OPINION SPAM DETECTOR: DETECTING SPAM REVIEWS

Felarca, John Denver P. [1], Militar, Jessica Justine B.[1], Sagum, Ria A. [2] ,
Santos, Clint Lennard B.[1], Zubiri, Ceasar Jemrick D.[1]

[1] Bachelor of Computer and Information Sciences, College of Computer and Information Sciences, Polytechnic University of the Philippines, Sta. Mesa, Manila, Philippines

[2]Master of Science in Computer Science, De la Salle University
Email: rasagum@pup.edu.ph

ABSTRACT

Consumers increasingly rate, review and research products online. Consequently, websites containing consumer reviews are becoming targets of opinion spam. While recent work has focused primarily on manually identifiable instances of opinion spam, in this work we study deceptive opinion spam—fictitious opinions that have been deliberately written to sound authentic. Integrating work from behavioral features and text analysiis, we develop and compare three approaches to detecting deceptive opinion spam, and ultimately develop a classifier from our spam dataset.

Categories and Subject Descriptors

I.2.7: **[Natural Language Processing**]: Text Analysis H.3.3 **[Information Storage and Retrieval]**: Information Search and Retrieval – Information filtering. H.2.8: [**Database Management]**: Database Applications – Data mining

General Terms

Reviews, Algorithms, Dataset

Keywords

Review Spam, Spam detection, Natural Language Processing, Information Extraction, Data Mining

1. INTRODUCTION

Consumer's purchase decisions are increasingly influenced by user-generated online reviews of products and services [1]. The number of consumers that first read reviews about a product they wish to buy is constantly on the rise. Technology research company Gartner Inc. claims 31% of consumers read online reviews before actually making a purchase Gartner [2]. As consumers increasingly rely on these ratings, the incentive for companies to try to produce fake reviews to boost sales is also increasing. Gartner predicts in 2014 as much as 15 percent of all social media reviews will consist of company paid fake reviews, Gartner [3]. Deceptive reviews have at least

two major damaging effects for the consumers. First, they lead the consumer to make bad decisions when buying a product. After reading a bunch of reviews, it might look like a good choice to buy the product, since many users praise it. After, it turns out the product quality is way below expectations and the buyer is disappointed. Second, the consumer's trust in online reviews drops. Accordingly, there is a growing incentive for businesses to solicit and manufacture DECEPTIVE OPINION SPAM— fictitious reviews that have been deliberately written to sound authentic and deceive the reader [1]. For example, Ott et al. (2012) has estimated that between 1% and 6% of positive hotel reviews appear to be deceptive, suggesting that some hotels may be posting fake positive reviews in order to hype their own offerings. While previous related work has explored characteristics of positive deceptive opinion spam, the complementary problem of negative deceptive opinion spam remains largely unstudied. Opinion spam can range from annoying self-promotion of an unrelated website or blog to deliberate review fraud, as in the recent case of a Belkin employee who hired people to write positive reviews for an otherwise poorly reviewed product. While other kinds of spam have received considerable computational attention, regrettably there has been little work to date on opinion spam detection. Furthermore, most previous work in the area has focused on the detection of DISRUPTIVE OPINION SPAM—uncontroversial instances of spam that are easily identified by a human reader, e.g., advertisements, questions, and other irrelevant or nonopinion text (Jindal and Liu, 2008). And while the presence of disruptive opinion spam is certainly a nuisance, the risk it poses to the user is minimal, since the user can always choose to ignore it. Lastly, our study of deceptive opinion spam detection as a genre identification problem reveals relationships between deceptive opinions and imaginative writing, and between truthful opinions and informative writing.

2. RELATED WORK

The opinion spam problem was first formulated by Jindal and Liu in the context of product reviews, Jindal and Liu [4]. By analyzing several million reviews from the popular Amazon.com, they showed how widespread the problem of fake reviews was. The existing detection methods can be split in the context of machine learning into supervised and unsupervised approaches. Second, they can be split into three categories by their features: behavioral, linguistic or those using a combination of these two. They categorized spam reviews into three categories: non-reviews, brand-only reviews and untruthful reviews. The authors ran a logistic regression classifier on a model trained on duplicate or near-duplicate reviews as positive training data, i.e. fake reviews, and the rest of the reviews they used as truthful reviews. They combined reviewer behavioral features with textual features and they aimed to demonstrate that the model could be generalized to detect non-duplicate review spam. This was the first documented research on the problem of opinion spam and thus did not benefit from existing training databases. The authors had to build their own dataset, and the simplest

approach was to use near-duplicate reviews as examples of deceptive reviews. Although this initial model showed good results, it is still an early investigation into this problem.

Other researches, Ott et al. [1] used a bag-of-words approach and calculated the frequency of certain words from the review text. They then classified some reviews as suspicious if the text contained a high number of predefined suspicious words. This led to more subjective conclusions that spammers prefer to use more personal pronouns than genuine reviewers or they usually write reviews of more than 150 characters on average. The authors cataloged some words, e.g. “vacation” and “husband” as highly suspicious. They concluded these couple of words appeared more often in the fake reviews created through Amazon Mechanical Turk, but one can hardly say that a review containing the word “vacation” is 100% fake. An obvious aspect is that once the spammers find out about these textual frequency traps which cause suspicion, they will simply avoid them.

