can have a quick browse of the detected spam without invoking the spam content analysis component. Our prototype system was developed using Java (J2SE v 1.4.2), Java Server Pages (JSP) 2.1, and Servlet 2.5. The system is hosted on a DELL 1950 III Server with Quad-Core Xeon 2.33GHz Processors, 16GB main memory, and 6TB secondary storage.

A Computational Model for the Detection of Untruthful Reviews

The Intuitions of the Untruthful Review Detection Method

In this paper, untruthful reviews loosely refer to spammers' false comments (opinions) about some products or services (Jindal and Liu 2008). To directly measure the concept of "untruthfulness" is not practical because computers cannot read a reviewer's mind. Alternatively, we propose an approximation method which indirectly estimates "untruthfulness" based on the "similarity" of semantic content among reviews. If the semantic contents of two reviews are very similar, there is a good reason to believe that the content of one review is just copied from another one, and so that review does not sincerely reflect the reviewer's true opinions. Borrowing the ideas from existing Web spam research which can be broadly classified as content-based or link-based approaches (Gyöngyi and Garcia-Molina 2005), we develop a novel computational method to detect untruthful reviews based on the content of a review and its "conceptual" similarity links to other reviews. Unlike ordinary Web pages, direct hyperlinks among consumer reviews are uncommon (except some commercial spam). We extend the well-known language modeling (LM) framework (Lafferty and Zhai 2001; Liu and Croft 2004; Ponte and Croft 1998) to develop a novel semantic-based smoothing method to estimate the likelihood of semantic content generation between two reviews. Then, Kullback-Leibler (KL) divergence (Kullback and Leible 1951) is applied to measure the distance of these probability distributions to estimate the conceptual similarity between the reviews.



a. Spam Review (0345410092) | **b. Spam Review (0061564702)**

c. Spam Review (B0016O5QN0) | **d. Spam Review (B0016OCUOI)**

Figure 3. Examples of Obfuscation

It has been pointed out that spammers tend to adopt an obfuscation strategy to deliberately modify online comments (Abbasi et al. 2008). Our proposed semantic language modeling approach is different from the traditional

plagiarized content detection method (David and Pinch 2006) in that "substituted" terms can be taken into account when the similarity of review contents is estimated. Figure 3 depicts two cases of obfuscation. Figures 3a and 3b show that the spammer deliberately modifies some of the words in the fake reviews. Figures 3c and 3d show that the spammer deliberately modifies the titles of the fake reviews. To effectively detect these fake reviews, the proposed semantic language models should take into account the relationship ("love"→"like") when semantic content similarity is estimated between these two reviews. In fact, the term "like" is a synonym of the term "love" according to WordNet (Miller et al. 1990). For the term association relationships like ("fabulous"→"fantastic"), they are dynamically discovered via the contest-sensitive text mining method which has been successfully applied to information retrieval tasks before (Lau 2003; Lau et al. 2008). The computational details of the proposed semantic language model for untruthful review detection will be discussed in the following sub-section.

The Semantic Language Models

We propose a novel semantic language model to estimate the similarity between any pairs of reviews in terms of the likelihood of a review "generating" the semantic content of another review. The term "language model" is widely used by the speech recognition community to refer to a probability distribution *M* which represents the statistical regularities for the generation of the language (Nadas 1984). In other words, a language model is a probabilistic function that assigns a probability to a string *t* drawn from some vocabulary *T*. Language modeling has been applied to estimate the relevance of a document *d* with respect to a query *q* in the field of information retrieval (IR) (Liu and Croft 2004; Ponte and Croft 1998). Moreover, language modeling approaches have been successfully applied to Web spam detection (Martinez-Romo and Araujo 2009), blog spam detection (Mishne et al. 2005), and opinion mining respectively (Lau et al. 2009c). However, the aforementioned language modeling approaches did not take into account the term relationships when the document generation probability was estimated. The basic unigram language model is defined by (Liu and Croft 2004; Ponte and Croft 1998):

$$P(q \mid d) \propto P(q \mid M_d) = \prod_{t_i \in q} P(t_i \mid M_d) \tag{1}$$

$$P(t_i \mid M_d) = (1-\lambda)\left((1-\gamma)P_{ML}(t_i \mid M_d) + \gamma P_{SEM}(t_i \mid M_d)\right) + \lambda P_{ML}(t_i \mid M_D) \tag{2}$$

$$P_{ML}(t_i \mid M_d) = \frac{tf(t_i, d)}{|d|} \tag{3}$$

In Eq.1, the term $P(q \mid d)$ represents the likelihood of document *d* is relevant with respect to the query *q*, and this likelihood is approximated by the probability that the document language model M_d "generate" the query *q*, that is, $P(q \mid M_d)$. This generation probability turns out to be the product of the probability of M_d generating the individual term t_i of the query *q*, that is, $P(t_i \mid M_d)$. One important element of the language modeling approach is the "smoothing" of term probability (Zhai and Lafferty 2004). The main intuition is that if a query term t_i is not found in the document *d*, it may not necessarily mean that the document is not about t_i because semantically similar terms such as synonyms could be used to compose the document. Accordingly, the objective of smoothing a document model is not to over estimating the generation probability for terms observed in the document by applying a factor $(1-\lambda)$ to the maximum likelihood language model $P_{ML}(t_i \mid M_d)$ and the semantic language model $P_{SEM}(t_i \mid M_d)$ defined in Eq.2.

