Detecting Review Spam: Challenges and Opportunities

Yingying Ma and Fengjun Li
Department of Electrical Engineering and Computer Science
The University of Kansas, Lawrence, KS 66045
Email: {elainema, fli}@ku.edu

***Abstract*—Online customer reviews for both products or merchants have greatly affected others' decision making in purchase. Considering the easily accessibility of the reviews and the significant impacts to the retailers, there is an increasing incentive to manipulate the reviews, mostly profit-driven. Without proper protection, spam reviews will cause gradual loss of credibility of the reviews and corrupt the entire online review systems eventually. Therefore, review spam detection is considered as the first step towards securing the online review systems. In this paper, we aim to overview existing detection approaches in a systematic way, define key research issues, and articulate future research challenges and opportunities for review spam detection.**

***Index Terms*—Review spam, review spammer, spam behavior.**

I. Introduction

People's attitudes and opinions are highly influenceable by others, which is known as the *word-of-mouth* effect in shaping decision making. The Internet and Web-based technologies have created vast opportunities to enable online word-of-mouth carriers that play a critical role in influencing consumer purchase decision in electronic commerce. People exchange opinions about products or merchants in online blogs, forums, social media, or directly post reviews in various reputation systems provided by individual online retailers, mega-retailers (e.g., eBay, Amazon), or third-party sites (e.g., Bizrate, resellerrating.com, Google+ Local, Yelp, etc.). Recent surveys show that 83% of the consumers check out online reviews to know about the products or businesses they are buying from [1], and 80% of the consumers have changed purchase decision due to negative reviews [2].

Given the user-generated nature of online reviews and the increasing impact on purchase decision making, the quality and credibility of the reviews becomes a primary concern. Many review sites allow consumers to rate products or stores, write detailed comments, or assess others' reviews (e.g., labeling as "helpful") to express individual opinions. The ratings and reviews are highly subjective and often individually biased. Moreover, with profit incentives, unreliable reviews with dishonest ratings, known as *review spams*, are intentionally inserted into online review systems, e.g., product manufacturers, competitors or professional online reputation management companies may fake false positive (a.k.a *ballot stuffing*) or maliciously negative reviews (a.k.a. *bad mouthing*) to promote or demote a product, or to attract customers to a store or distract them from competitors. Driven by profits, a large number of spam reviews inserted manually or automatically by professional review management companies have been observed on many well-known online reputation systems. The gradual loss of credibility of online reviews will keep confusing the consumers with poor or wrong assessment and eventually cause the corruption of the review system. Review spam detection is the first step towards securing the online review systems. In this paper, we aim to overview existing detection approaches in a systematic way, define key research issues, and articulate future research challenges for review spam detection.

II. Review Spam Detection: Overview

A. Online Reviews and Review Spams

Reviews are the *user generated contents* provided by *users* to express personal opinions about *objects*. Reviews about merchandise, books, movies, news, services, etc. are considered as *product* reviews [3]–[5], and reviews that express overall assessments of organizations (or stores) are classified as *store* reviews [6]–[8]. Different behaviors of the spammers have been observed when inserting fraudulent reviews into two types of review systems. A review typically includes a rating, a narrative comment about the object, or both. Correspondingly, there exists two types of review spams, inserting dishonest ratings or inserting unreliable comments.

Considering the increasing damage caused by review spam, it is a critical and urgent task to detect review spam automatically. However, this is unsurprisingly difficult since it is hard to filter out, even manually, a spam review or capture a spammer behavior. The reason may be two-fold, the *subjective nature of the reviews* and the *human-generated contents and patterns* that disguise spam behavior.

As a subjective concept, review varies among individuals. The display order of reviews often introduces a presentation bias. Studies showed that reviewers (with no strong opinions) are often influenced by previous reviews

978-1-936968-36-7 © 2012 ICST
DOI 10.4108/icst.collaboratecom.2012.250640

(by simply"echoing" without new input). Also, the reviewers may form an "expectation" for a product or service from priori reviews so that their judgement guided by the expectation may be biased [9]. Hence, the biased or poor-qualities reviews should be distinguished from fraudulent reviews that are often added to the review system with a clear intention or goal to achieve.

Most spam reviews currently observed are written by human spammers who intentionally write in a similar way as genuine reviewers to make fictitious contents sound authentic so as to avoid being detected [5].

B. Existing Review Spam Detection Approaches

Comparing with other spam detection approaches, such as email spam, Web spam and social spam, review spam detection mechanism has not been extensively studied until recently. In [10], the problem of product review spam was first introduced, different approaches have been proposed to handle product review spam detection [3]–[5], [11]–[13], most of them focuses on product review spam detection.

