FAKE REVIEW DETECTION USING TIME SERIES

MOHAMMADALI TAVAKOLI

A dissertation submitted in partial fulfillment of the
requirements for the award of the degree of
Master of Science (Information Technology-Management)

Faculty of Computing
Universiti Teknologi Malaysia

January 2014

I dedicate this thesis to my beloved wife Atefeh for her endless love, patience, support and encouragement. It is her unconditional love that motivates me to set higher targets.

ACKNOWLEDGEMENT

I would like to express my special appreciation and thanks to my supervisor Professor Dr.Naomie Salim, you have been a tremendous mentor for me. I would like to thank you for encouraging my research and for your brilliant comments and suggestions. Furthermore I wish to express my deep gratitude to Professor Dr. Alfred Cobsa, who gave me valuable suggestions on my research. A special gratitude I give to Dr.Mathias Joust, whom I consulted and discussed with on various research topics.

I would like to thank my family for their understanding. Finally and most importantly, I would like to thank my wife, Atefeh, for her love and support during my Master. I would never have been able to finish my dissertation without all of the sacrifices that she has made on my behalf.

ABSTRACT

Today's e-commerce is highly depended on online customers' reviews posted in opinion sharing websites that are growing incredibly. These reviews are important not only effect on potential customers' purchase decision but also for manufacturers and business holders to reshape and customize their products and manage competition with rivals throughout the market place. Moreover opinion mining techniques that analyze customer reviews obtained from opinion sharing websites for different purposes could not reveal accurate results for combination of spam reviews and truthful reviews in datasets. Thus employing review spam detection techniques in review websites are highly essential in order to provide reliable resources for customers, manufacturers and researchers. This study aims to detect spam reviews using time series. To achieve this, the novel proposed method detects suspicious time intervals with high number of reviews. Then a combination of three features, i.e. rating of reviews, similarity percentage of review contexts and number of other reviews written by the reviewer of current review, will be used to score each review. Finally a threshold defined for total scores assigned to reviews will be the border line between spam and genuine reviews. Evaluation of obtained results reveals that the proposed method is highly effective in distinguishing spam and non-spam reviews. Furthermore combination of all features used in this research exposed the best results. This fact represents the effectiveness of each feature.

ABSTRAK

Hari ini e -dagang adalah sangat bergantung kepada ulasan pelanggan talian ' yang dicatatkan pada laman web perkongsian pendapat yang berkembang sangat. Ini ulasan adalah penting bukan sahaja memberi kesan pada keputusan pembelian pelanggan berpotensi ' tetapi juga untuk pengeluar dan pemegang perniagaan untuk membentuk semula dan menyesuaikan produk mereka dan menguruskan persaingan dengan pesaing di seluruh pasaran. Teknik perlombongan pendapat lebih-lebih lagi yang menganalisis pelanggan yang diperolehi daripada laman web perkongsian pendapat untuk tujuan yang berbeza tidak boleh mendedahkan keputusan yang tepat untuk kombinasi ulasan spam dan ulasan benar dalam dataset. Oleh itu menggunakan kajian spam teknik pengesanan dalam kajian laman web adalah sangat penting untuk menyediakan sumber-sumber yang boleh dipercayai untuk pelanggan, pengeluar dan penyelidik. Kajian ini bertujuan untuk mengesan ulasan spam menggunakan siri masa. Untuk mencapai matlamat ini , kaedah yang dicadangkan novel mengesan jarak masa yang mencurigakan yang mempunyai bilangan ulasan. Kemudian gabungan tiga ciri-ciri, iaitu penarafan ulasan, peratusan persamaan kajian konteks dan beberapa ulasan lain yang ditulis oleh pengulas kajian semasa, akan digunakan untuk menjaringkan setiap kajian semula . Akhirnya ambang yang ditetapkan untuk jumlah markah yang diberikan kepada ulasan akan garis sempadan antara spam dan ulasan tulen. Penilaian keputusan yang diperolehi menunjukkan bahawa kaedah yang dicadangkan adalah amat berkesan dalam membezakan spam dan bukan spam - ulasan. Tambahan pula gabungan semua ciri-ciri yang digunakan dalam kajian ini didedahkan hasil yang terbaik. Fakta ini mewakili keberkesanan setiapciri.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1. Introduction

With the development of internet, people became more confident to explain their thoughts on websites and share them with millions of people. Web 2.0 slowly changed different aspects of people living. For instance, by creating online groceries, a huge number of daily trades are virtualized. Nowadays people are more dependent to the internet for purchasing products and services. Long time ago, when they wanted to purchase a product, the best method was asking other customers who have purchased it before and know about the quality of that product very well to ensure that they will have a successful transaction. Similarly now they can visit customer reviews about various products or services that they tend to purchase via opinion sharing websites. Hence they can easily trade off the pros and cons of a specific good.