In Eq.2, the generation probability of an observed query term t_i is estimated according to the maximum likelihood document language model defined in Eq.3, where $tf(t_i, d)$ is the occurrence frequency of t_i in *d*, and $|d|$ is the document length (i.e., the number of terms in *d*). On the other hand, for the unobserved terms, the smoothing process tries to adjust (i.e., increase) their generation probability by applying the factor $\lambda P_{ML}(t_i \mid M_D)$. The term λ is called the Jelinek-Mercer smoothing parameter which usually assumes values in the range of [0.1, 0.7] (Nie et al. 2006; Zhai and Lafferty 2004). For Jelinek-Mercer smoothing, $P_{ML}(t_i \mid M_D) = \frac{tf(t, D)}{|D|}$ is usually applied, and it

means that the probability of an unobserved term *t* is estimated according to its document frequency in the entire document collection *D*. In particular, $tf(t, D)$ represents the document frequency of *t*, and $|D|$ is the number of documents of the collection *D*. The term γ is a second Jelinek-Mercer parameter applied to smooth the generation probabilities between $P_{ML}(t_i \mid M_d)$ and $P_{SEM}(t_i \mid M_d)$. Our novel semantic language model is defined by:

$$P_{SEM}(t_i \mid M_d) = \frac{\sum_{t_i, t_j \in R} P(t_i \mid t_j) P_{ML}(t_j \mid M_d)}{|R|} = \frac{\sum_{t_i, t_j \in R} P(t_j \rightarrow t_i) P_{ML}(t_j \mid M_d)}{|R|} \quad (4)$$

where $P_{SEM}(t_i \mid M_d)$ is the proposed semantic language model, and $P(t_j \rightarrow t_i)$ is the certainty of the term association between t_i and t_j. The basic intuition of Eq.4 is that if a query term such as "like" is not found in the document, but the term "love" is found in the document and the term association such as "love" → "like" is established (according to WordNet or context-sensitive text mining), the generation probability of $P_{ML}(\text{"love"} \mid M_d)$ can be used to estimate $P_{ML}(\text{"like"} \mid M_d)$. In Eq.4, the term *R* represents the set of term relationships in the form of $t_j \rightarrow t_i$ and $|R|$ is the cardinality of the set *R*. Since there may be quite a number of term associations discovered via context-sensitive text mining, only the top five associations ranked by $P(t_i \mid t_j)$ for each term t_i is considered. For the synonym relations extracted from WordNet, $P(t_i \mid t_j) = 1.0$ is assumed because they are defined by human experts.

For review spam detection, a pair of reviews denoted d_1 and d_2 will be compared each time. Essentially, one of the reviews is seen as a long query. For the proposed computational method, the longer review is always denoted as d_1. If two reviews have the same length, either one can be taken as d_1. KL divergence (Kullback and Leibler 1951) is a well-known measure commonly used to estimate the distance between two probability distributions, and it has been successfully applied to Web spam detection before (Martinez-Romo and Araujo 2009; Mishne et al. 2005). Accordingly, we apply KL divergence to measure the distance between the pair of language models such as M_{d_1} and M_{d_2}. If the KL divergence value of the two language models is very small, it suggests that the semantic contents of the pair of reviews are quite similar, and they are likely to be spam. The KL divergence measure can also be seen as a kind of normalization applied to the review generation probabilities derived by our semantic language models. The final equation for the untruthful review detection method which is underpinned by LM and KL divergence is defined by:

$$KL(M_{d_1} \,||\, M_{d_2}) = \sum_{t_i \in \{d_1 \cup d_2\}} P(t_i \mid M_{d_1}) \times \log_2 \frac{P(t_i \mid M_{d_1})}{P(t_i \mid M_{d_2})} \quad (5)$$

where t_i is a term appears in d_1 or d_2. Only one KL computation is required for our approach and d_1 is assumed to be the longer review for each pair.

Design Evaluation

Evaluation Procedures

In order to evaluate the design of the proposed review spam detection methodology, an instantiation (i.e., a prototype system) was developed. Basically, we adopted an evaluation approach similar to that used by the well-known spam detection benchmarking forum, the TREC Spam Track (Cormack and Lynam. 2005; Cormack 2007).

In particular, our review spam detection system automatically scanned through a collection of consumer reviews to identify the spam (e.g., untruthful reviews and non-reviews). Our system's classification results would then be compared with the gold standard (i.e., the spam and the legitimate reviews confirmed by human experts). Standard effectiveness measures commonly used by the Web spam research community would then be applied to assess the effectiveness of our instantiation, and hence the effectiveness of the proposed review spam detection methodology. As the TREC Spam Track evaluation dataset comprised email messages only, we built our own evaluation dataset based on the reviews downloaded from amazon.com. We utilized the Amazon Web services to extract 1,484,485 reviews from six product categories during January 2010. A subset from this collection was then used to build our evaluate dataset. As it would be extremely labor intensive if we asked our human annotators to inspect all the downloaded reviews to identify the spam and the ham (i.e., legitimate) reviews for each product category, a semi-automatic method was applied to build the evaluation set. Similar to the method adopted by Jindal and Liu (2008), we used the Jaccard ratio to identify pairs of suspicious spam reviews for human inspection. If the Jaccard ratio of a pair of reviews was greater than or equal to a pre-defined threshold value (e.g., 0.7), these reviews would be added to a candidate spam set. Two human annotators would then inspect the candidate spam set. If both of them confirmed a spam case, it would be included in our evaluation dataset as a spam. If there is a disagreement between the annotators, the potential spam review would not be added to our evaluation dataset. Similarly, the Jaccard ratio was applied to select ham reviews to build our evaluation dataset. For instance, if the maximal Jaccard ratio between a review and all the other reviews of a product category was below a pre-defined threshold (e.g., 0.1), it would be included in our evaluation dataset as a ham. Similar to Jindal and Liu's (2008) approach, the ham reviews were not inspected by our human annotators because of its large volume. The spam-ham ratio was set to 2.3% to simulate a realistic and highly skewed spam distribution according to a previous empirical study (Jindal and Liu 2008). The details of our evaluation dataset are depicted in Table 1. The reviews from the product category of "Music" were used to calibrate a KL divergence threshold and empirically establish the Jelinek-Mercer smoothing parameters; these parameters would then be applied to spam detection for the other product categories.