1) Product Review Spam Detection: For product reviews, spam detection is either (1) content based – finding (near) duplicate opinion items or (2) behavior based – finding the deviation of ratings from authentic ratings or patterns that match the behavior assumption of the spammers.

In [10], a large number of *duplicate* or *near-duplicate* reviews have been observed at online review sites. These duplicate or near-duplicate reviews are either written by a same reviewers on different products or by different reviewers on a same product, or by different reviewers on different products with highly similar reviews. Similarity-based methods were adopted to identify near-duplicate content, which was considered as spam reviews and used as labeled training data to build the discriminator for 2-class classification. More types of reviews such as non-review and reviews only on brand have been considered as spam in [14] and automatically labeled out to construct the classifier for supervised learning based detection similar to [10].

The unexpected pattern in rating distribution [4] or rating changes [3] have been investigated to explore the existence of strange behaviors that imply abnormal ratings or an unexpected rapid boosting or downgrading change in ratings and a suspicious time interval in which such change occurs. These detection approaches believe that spam is inserted to fulfil a certain goal, which in turn causes "unexpectedness", and use rule mining method to detect the anomaly. [15] studied typical spammer behavior and came up with four types of suspicious spamming behaviors. While most of them have also been observed by previous work, [15] constructed corresponding evaluations using unsupervised learning to measure the degree of spam. Another observation about the spammer is the group collusion phenomenon. To quickly achieve the goal of spamming – manipulating the rating or the sentiment on the target product – a set of spammer will behave in a similarly suspicious way. The collective actions of grouped spammers actually provide more hints to dig out the correlations. Hence, the detection of group spam involves finding patterns of suspicious group behaviors and mining candidate groups and their members [11]. In [16], typical behaviors of spam group members such as content similarity and rating deviation are categorized into eight indicators to calculate group spamicity atop three types of inter-relationships between product, groups, and group members.

2) Store Review Spam Detection: Unlike product reviews that each review targets a particular item, store reviews introduce more chaos: reviews of different items co-exist under a same object, reviews from a same reviewer or similar reviews from different reviewers should not be considered as duplicates, the review content allows to be from most specific to highly abstract, etc. All these characteristics increase the difficulty in detecting store review spams. As it is hard to obtain features from text mining, [6], [7] proposed a review graph model to capture the relationships among three entities, *review*, *reviewer*, and *store*. A trustiness (or reliability) score is assigned to each of the three entities, and the mutual impact of one score on the other two is addressed in an iterative algorithm. Though the review graph model is new in detecting review spam, it is similar to traditional reputation systems where the score indicates the reputation of the node.

Conceptually the review graph incorporates trust for each entity in the system and models the interaction among them. However, in real store review datasets, a large portion of singleton reviews exists, which indicates a large number of isolated reviewer-review relations in the review graph. As the review graph may not work properly, features extracted from suspicious spamming behaviors should be adopted for detection. Based on an assumption that the bursty arrival pattern is correlated to a positive or negative burst in ratings, [8] suggested to use time series pattern to find bursts in rating as well as the bursts in the number of received singleton reviews in given time windows.

III. Spam Detector Design Challenges

As the study of review spam and its detection is still in its early stage, current design of detection mechanisms is associated with a number of challenges. In order to develop a complementary solution, we need to understand these major obstacles.

A. Data Acquisition

A useful tool for tackling spam review detection problem is machine learning, which has been adopted in almost all of the existing detection attempts. The hearth of learning process is data, more specifically an accurate and diverse ground-truth data set.

However, one of the fundamental problems in applying machine learning methods (especially the supervised learning methods) to spam detection is the shortage of labeled data (or gold-standard annotated data [13]). In review spam detection, it is easy to collect a huge amount of unlabeled data (i.e., raw reviews) from online review systems. However, the acquisition of labeled data is difficult and costly. On the other hand, the quality of training data plays a critical role in developing accurate classifiers.

A common method used in most of the existing approaches [15], [16] for acquiring labeled data involves recruiting one or multiple skilled human agents. For example, eight expert judges were employed in [16] to label out 2431 reviews as spam, non-spam, and borderline. Though the judges are assumed to have domain expertise (e.g., [15] selected tertiary students familiar with searching and reading product reviews), the difficulty in correctly labeling spam and non-spam reviews is significantly larger than the ones in other applied machine learning approaches since the reviews are subjective descriptions in which the boundary between personalized review, poor-qualitied review and fake review is unclear. In this sense, the reliability of human judges to correctly tell the fake reviews from the real reviews is the primary challenge for all schemes based on supervised learning.