3. METHODOLOGY

This study aims to classify spam vs non-spam reviews in blogs and forum. This paper analyzes such spam activities and presents some novel techniques to detect them.

3.1 Data Volume

Our dataset is shared by Bing Liu, Amazon Product Review Data (Huge) used in (Jindal and Liu, WWW-2007) for review spam (fake review) detection. It has information about reviewers, review text, ratings, product info, etc. These additional data allow us to create more useful features for building machine learning models to spot review spammers.

3.2 Data Richness

Compared with other datasets, reviews in our dataset come with a much richer context. These additional data allow us to create more useful features for building machine learning models to spot review spammers.

We have used supervised learning, pattern discovery, and graph-based methods, and relational modeling to solve the problem.

Review content:

- Lexical features such as term-frequency features word n-grams, part-of-speech n-grams, and other lexical attributes.

4. RESULTS

Reviewers can use various devices. In the main site of Amazon, spammers can quickly start writing fake reviews once registered. Consequently, the percentage of reviews with fake review labels posted from Amazon main site is relatively higher than other sites like Yelp and Danping.

Experimental results showed that semantic similarity can outperform the vectorial model in detecting deceptive reviews, capturing even more subtle textual clues. The precision score of the review classifier showed high results, enough to make the method viable to be integrated into a

production detection system. The results of the topic modeling method showed that combining opinion spam detection with topic modeling may offer, for some number of topics, results in the vicinity of the semantic similarity models.

5. DISCUSSION

There are two directions where the research on opinion spam has focused on so far: behavioral features and text analysis. Behavioral features represent things like the user's rating of a product, the date when the user posted the review, the IP from where the review was posted, how many previous reviews the user made before and so on. Textual analysis refers to methods used to extract clues from the review content, such as the frequency of personal pronouns in the text, or if a high number of predefined suspicious words are used.

The first method so far seems to be more reliable and can be more easily put into practice. It also offers very good results as a standalone method, although the textual features do bring little overall improvement. This is first of all due to the complexity and hardness of implementing language processing methods in general. The textual analysis techniques have shown less precision on detecting opinion spam, though they improve the overall accuracy when added on top of the behavioral features. The linguistic techniques used so far mostly consisted of computing cosine similarity between the contents of the reviews. In a new study Zengin and Carterette [5], the authors concluded that human judgment used to detect semantic similarity of web document does not correlate well with cosine similarity.

This project is focused on a potentially more insidious type of opinion spam: DECEPTIVE OPINION SPAM—fictitious opinions that have been deliberately written to sound authentic, in order to deceive the reader. For example, one of the following two hotel reviews is truthful and the other is deceptive opinion spam (Liu, 2008).

1. I have stayed at many hotels traveling for both business and pleasure and I can honestly say that The James is tops. The service at the hotel is first class. The rooms are modern and very comfortable. The location is perfect within walking distance to all of the great sights and restaurants. Highly recommend to bot business travelers and couples.

2. My husband and I stayed at the James Chicago Hotel for our anniversary. This place is fantastic! We knew as soon as we arrived we made the right choice! The rooms are BEAUTIFUL and the staff very attentive and wonderful!! The area of the hotel is great, since I love to shop I couldn't ask for more!! We will definitely be back to Chicago and we will for sure be back to the James Chicago.

6. CONCLUSION

Our results demonstrate that while human deception detection performance is greater for negative rather than positive deceptive opinion spam, the best detection performance is still achieved through

automated classifiers, with approximately 86% accuracy.

Standard n-gram–based text categorization techniques have been shown to be effective at detecting deception in text. Following Ott et al. (2011), we evaluate the performance of linear Support Vector Machine (SVM) , Naïve Bayes, Linear Regression classifiers, trained with term-frequency features on our novel negative deceptive opinion spam dataset.

7. RECOMMENDATIONS

Possible directions for future work include an extended evaluation of the methods proposed in this work to both negative opinions, as well as opinions coming from other domains. Many additional approaches to detecting deceptive opinion spam are also possible, and a focus on approaches with high deceptive precision might be useful for production environments.

8. REFERENCES

[1] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Lin

[2] Gartner. Solve the problem of fake online reviews or lose credibility with consumers, 2013. URL http://www.gartner.com/id=2313315.

[3] Gartner. Gartner says by 2014, 10-15 percent of social media reviews to be fake, paid for by companies, 2012. URL http://www.gartner.com/newsroom/id/2161315.

[4] Nitin Jindal and Bing Liu. Opinion spam and analysis. In Proceedings of the 2008 International Conference on Web Search and Data Mining, WSDM '08, pages 219–230, New York, NY, USA, 2008. ACM. ISBN 978-1-59593927-2. doi: 10.1145/1341531.1341560.

[5] Mustafa Zengin and Ben Carterette. User judgements of document similarity. Clarke et al.[19], pages 17–18, 2013.