Table 1. The Characteristic of the Amazon Evaluation Dataset

Product Category	Amazon Browse Node ID	No. of Untruthful Reviews	No. of Legitimate Reviews	Total No. of Reviews	Spam-Ham Ratio
Grocery	3760931	180	8,000	8,180	2.25%
Electronics (Digital Cameras)	281052	200	8,800	9,000	2.27%
PC Hardware	541966	350	15,400	15,750	2.27%
Books (Entertainment and History)	86 and 9	300	13,200	13,500	2.27%
DVD (Animation)	163416	280	12,400	12,680	2.26%
Music (Jazz)	34	200	8,800	9,000	2.27%
Total		1,510	66,600	68,110	

Two versions of the proposed LM and KL based untruthful review detection method were implemented and evaluated. The first experimenting system (KL) was implemented according to Eq.1 and Eq.5 only. In other words, no semantic smoothing was applied. The second experimenting system (KLSS) was developed according to Eq.1 to Eq.5, and it was underpinned by our novel semantic language model which could take into account term substitutions in fake reviews due to spammers' obfuscation actions. On the other hand, a baseline review spam detection system was also developed according to a supervised classification model proposed by Jindal and Liu (2008). This baseline system (LR) was underpinned by the logistic regression model and made use of three categories of features such as content-based features, reviewer-based features, and product-based features to predict if a review is an untruthful review or not. There were 21 content-based features including features such as length of review title, length of review, percentage of helpful feedback, percentages of positive or negative sentiment indicators, and so on. For counting the percentages of positive or negative sentiment indicators, we utilized the OpinionFinder sentiment lexicon (Wilson et al. 2005). In addition, there were 11 reviewer-based features like

whether the reviewer had only written one review, the average rating of the reviewer, and so on. Finally, there were four product-based features such as price of the reviewed product, sales rank of the reviewed product, the average rating and the standard deviation of the ratings of the reviewed product. In addition, we also applied the vector space model (Salton et al. 1975) to build another baseline system (VS). In particular, the Term Frequency Inverse Document Frequency (TFIDF) term weighting scheme (Salton and McGill 1983) was used to construct weighted vectors to represent consumer reviews, and the cosine similarity measure (Salton and McGill 1983) was applied to compare the similarity of reviews. For each product category, if the cosine score of a pair of review was greater than a threshold value (e.g., 0.8), they would be classified as untruthful reviews. A third baseline system (JR) was implemented based on the Jaccard ratio. For this baseline system, a bag of words approach was used to represent a review. The Jaccard ratio was then computed for each pair of reviews. If the Jaccard ratio was greater than a pre-defined threshold, the pair of reviews would be considered as spam. Basically, both the VS and the JR baseline systems adopted an un-supervised classification approach. All the implemented systems used the same threshold calibrate strategy.

The Performance Measures

We employed the evaluation measures adopted in the TREC Spam Track (Cormack and Lynam. 2005; Cormack 2007) to evaluate the performance of the various review spam detection methods. Although these measures were originally used to examine the effectiveness of email spam filters in the TREC Spam Track, they were also widely used to evaluate other kinds of Web spam (Martinez-Romo and Araujo 2009). With reference to the confusion matrix depicted in Table 2, the various effectiveness measures can be defined by:

Table 2. A Confusion Matrix for the Definition of the Effectiveness Measures

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

$$hm = \frac{b}{b+d} \tag{6}$$

$$sm = \frac{c}{a+c} \tag{7}$$

$$lam = \text{logit}^{-1}\left(\frac{\text{logit}(hm)+\text{logit}(sm)}{2}\right) \tag{8}$$

$$tp = \frac{a}{a+c} \tag{9}$$

where *a, b, c*, and *d* refer to the number of reviews falling into each category. The ham misclassification rate (*hm*) is the fraction of all ham misclassified as spam; the spam misclassification rate (*sm*) is the fraction of all spam misclassified as ham. As there is a natural tension between ham and spam misclassification rate, a spam detection system (i.e., a classifier) can always improve the *hm* rate at the expense of the *sm* (e.g., by increasing the spam classification threshold *t*) or vice versa. It is desirable to have a single measure which combines both of the above measures. Therefore, the TREC Spam track also made use of the logistic average misclassification rate (*lam*) to measure the effectiveness of spam detection systems, where $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$ and $\text{logit}(x) = \ln\left(\frac{x}{1-x}\right)$. Since *hm*, *sm*, and *lam* are the measures for failure rather than effectiveness, the lower scores imply a better detection performance. The true positive rate (*tp*) is the fraction all spam identified by the system. On the other hand, the common effectiveness measure $\text{accuracy} = \frac{a+d}{a+b+c+d}$ may not be a good measure for spam detection applications.

Given a skewed distribution of spam-ham reviews (e.g., a large number ham reviews and only a small number of spam reviews), a spam detection system can simply classify all reviews as ham (i.e., category *d*) and trivially achieves a relative high accuracy score. Nevertheless, to make it easier to compare with the results of some earlier studies which also utilized the accuracy measure, we also report the accuracy figures in this paper.

Receiver Operating Characteristic (ROC) curve (Hand and Till 2001), which is the graphical representation of the *tp* (i.e., 1 - *sm*) as a function of *hm*, has also been applied to evaluate spam detection systems (Cormack 2007; Cormack et al. 2007b). The advantage of the ROC curve is that the evaluation (or comparison) of spam detection systems will not depend on the choice of a particular threshold value (Hand and Till 2001). The top left corner of a ROC plot represents a good system classification performance. The area under the ROC curve (AUC) can be interpreted as the probability that a random ham will receive a lower spamminess score than a random spam (Cormack 2007; Cormack et al. 2007b). AUC is equivalent to the Mann-Whitney-Wilcoxon signed rank test (Hand and Till 2001). To be consistent with the *hm* and the *sm* rates which measure failure rather than effectiveness, the TREC spam track also employed the measure "Area Above the ROC Curve", that is, $(1 - AUC)$, to evaluate spam filters.