The increasingly propensity of people to use online opinion sharing websites has created a challenging situation for manufacturers, business holders and stores. Hence dishonest producers who tend to control and optimize the customers' opinions flow on their products and brand attempt to publish fake reviews among review websites. Sometimes they hire individual or in some cases groups of spammers to create not only glamorized positive reviews on their products but also harmful negative reviews on competitors'. These types of non truthful reviews motivate

customers to find their products the best option to purchase among similar products offered by different brands.

Fake opinions are extremely harmful for business holders. Therefore opinion mining techniques are assisting business to analyze posted customers' opinions on offered products to detect spam reviews and proffer truthful reviews to purchasers. However research in this area is not adequate and many critical problems related to spam detection are not solved yet.

1.2. Problem Background

In comparison with other types of spam such as e-mail spam and web spam, review spam detection is very complicated because manual evaluation of reviews and distinguishing fake reviews from real opinions is very hard, if not impossible (Jindal and Liu, 2008). Hence state-of-the art methods in detecting various types of spam are not applicable in review domain. Accordingly review spam detection is a different and complex problem in Natural Language Processing area.

Various researchers proposed different methods and algorithms to detect fake reviews. Algur *et al.*(2010) used conceptual feature similarity to detect spam reviews. They extracted features from review database and store them in a feature database. Then extracted features were used in constructing a feature matrix *M* with *n* columns and *m* rows. Where *n* indicates the number of reviews and *m* represents extracted features from them. Consequently they categorized reviews into four groups i.e. duplicate reviews, near duplicate reviews, partially related reviews and unique reviews. This categorization was based on similarity of features among them. First two categories in their approach were considered as spam and the rest of them as truthful reviews. Further they classified reviews to spam and non-spam by defining rules to separate spam reviews, specifying a threshold, and analyzing the matrix. In this research they assumed products features similarity as a factor to detect

spam reviews. The purpose of product features is specific parts, attributes and aspects of a product that is mentioned in the content of a review. However by revising reviews of a product one can observe that many opinion holders talked about one or more similar features which are disaster elements or privileged aspects of a product (e.g. battery life in a digital camera) which this method consider them as spam reviews. In the other words this method measures the similarity of product features between reviews and considers two reviews with high level of similarity as spam reviews.

Likewise, an important initial research on spam detection which similarly used duplication reviews is done by Jindal and Liu(2007b,2008). The duplication in aforementioned study was in product features. However in Jindal and Liu focused on duplication in context of the reviews. Firstly they used Shingle method approach proposed in (Broder, 1997) to find duplicate and near duplicate reviews. A 2-gram based review content comparison was used in their approach to identify the similarity of review contents. Then they performed logistic regression to detect spam reviews on brands and non review texts among reviews by manually labeling 470 fake reviews. 36 features were used in their approach in order to classifying spam and non-spam reviews. A critical point in their approach is that if a duplicate review is from a same person but on two models of a same product (e.g. a Samsung DVD player and a Samsung T.V) it will be considered as spam, yet it might be a truthful review. In addition 100% duplicate reviews might be the result of pressing submit button frequently by an innocent opinion holder.

In an approach proposed by Li *et al.* (2010) a co-training algorithm with two views was designed to skip the onerous task of manual annotation in their supervised learning framework. The views for each review were review features and reviewer features. The other reason of using co-training algorithm was utilizing unlabeled reviews to as train data during run time. Assuming that the more malicious reviews have less helpfulness rate was the foundation of their approach. They annotated 1398 spam reviews which had low-helpfulness rate out of 6000 reviews. For the supervised method they used public machine learning software Weka to perform

SVM, Bayes and logistic regression. The authors argued that Bayes achieved best results on their dataset.

On one hand, in their approach unlike single heuristic methods used in previous studies, authors used a two view co-training method to detect spam reviews. On the other hand, one cannot assume that low-helpfulness feature is the main factor of detecting spam reviews. Additionally more than 10% of their manually detected spam reviews are from top and middle helpful set reviews.