Experimental Results and Discussions

We applied the experimenting systems and the baseline systems to the evaluation dataset which contained reviews extracted from several product categories of Amazon. The detailed performance data of all the systems is depicted in Table 3. The KLSS system achieved the lowest error rate in terms of lam% (6.1%) and (1-AUC)% (3.8%); it also produced the highest accuracy (94.5%) and true positive rate (93.2%). According to the ROC curve analysis shown in Figure 4, the KLSS system consistently performs better than other baseline systems at all possible threshold levels (e.g., its ROC curve is above all the other curves).

Table 3. Comparative Performance of the Untruthful Review Detection Methods

Method	tp%	hm%	sm%	lam%	(1-AUC)%	Accuracy
KLSS	93.2%	5.5%	6.8%	6.1%	3.8%	94.5%
KL	91.1%	8.5%	8.9%	8.7%	5.5%	91.5%
JR	90.9%	9.3%	9.1%	9.2%	5.6%	90.7%
VS	89.5%	12.9%	10.5%	11.6%	9.1%	87.2%
LR	35.1%	27.5%	64.9%	45.6%	48.3%	71.7%

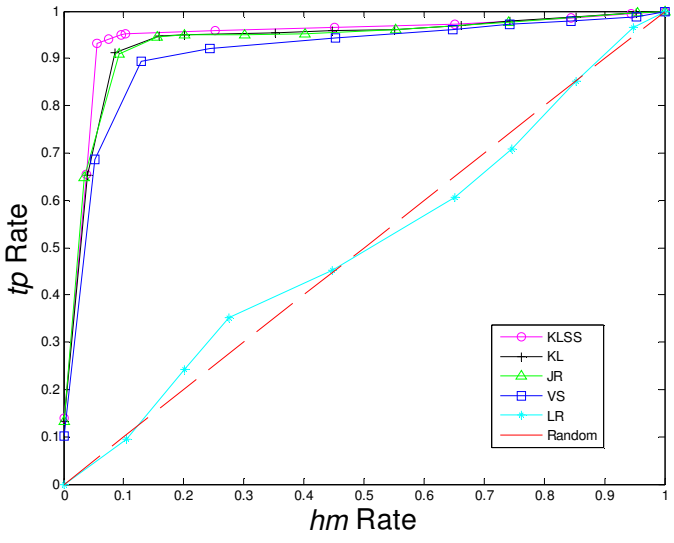


Figure 4. The ROC Curves of Various Spam Detection Methods

The KLSS system achieved the best performance because it could take into account term substitutions in fake reviews; term substitutions occurred when spammers exercised the obfuscation strategy (Abbasi et al. 2008). The VS system is not as effective as the proposed language modeling based detection systems such as KL and KLSS even though all of them are based on un-supervised classification approach. After our in-depth analysis of the detection results, we found that the VS system performed very poorly for short reviews. Table 4 shows one of the typical examples that the VS system mistakenly assign a relatively high similarity (spamminess) score (e.g., 0.902) to two semantically different reviews (i.e., ham misclassification) in the product category "PC Hardware". The first row of Table 4 shows our internal XML encoding of the contents of the reviews including their respective ASINs from Amazon. The <Content> tag is used to encode the main content of a review, and the <Summary> tag is employed to delimit the heading of a review. The second row of Table 4 depicts the normalized TFIDF vectors of the corresponding reviews. As can be observed, the semantic contents of these reviews are quite different. However, as one of the overlapping terms "quickbook" (highlighted) has a relatively high TFIDF weight in the respective review vectors, the resulting cosine score becomes quite high. The reason of a high TFIDF weight for the term "quickbook" is that it is a rare term in the review collection, and this term appears twice in both reviews (one occurrence in the main content and another occurrence in the review heading). In contrast, the SLM detection method will not be confused by the relatively high term weight of a rare term because term probability distributions in reviews rather than term weights are utilized to estimate the degree of semantic content similarity between two reviews.

Table 4. The Problem of the Vector Space Model for Untruthful Review Detection

<Review> <ASIN>B000RZTDZ6</ASIN> <DocID>B000RZTDZ6-046</DocID> <Summary> quickbooks </Summary> <Content> This product seems ok,,, but I couldn't use is with my laptop when needed to work on quickbooks. Don't understand the exact reason. </Content> </Review>	<Review> <ASIN>B000RZTDZ6</ASIN> <DocID>B000TKHBDK-506</DocID> <Summary> Not compatible with quickbooks </Summary> <Content> I really wanted this product to help me with quickbooks on my laptop but it wasn't compatible... I wish there was one that I could use. It would help me work faster when I needed to do data entry. </Content> </Review>
Normalized TFIDF vector for B000RZTDZ6-046: <**quickbook** 0.98602 exact 0.08341 **product** 0.08150 reason 0.07138 understand 0.06753 **work** 0.04707 **laptop** 0.04527 **thi** 0.01571>	Normalized TFIDF vector for B000TKHBDK-506: <**quickbook** 0.90647 compat 0.21545 entri 0.20541 data 0.19632 faster 0.19136 realli 0.0727 **product** 0.06318 wa 0.05049 **work** 0.03649 **laptop** 0.03508 **thi** 0.01218>

The effectiveness of the JR system is similar to that of the KL system (i.e., language modeling without semantic smoothing) because both systems estimate content similarity based on the set of overlapping terms. Moreover, the JR system performs better than the VS system because it will not be confused by the extraordinary high TFIDF weights of the rare terms. On the other hand, the LR system performed poorly, and it was not much better than a random guest according to its ROC curve depicted in Figure 4. The reason for such a poor performance is that there is not prominent feature which can clearly distinguish the untruthful reviews from the legitimate reviews.

Application of the Designed Artifacts

An Empirical Study of Online Review Spam at a Typical e-Commerce site

Given the sheer volume of online consumer reviews, our designed artifacts enable business managers or individual consumers to efficiently detect and analyze fake reviews presented at e-Commerce Web sites. Such an analysis can help marketers or business managers develop a real picture of the consumers' opinions or preferences about their

products and services. Since amazon.com is the largest e-Commerce Web site, we conduct our empirical analysis of review spam based on the random sample of 1,484,485 online consumer reviews downloaded from amazon.com during January 2010. Table 5 summarizes the details of our spam detection results. The percentages of the adjusted spam detection figures range from 0.99% to 2.99%. The average spam percentage detected is 2.08% which is slightly higher than that reported in a previous study (Jindal and Liu 2008). This could be due to the improvement on detection effectiveness by using our novel LM and KL based detection methodology when compared with the LR-based supervised classification method used in the previous study (Jindal and Liu 2008). Apparently, untruthful reviews are the main source of spam; there are only a handful of non-reviews detected by our SVM-based detection component in each product category. The reason of this phenomenon may be that the moderation procedure adopted at amazon.com can filter most of the non-reviews based on the explicit features demonstrated in this kind of reviews. However, it is very difficult to detect untruthful reviews based on the current state-of-the-art adopted at amazon.com. On average, a marketer or consumer will only refer to two fake reviews out of one hundred references to online reviews. Although the review spam rate is not very high when compared to the average of 10-15% spam on the Web (Gyöngyi and Garcia-Molina 2005), marketers or consumers should be cautious if they want to make business or purchase decisions solely based on the online consumer reviews. Whether the 2.08% spam reviews will cause a serious consequence to merchants or individual consumers can be evaluated based on an econometric analysis presented in the following sub-section.

Table 5. Review Spam Detected Based on the Amazon Sample

Product Category	No. of Products	No. of Products with Spam	No. of Reviews	No. of Untruthful Reviews	No. of Non-reviews	% of Spam
Grocery	16,008	1,484	122,474	3,604	56	2.99%
Electronics (Digital Cameras)	2,318	601	118,892	1,114	64	0.99%
PC Hardware	16,866	1,931	380,733	10,795	114	2.87%
Books (Entertainment and History)	33,328	2,784	481,291	7,325	85	1.54%
DVD (Animation)	3,115	1,121	151,941	3,263	97	2.21%
Music (Jazz)	26,529	1,550	229,154	4,228	88	1.88%
Average/Cat	16,361	1,579	247,414	5,055	84	2.08%

Econometric Analysis

Our designed artifacts make it feasible to conduct an econometric analysis to assess the impact of review spam on product sales. According to the previously established linear relationship between sales rank and actual sales of products (Chevalier and Mayzlin 2006), it is possible to estimate the economic impact of spam on the actual sales of products. In general, a linear regression model between a product's sales rank and its product related attributes can be described by (Archak et al. 2007; Chevalier and Mayzlin 2006; Ghose and Ipeirotis 2007):

$$\begin{aligned}\ln(SalesRank_p) = {} & \nu_p + \mu_p + \beta_p^1 \cdot \ln(Price_p) + \\ & \beta_p^2 \cdot Shiptime_p + \\ & \beta_p^3 \cdot Rating_p + \\ & \beta_p^4 \cdot \ln(Nreviews_p) + \\ & \beta_p^5 \cdot \ln(Lreviews_p) + \\ & \varepsilon_p\end{aligned} \tag{10}$$

where v_p is a fixed effect of product p and it may be related to factors such as the quality of the product, the brand loyalty of the product, the popularity of the manufacturer, etc. On the other hand, μ_p is the Website-product effect for product p and it may be related to the fit between the product and the preferences of the consumers transacting on the Website (Chevalier and Mayzlin 2006). Moreover, the variable ε_p represents a random disturbance factor which is assumed to be normally distributed (Archak et al. 2007). The dummy variable $Shiptime_p$ encodes the estimated shipment time of the product by the Web site. For instance, the estimated shipment time by Amazon can be expressed in terms of hours, days, weeks, months, or no estimated delivery period. We then encoded the shipment dummy by the values of 1, 2, 3, 4, and 5 respectively. The variable $Rating_p$ represents the average review rating of the product. The variables $Nreview_p$ and $Lreview_p$ represent the number of reviews and the length of reviews (in characters) pertaining to the product. If the sales rank of a product is not available at amazon.com, that product was not included in our regression analysis.

Table 6. Impact Analysis

Variable	Grocery		Electronics		PC Hardware		Books		DVD		Music	
	ALL	SF	ALL	SF	ALL	SF	ALL	SF	ALL	SF	ALL	SF
ln(Price)	-0.132** (.033)	-0.134** (.033)	0.086*** (.023)	0.087*** (.023)	-0.067** (.032)	-0.063** (.032)	-0.143** (.060)	-0.160*** (.060)	0.324*** (.048)	0.320*** (.048)	0.276*** (.059)	0.271*** (.058)
Shiptime	0.183*** (.024)	0.183*** (.024)	0.343*** (.027)	0.343*** (.027)	0.187*** (.023)	0.193*** (.023)	0.141*** (.009)	0.143*** (.009)	0.199*** (.011)	0.198*** (.011)	0.165*** (.008)	0.161*** (.008)
Rating	-0.211** (.056)	-0.218** (.059)	-0.111 (.101)	-0.198* (.102)	-0.193** (.093)	-0.162* (.096)	-0.227*** (.053)	-0.135*** (.043)	-0.359*** (.061)	-0.318*** (.051)	-0.225*** (.061)	-0.243*** (.050)
ln(Nreviews)	-0.343** (.057)	-0.341** (.056)	-0.389*** (.047)	-0.386*** (.047)	-0.307*** (.052)	-0.243*** (.052)	-0.477*** (.024)	-0.458*** (.024)	-0.545*** (.023)	-0.501*** (.023)	-0.604*** (.025)	-0.545*** (.023)
ln(Lreviews)	0.133 (.069)	0.124 (.052)	0.233 (.142)	0.123 (.080)	-0.590*** (.139)	-0.059 (.070)	0.075 (.065)	0.159*** (.037)	0.128** (.060)	0.190*** (.039)	0.181*** (.062)	0.142*** (.038)
R^2	0.355	0.358	0.309	0.315	0.190	0.184	0.377	0.373	0.369	0.372	0.458	0.459
No. of Observations	617		500		594		2,774		1,099		1,545	

Notes: * indicating $p < .10$, ** indicating $p < .05$, *** indicating $p < .01$; ALL indicating all the reviews were included; SF (Spam Free) indicating only ham reviews were included; standard errors are in parentheses.