Two other studies (Yoo and Gretzel, 2009; Ott *et al.*, 2011) had worked on datasets of hotels reviews crawled from TripAdvisor.com. The first one studied on lexical complexity differences and using brand names and first person pronouns between fake and honestly reviews. They measured quantity (number of words included in a review) and lexical complexity of reviews (average length of each word) using Microsoft Word word count tool and rate of recurrence of unique words, pronouns and brand name using CATPAC. They also used General Inquirer to percentage positive and negative words in sentences. Consequently the second study used mentioned method to obtain 400 truthful reviews about top 20 hotels from a review website. Then they created 400 fake reviews for those hotels by exploiting Amazon Mechanical Turk that is a service provided by Amazon. It performs human needed computational tasks such as data annotation, reviewing products, and creating texts. They defined three tasks: firstly they compare distribution of POS tags between fake and ingenuous reviews by making features for each review depends on frequency of POS tags in it. Secondly to detect personality traits or Psycholinguistic deception detection they used the Linguistic Inquiry and Word Count (LIWC) software. Then they created classifier from its output features. Thirdly they used n-gram-based classifier in content and concept of reviews to label honest and fake ones. Finally they used these three approaches to train SVM and Bayes classifiers. They argued that using standard n-gram-based categorization approaches perform better in deception detection than keyword based deception cues like LIWC. It is even more effective by combination with psycho linguistically motivated features.

The three types of abovementioned approaches have mainly focused on content and context of a review. Yet there are many spammers that write their genuine experience about a really purchased product for a non-purchased product in order to spam it (e.g. the spammer has a Canon camera and write positive spam reviews for Nikon camera based on his experience of Canon camera). In these common cases focusing on context and content of reviews is not efficient any more.

Additionally, a novel study in this area is done by Xie *et al.*(2012a). The approach attempts to detect singleton spam reviews. A singleton review is the only review written by a reviewer. The authors assumed that reviewers' behaviors can be divided into two phases: arrival phase, when a customer purchase a product or a spammer hire and writing phase, when they start developing reviews. They analyzed spammers and customers behaviors in normal arrival, promotion arrival and spam attack arrival. Accordingly they found that spammers start writing phase immediately after arrival but customers have delay for receiving product and testing it. In other words attacks tend to create a burst on review arrival process which is dissimilar with customers. Therefore the authors focused on joint nonstandard patterns in arrival phase and rating to do their task. The scope of their study is on cases that rating is promoted dramatically. They formed a three dimensional time series to capture behaviors. Then they tried to find unusual blocks in the time series using a three part algorithm obtained from previous researches. Finally they proposed a framework to detect singleton spam reviews. Consequently in another method proposed in (Fei *et al.*, 2013) review burst pattern were used to detect spammers. The authors generated 5 new spammer behavioural features as indicators to be used in review spammer detection. 1) Ratio of Amazon verified purchase (AVP) 2) Rating deviation 3) Burst review ratio 4) Review content similarity 5) Reviewer burstiness. All mentioned features are demonstrated in Chapter 2. Their method reveals more accurate results comparing with abovementioned approaches (Xie *et al.*2012a). However one of these 5 used features is 'Ratio of Amazon verified purchase' which possibility of using this feature in any detection technique optimizes the accuracy of the method profoundly.

All in all, various approaches are proposed to detect spam reviews, individual and group spammers and suspicious behavior in reviews by focusing on different aspects such as reviewer behavior, review content, comparing review features with surrounding reviews, reliability of the product, and so on. However there are many problems in this area to classify suspicious and truthful reviews. Accuracy of some methods represents that it is not competent in filtering maximum spam or in preventing truthful reviews to be detected as spam. Furthermore some aspects are not studied by researchers or few researches are done on them. The following Table (Table 1.1) represents the major proposed techniques in review spam detection. The precision is computed using a micro-average, i.e., from the aggregate true positive, false positive and false negative rates, as suggested by Forman and Scholz (2010).