Table 6 depicts the results of our linear regression analysis when all the reviews, or when only the ham reviews are included. As a whole, our results are consistent with that of the previous regression analysis in that both the volume and the valence of reviews are significant predictors for product sales (via sales ranks) in different product categories (Chevalier and Mayzlin 2006; Dellarocas et al. 2007; Godes and Mayzlin 2004; Liu 2006). Both of these factors are negatively correlated to sales rank. In other words, the higher average rating of a product is, the lower its rank number will be (i.e., ranked close to the top position). Sales prices and shipment time are also found significant factors affecting product sales. On the other hand, review lengths seem not a significant factor to predict sales. It is obvious that the presence of review spam does influence the predictive power of the independent variables because both the regression coefficients and the R^2 are changed after the removal of spam.

To further assess the impact of fake reviews on product sales, we extend our regression model by incorporating the spam review factors. As the number of positive spam $NposSpam_p$ and the number of negative spam $NnegSpam_p$ may affect consumers' perception about a product and eventually its sales, these factors are included in our extended regression model. A review with ratings of 1 and 2 is treated as negative, and a review with ratings of 4 and 5 is treated as positive in this analysis. Similarly, $NposHam_p$ and $NnegHam_p$ represent the number of positive ham reviews and the number of negative ham reviews. The factors $HamRating_p$ and $SpamRating_p$ represent the average

rating of ham reviews and the average rating of spam reviews respectively. The results of our regression analysis for the Amazon Book category are shown in Table 7. As the linear relationship between the sales rank and the actual sales of books, that is, $\ln(sales) = 9.825 - .78\ln(rank)$ has been established (Chevalier and Mayzlin 2006), we focus on the analysis of the book data only.

$$\begin{aligned}\ln(SalesRank_p) = {} & \nu_p + \mu_p + \beta_p^1 \cdot \ln(Price_p) + \\ & \beta_p^2 \cdot Shiptime_p + \\ & \beta_p^3 \cdot HamRating_p + \\ & \beta_p^4 \cdot \ln(NposHam_p) + \beta_p^5 \cdot \ln(NnegHam_p) + \\ & \beta_p^6 \cdot SpamRating_p + \\ & \beta_p^7 \cdot \ln(NposSpam_p) + \beta_p^8 \cdot \ln(NnegSpam_p) + \\ & \varepsilon_p\end{aligned} \tag{11}$$

Table 7. Cross-Sectional Econometric Analysis

ln(Price)	Shiptime	HamRating	ln(NposHam)	ln(NnegHam)
0.133*** (0.039)	0.178*** (0.006)	-0.127** (0.026)	-0.584*** (0.024)	0.072*** (0.028)
SpamRating	**ln(NposSpam)**	**ln(NnegSpam)**	**R^2**	**# of Observations**
-0.148*** (0.040)	0.152*** (0.040)	0.269*** (0.097)	0.404	2,774

Notes: ** indicating $p < .05$, *** indicating $p < .01$; standard errors are in parentheses.

As the coefficient of the variable $NposSpam$ has a positive sign, it seems that promotional spam may not able to increase product sales. On the contrary, spreading negative spam may cause damage to a firm's product sales. Suppose that a book currently ranked at 100 were attacked by one negative spam, its sales rank would increase by 0.269 unit (the magnitude of the coefficient of $NnegSpam$) given that the other factors remain the same. In other words, its sales rank would rise to 100.269. This change of sales rank implies that the actual sales of the book may be reduced. According to the linear relationship between sales and sales rank, the previous sales volume and the new sales volume can be estimated by: $\ln(sales_{prev}) = 9.825 - .78\ln(100) = 6.233$ and $\ln(sales_{new}) = 9.825 - .78\ln(100.269) = 6.231$. As a result the actual sales of the book would drop by 1.02 units per week (i.e. 509.28 – 508.26). So, if a firm's product were attacked by a negative spam, the economic impact of this spam would be the reduction of sales by one unit per week, or 4 units per month. According to our dataset of Amazon book, the average price of a book is $19.1. As a result, the economic loss of a firm caused by the attack of each negative spam review is $76.4 per month!

Conclusions and Future Work

Few empirical studies have been conducted to examine the trustworthiness of online consumer reviews because of the lack of an effective and efficient methodology to automatically scan through a sheer volume of online consumer reviews. Driven by the design science research methodology, one of the main contributions of our research work is the development of a novel methodology to combat online review spam. Through the development of an instantiation, the proposed design has been evaluated based on a TREC like evaluation procedure. Our experimental results confirm that the semantic LM and KL divergence based computational model is effective for the detection of untruthful reviews; our proposed computational model outperforms other well-known baseline models in the Amazon review dataset. Empowered by the designed artifacts, our empirical study found that around 2% of the online consumer reviews are spam. An econometric analysis has also been performed to assess the impact of fake

reviews on product sales. To the best of our knowledge, this is the first empirical study to examine the economic impact of fake consumer reviews on product sales. Based on the book data collected from Amazon, it is found that a firm may lose as much as $76.4 per month because of the attack by each negative spam review. The managerial implication of our research work is that business managers or marketers can apply our proposed methodology to identify and analyze fake reviews related to their products and services. Accordingly, they can develop more effective product design strategies and marketing plans based on genuine consumer feedback. In addition, our designed artifacts can help individual consumers assess the true quality of products and facilitate their comparison shopping processes. Future work involves the evaluation of both the effectiveness and the efficiency of the proposed methodology based on a larger dataset (e.g., the entire review collections from different e-Commerce Web sites). Moreover, more sophisticated language modeling approaches such as n-gram language models will be examined to improve the effectiveness of the review spam detection method. A larger scale of econometric analysis will be conducted to assess the impact of fake reviews on different kinds of products and services transacted over the Internet.

References

Abbasi, A., Chen, H., and Nunamaker Jr., J. F. 2008. "Stylometric Identification in Electronic Markets: Scalability and Robustness," *Journal of Management Information Systems* (25:1), pp. 49-78.

Arazy, O., and Woo, C. 2007. "Enhancing Information Retrieval Through Statistical Natural Language Processing: A Study of Collocation Indexing," *MIS Quarterly* (31:3), pp. 525-546.

Archak, N., Ghose, A., and Ipeirotis, P. 2007. "Show me the money!: deriving the pricing power of product features by mining consumer reviews," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, California, pp. 56 - 65.

Chang, M.W., Yih, W.T., and Meek, C. 2008. "Partitioned logistic regression for spam filtering," in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 97-105.

Chen, F., Tan, P., Jain, A. 2009. "A co-classification framework for detecting web spam and spammers in social media web sites," in *Proceeding of the 18th ACM conference on Information and knowledge management*, pp. 1807-1810.

Chen, Y., Xie, J. 2008. "Online Consumer Review: Word-of-Mouth as a New Element of Marketing Communication Mix," *Management Science* (54:3), pp. 477-491.

Cheung, M. Y., Luo, C., Sia, C.L., Chen, H. 2009. "Credibility of Electronic Word-of-Mouth: Informational and Normative Determinants of On-line Consumer Recommendations," *International Journal of Electronic Commerce* (13:4), pp. 9-38.

Chevalier, J. A., Mayzlin, D. 2006. . "The Effect of Word of Mouth on Sales: Online Book Reviews," *Journal of Marketing Research* (43:3), pp. 345-354.

Cormack, G. V., and Lynam, T.R. 2005. "TREC 2005 Spam Track Overview," Available at: http://plg.uwaterloo.ca/~gvcormac/trecspamtrack05.

Cormack, G. V. 2007. "TREC 2007 Spam Track Overview," Available at: http://trec.nist.gov/pubs/trec16/papers/SPAM.OVERVIEW16.pdf.

Cormack, G. V., Hidalgo, J., Sánz, E. 2007a. "Spam filtering for short messages," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 313-319.

Cormack, G. V., Hidalgo, J., Sánz, E. 2007b. "Online supervised spam filter evaluation," *ACM Transactions on Information Systems* (25:3), pp. 11.1-11.31.

Danescu-Niculescu-Mizil, Kossinets, C., Kleinberg, J., Lee, L. 2009. "How opinions are received by online communities: a case study on amazon.com helpfulness votes," in *Proceedings of the 18th international conference on World Wide Web*, pp. 141-150.

David, S. and Pinch, T. 2006. "Six degrees of reputation: The use and abuse of online review and recommendation systems," *First Monday*, Special Issue #6: Commercial Applications of the Internet. Available online at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/issue/view/221.

Dellarocas, C. 2003. "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," *Management Science* (49:10), pp. 1407-1424.

Dellarocas, C. 2006. "Strategic Manipulation of Internet Opinion Forums: Implications for Consumers and Firms," *Management Science* (52:10), pp. 1577-1593.

Dellarocas, C., Zhang, X., Awad, N. 2007. "Exploring the Value of Online Product Reviews in Forecasting Sales: The Case of Motion Pictures," *Journal of Interactive Marketing* (21:4), pp. 23-45.

Fetterly, D., Manasse, M., Najork, M. 2005. "Detecting phrase-level duplication on the world wide web," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 170-177.

Gefen, D., Benbasat, I., and Pavlou, P. A. 2008. "A Research Agenda for Trust in Online Environments," *Journal of Management Information Systems* (24:4), pp. 275-286.

Ghose, A., Ipeirotis, P. G. 2007. « Designing novel review ranking systems: predicting the usefulness and impact of reviews," in *Proceedings of the ninth international conference on electronic commerce*, pp. 303-309.

Godes, D. and Mayzlin, D. 2004. "Using Online Conversations to Study Word-of-Mouth Communication," *Marketing Science* (23:4), pp. 545-60.

Gyöngyi, A., and Garcia-Molina, H. 2005. "Web Spam Taxonomy," in *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, pp. 39-47.

Hand, D. and Till, R. 2001. "A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems," *Machine Learning* (45:2), pp. 171-186.

Hevner, A., March, S., Park, J., Ram, S. 2004. "Design science in information systems research," *MIS Quarterly* (28:1), pp. 75-105.

Jindal, N. and Liu, B. 2008. "Opinion spam and analysis," in *Proceedings of the 2008 international conference on Web search and web data mining*, pp. 219-229.

Jindal, N. and Liu, B. 2007a. "Analyzing and Detecting Review Spam," in *Proceedings of the Seventh IEEE International Conference on Data Mining*, pp. 547-552.

Jindal, N. and Liu, B. 2007b. "Review spam detection," in *Proceedings of the 16th International Conference on World Wide Web*, pp. 1189-1190.

Kim, S-M., Pantel, P., Chklovski, T., Pennacchiotti, M. 2006. "Automatically assessing review helpfulness," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 423-428.

Kullback, S. and Leibler, R.A. 1951. "On Information and Sufficiency," *The Annals of Mathematical Statistics* (22:1), pp. 79–86.