Table 1.1: List of major review spam detection methods

Study	Title	Year	Methods & Techniques	Results
Jindal and Liu, 2008	Opinion Spam and Analysis	2008	Detect & labeling duplicate reviews as spam. Then using SVM, Naive Bayes and logistic regression to classify spam and non spam reviews	Precision 85%
Algur et al., 2010	Conceptual level Similarity Measure based Review Spam Detection	2010	constructing the matrix of product features and detecting similar reviews as spam	precision 43.6%
Li et al., 2011	learning to identify review spam	2011	using co-training algorithm with two views i.e. review features and reviewer features	Precision 64%
Ott et al., 2012	Estimating the prevalence of deception in online review communities	2012	classification with linguistic features	precision 83.3%
Wang et al., 2012	Identify Online Store Review Spammers via Social Review Graph	2012	graph, iterative algorithm	Precision 49%,
Xie et al., 2012	Review Spam Detection via Temporal Pattern Discovery	2012	Three dimensional time series(ratio of singleton reviews, rating, number of reviews)	Precision 61.11%
Lim et al., 2010	Detecting Product Review Spammers using Rating Behaviors	2010	rating behaviors	Precision 78%,
Fei et al., 2013	Exploiting Burstiness in Reviews for Review Spammer Detection	2013	Kernel Density Estimation techniques with proposed features	Precision 83.7%,

According to the table, the most accurate result belongs to (Jindal and Liu, 2008). However the authors considered duplicate reviews as spam reviews. Duplication might be a mistake from an innocence reviewer, yet will be considered

as spam in their method. Therefore the accuracy of their proposed technique will be profoundly apposed by this fact. Another high accurate method in the table is the one proposed by Ott et al.(2012). They have produced the fake reviews using Amazon Mechanical Turk. Employing human resources to do this task will be highly effective on the obtained result. Additionally, there were not any features to detect spam reviews except content based features that relying on them is not adequate to detect spam reviews in real situation. Therefore the method might be fragile in detecting real spam reviews. Finally it is demonstrated that methods employing posting time factor performed well. The focus of this study will be on detecting the burst pattern of spam attacks as a strong evidence for detecting fake reviews fallen in attacks durations.

1.3. Problem Statement

It is generally accepted that annotating a 100 percent accurate spam reviews dataset collected from opinion sharing websites is impossible. Therefore a genuine review not only might be annotated as spam but also might be detected as spam in proposed methods. Distribution of spam reviews among all reviews could be closely related to points of time that spammers are hired and started attacks. Therefore detecting abnormal oscillations in reviewing flow for a product or brand could be a strong evidence of spam attacks. Assessing reviews fallen in spam attacks duration aggrades the accuracy of a method in detecting spam reviews.

1.4. Objectives of the Study

1. Detecting spam reviews using time series.
2. Identifying spam attacks using oscillations of number of reviews over the time.
3. Scoring reviews fallen in time intervals with high oscillations based on percentage of similarity of the review text with other reviews, rating deviation of the review and number of reviews written by a person fallen in a similar interval.
4. Detecting spam reviews using assigned spam scores.

1.5. Scope of the Study

In this study the focus will be on a dataset collected from a unique review website. Among various numbers of products and brands in the dataset, reviews of products that are produced by Nikon Company found the corpus of this project. Furthermore regarding to difficulty of manual annotation of spam review datasets which is almost impossible in majority of review datasets, the corpus is limited to number of 244 reviews. The corpus is selected from Nikon brand reviews with searching and reviewing all the reviews and discarding irrelevant parts.

1.6. Significant of Study

Today methods of purchasing products by people are profoundly different from erstwhile. Most of the customers review purchased products online and others who tend to buy a similar product will search opinion sharing websites to make the best decision. This situation promotes competition between merchants, business holders, manufacturers and even famous stores. Thus some of them attempt to perk

between competitors and highlight their products and corresponding features and aspects on review websites. Promoting their products against competitors or vise versa is the reason of hiring review spammers to do this task.

Opinion mining researchers in contrast, focused on detecting and discarding these spam reviews to moderate the market in a real and fare situation. Many approaches are proposed by them using various aspects of reviews to detect spam. Spammers in the other hand are becoming smarter and optimize their methods as they cannot be detected by majority of approaches.

The most important point in review spamming is the role of producers in this game. Times of hiring spammers by them which is a critical method to detect spam reviews could be detected by following proportion of their products in reviews and between competitors.

1.7.Conclusion

All things considered, one can say that considering the appearance of smarter spammers, review spam detection research needs more attention from researchers. New methods of spamming cannot be captured by majority of state of the art spam detection approaches. However spam attacks that will be starts a bit after hiring spammers by business holders could be detected using abnormal oscillations in number of reviews for a product or brand over the time supported with spammers' atypical behaviors.