Lafferty, J. and Zhai, C. 2001. "Document language models, query models, and risk minimization for information retrieval," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 111-119.

Lau, R.Y.K., Bruza, P.D., and Song, D. 2008. "Towards a Belief Revision Based Adaptive and Context Sensitive Information Retrieval System," *ACM Transactions on Information Systems* (26:2), pp. 8.1-8.38.

Lau, R.Y.K. and Lai, C.L. 2008. "Information Granulation for the Design of Granular Information Retrieval Systems," in *Proceedings of the 2008 International Conference on Information Systems*, Paris, France, Completed Research Paper 179.

Lau, R.Y.K., Song, D., Li, Y., Cheung, C.H., Hao, J.X. 2009a. "Towards A Fuzzy Domain Ontology Extraction Method for Adaptive e-Learning," *IEEE Transactions on Knowledge and Data Engineering* (21:6), pp. 800-813.

Lau, R.Y.K., Lai, C.L., Ma, J., and Li, Y. 2009b. "Automatic Domain Ontology Extraction for Context-Sensitive Opinion Mining," in *Proceedings of the Thirtieth International Conference on Information Systems*, Phoenix, Arizona, Completed Research Paper 171.

Lau, R.Y.K., Lai, C.L., and Li, Y. 2009c. "Leveraging the Web context for context-sensitive opinion mining," in *Proceedings of the 2009 IEEE International Conference on Computer Science and Information Technology,* Beijing, China, pp. 467-471.

Lau, R.Y.K. 2003. "Context-Sensitive Text Mining and Belief Revision for Intelligent Information Retrieval on the Web," *Journal of Web Intelligence and Agent Systems* (1:3-4), pp. 151-172.

Lin, Y.R., Sundaram, H., CHi, Y., Tatemura, J., Tseng, B. L. 2008. "Detecting splogs via temporal dynamics using self-similarity analysis," *ACM Transactions on the Web* (2:1), Article 4.

Liu, X. and Croft, B. 2004. "Cluster-based retrieval using language models," *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.186-193.

Liu, Y., Huang, X., An, A., Yu, X. 2008. "Modeling and Predicting the Helpfulness of Online Reviews," *Proceedings of the Eighth IEEE International Conference on Data Mining*, pp. 443-452.

Macdonald, C. and Ounis, I. 2007. "Overview of the TREC 2007 Blog Track," in *Proceedings of the Sixteenth Text REtrieval Conference*, Gaithersburg, Maryland. Available at http://trec.nist.gov/pubs/trec16/.

Macdonald, C., Ounis, I., Soboroff, I. 2009. "Is spam an issue for opinionated blog post search?" in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 710-711.

March, S. T., Storey, V. C. 2008. "Design science in the information systems discipline," *MIS Quarterly*, (32:4), pp. 725-730.

Martinez-Romo, J., and Araujo, L. 2009. "Web spam identification through language model analysis," in *Proceedings of the Fifth International Workshop on Adversarial Information Retrieval on the Web*, pp. 21-28.

Mayzlin, D. 2006. "Promotional Chat on the Internet," *Marketing Science* (25:2), pp. 155-163.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. 1990. "Introduction to WordNet: An on-line lexical database," *Journal of Lexicography* (3:4), pp. 234–244.

Mishne, G., Carmel, D., Lempel, R. 2005. "Blocking Blog Spam with Language Model Disagreement," in *Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web*, pp. 1-6.

Nadas, A. 1984. "Estimation of probabilities in the language model of the IBM speech recognition system," *IEEE Transactions on Acoustics, Speech and Signal Processing* (32:4), pp. 859.

Nie, J.Y., Cao, G., Bai, J. 2006. "Inferential language models for information retrieval," *ACM Transactions on Asian Language Information Processing*, (5:4), pp. 296-322.

Ntoulas, A., Najork, M., Manasse, M., Fetterly, D. 2006. "Detecting spam web pages through content analysis," *Proceedings of the 15th international conference on World Wide Web,* pp. 83-92.

Oreilly, T. 2007. "What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software," *Communications & Strategies*, (1), pp.17. Available at SSRN: http://ssrn.com/abstract=1008839

Piskorski, J., Sydow, M., Weiss, D. 2008. "Exploring linguistic features for web spam detection: a preliminary study," in *Proceedings of the 4th international workshop on Adversarial information retrieval on the Web*, pp. 25-28.

Ponte, J. and Croft, B. 1998. "A language modeling approach to information retrieval," in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 275-281.

Porter, M. 1980. "An algorithm for suffix stripping," *Program* (14:3), pp. 130–137.

Raman, T. V. 2009. "Toward 2^{W}, beyond web 2.0," *Communications of the ACM*, (52:2), pp. 52-59.

Salton, G. and McGill, H.J. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983

Salton, G., Wong, A., Yang, C. S. 1975. "A vector space model for automatic indexing," *Communications of the ACM*, (18:11), pp. 613-620.

Wilson, T., Wiebe, J., and Hoffmann, P. 2005. „Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pp. 347–354.

Wernerfelt, B. 1994. "On the function of sales assistance," *Marketing Science*, (13:1), pp. 68–82.

Xiao, B. and Benbasat, I. 2010. "Product-Related Deception in E-Commerce: A Theoretical Perspective," *MIS Quarterly*. Forthcoming. Available at: http://www.misq.org/archivist/vol/Queue/XiaoBenbasat.html.

Zhai, C. and Lafferty, J. 2004. "A study of smoothing methods for language models applied to information retrieval," *ACM Transactions on Information Systems* (22:2), pp. 179-214.

Zheleva, E., Kolcz, A., Getoor, L. 2008. "Trusting spam reporters: A reporter-based reputation system for email filtering," *ACM Transactions on Information Systems* (27:1), Article 3.

Zhou, B., Pei, J. 2009. "Link spam target detection using page farms," *ACM Transactions on Knowledge Discovery from Data* (3:3), Article 13.