REFERENCES

Algur *et al.* (2010). "Conceptual level Similarity Measure based Review Spam Detection." International Conference on Signal and Image Processing: 8.

Bing and Minqing (2004). Mining and summarizing customer reviews. KDD.

Blitzer *et al.* (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. Annual Meeting-Association For Computational Linguistics.

Broder (1997). " On the resemblance and containment of documents " Compression and Complexity of Sequences: 9. IEEE Computer Society.

Dave *et al.* (2003). Mining the peanut gallery: opinion extraction and semantic classification of product reviews. WWW.

Etzioni (2005). Extracting Product Features and Opinions from Reviews. EMNLP.

Fei *et al.* (2013). Exploiting Burstiness in Reviews for Review Spammer Detection. Seventh International AAAI Conference on Weblogs and Social Media.

Ge and Smyth (2001). "Segmental Semi-Markov models for endpoint detection in plasma etching." IEEE Transactions on Semiconductor Engineering.

Hu and Liu (2004). Mining and summarizing customer reviews. KDD.

Jindal and Liu (2007a). "Analyzing and Detecting Review Spam." Seventh IEEE International Conference on Data Mining **68**: 6.

Jindal and Liu (2007b). "Review Spam Detection." World Wide Web Conference Series: 2.

Jindal and Liu (2008). "Opinion Spam and Analysis." Conference on Web Search and Web Data Mining: 11.

Jindal *et al.* (2010). Finding unusual review patterns using unexpected rules. Proceedings of the 19th ACM international conference on Information and knowledge management, ACM.

Lai *et al.* (2010). Toward A Language Modeling Approach for Consumer Review Spam Detection.

IEEE 7th International Conference**:** 8.

Li *et al.* (2010). Learning to Identify Review Spam. Twenty-Second International Joint Conference on Artificial Intelligence.

Lim *et al.* (2010). Detecting Product Review Spammers using Rating Behaviors. international conference on Information and knowledge management 10.

Liu (2012). "Sentiment analysis and opinion mining." Synthesis Lectures on Human Language Technologies **5**(1): 1-167.

Morales *et al.* (2013). Synthetic review spamming and defense. Proceedings of the 22nd international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee.

Mukherjee *et al.* (2012). Spotting Fake Reviewer Groups in Consumer Reviews. Proceedings of international world web conference

Mukherjee† *et al.* (2011). Detecting Group Review Spam. International conference on WWW, India.

Ott *et al.* (2012). Estimating the prevalence of deception in online review communities. Proceedings of the 21st international conference on World Wide Web, ACM.

Ott *et al.* (2011). "Finding Deceptive Opinion Spam by Any Stretch of the Imagination." 49th annual meeting of the association for the computational linguistics: 11.

Pennebaker *et al.* (2007). The development and psychometric properties of LIWC2007. Austin, TX, LIWC.Net.

Popescu and Etzioni (2005). Extracting product features and opinions from reviews. HLTEMNLP.

Rayson *et al.* (2001). "Grammatical word class variation within the British National Corpus sampler." Language and Computers **36**: 11.

S *et al.* (2008). A holistic lexicon-based approach to opinion mining. WSDM.

Shashirekha *et al.* (2009). Ontology based similarity measure for text documents. International Conference on Signal and Image processing (ICSIP-2009).

Sun *et al.* (2013). Synthetic review spamming and defense. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.

Wang *et al.* (2011). Review Graph based Online Store Review Spammer Detection. IEEE International Conference on Data Mining - ICDM**:** 6.

Wu *et al.* (2010). "Distortion as a Validation Criterion in the Identification of Suspicious Reviews." social media analytics: 4.

Xie *et al.* (2012a). Review Spam Detection via Temporal Pattern Discovery, 18th ACM SIGKDD international conference on Knowledge discovery and data mining 9.

Xie *et al.* (2012b). Review Spam Detection via Time Series Pattern Discovery. 21st international conference companion on World Wide Web. ACM New York, NY, USA**:** 2.

Yoo and Gretzel (2009). Comparison of Deceptive and Truthful Travel Reviews. Information and Communication Technologies in Tourism**:** 11.

Forman and Scholz (2010). "Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement." ACM SIGKDD Explorations Newsletter **12**(1): 49-57.