Several heuristic methods have been presented to improve the correctness of data labeling and thus the quality of training data. Guidelines on distinguishing legal and spamming reviews were collected from consumer studies with extensive domain knowledge to form a list of spamming signals or indicators. In [16], heuristic indicators (e.g., extreme ratings with empty adjectives) were provided to human judges to help making decisions. However, such indicators, if improperly abstracted or presented, may unavoidably introduce bias to the labeled data. Amazon Mechanical Turk was first used in [5] to collect arbitrary reviews for a set of target stores, where the reviews submitted by Turkers who are primarily driven by monetary rewards were considered a gold-standard suspicious reviews.

B. Behavior Patterns

From the overview of the existing detection approaches, we see that most schemes highly depend on the modeling of suspicious spammer behaviors, which is abstracted and derived from the observations. However, the observations may be too limited and the derivation may be too intuitive to provide a sound and comprehensive explanation. For instance, while the spam detection model in [8] is based on the correlations in the time series patterns, a recent study on daily deal sites [17] (e.g., Groupon and Livesocial) showed that the number of reviews (measured at Yelp) have increased significantly after daily deals, which obviously created a burst in the number of reviews but not caused by review spam. Therefore, the behavior of suspicious spammers need to careful studied to construct more accurate indicators.

IV. New Opportunities

A. Associate review credibility with verifiable actions

For product reviews, credible comments should be based on the experience with the product. Therefore, a strong tie between a review and the purchase action would reinforce the credibility. Based on this consideration, Amazon provides *Verified Purchase Review* service to confirm the reviewer has purchased the reviewed product through Amazon and explicitly show this linkage to help evaluating the review credibility. Similar feature has also been provided at third-party review sites, such as the "certified review" tag at demandforce.com attached to reviews where the reviewer's (most recent) visit to the particular store is verified. Such linkage has a convincing force to indicate which reviews are more creditable than others, and thus should be incorporated into the learning model. It is also worth noting that this does not mean reviews without verification are less credible since reviews can obtain experiences from other channels than what have been used for verification. Moreover, the trustworthiness introduced by verifiable linkage should not be considered arbitrary, since adversaries may be willing to insert fraudulent reviews at reasonable cost including real purchases, e.g., John Locke's case of manipulating of book reviews by paying for book purchase with unreliable positive reviews [1].

B. Associate review credibility with location information

As smart phones and other mobile handhold devices proliferate, many review sites provide applications and services to enable mobile access, especially for location-aware review searching. Location-enabled mobile device with the permission to access data about its location can discover and communicate realtime location information to remote servers and get relevant information such as map or nearby business rating back to the clients. To avoid fake location information provided by users with an incentive to cheat, various location proofing mechanisms such as trusted geotagging [18] or location-based authentication schemes [19], [20] have been proposed to prove the current and past locations of mobile devices. Another source for location verification is location-based social networking services, such as Foursquare or GetGlue that allow users to "check in" to a physical place of consumption with incentives. Some review sites provide an integrated review service with "social check-in", e.g., Yelp Check-ins, to broadcast reviews across online social networks.

Though no research has been observed from this aspect, we believe that the location information in-turn should also be utilized to assess the credibility of reviews submitted on-site. If the location information is trusted and within geographical proximity of the store being reviewed, the

[1]The New York Times report on "The Best Book Reviews Money Can Buy".

review is more likely to be selected into the ground truth data set. One concern is that the volume of mobile submitted reviews is considerably smaller than Web submitted reviews. This could be tackled by adopting semi-supervised learning that trains the discriminator with a small set of labeled data and a large set of unlabeled data.

C. Associate review credibility with social relationships

It is widely recognized that social influence plays a critical role in shaping people's decision making in product marketing. As such, recent researches on social network based recommendation systems [21]–[24] have already included social relations in the model to learn the correlations in preferences among friends and measure the influence from immediate and distant friends to the target's future decision.

Such explicit social relations among reviewers should also be taken into consideration in spam detection. Intuitively, the credibility of the reviews is closely related to the credibility of the reviewers. A review has more influential power to impact a reader's decision if it is written by a friend or a social friend or even a friend-of-friend. We believe such social relations can be utilized in a different direction to measure the credibility of one's opinion to his immediate and distant friends. In turn, we can reach an enriched credibility model for the reviews that reflects different levels of trust towards individual reviews based on the distance between the reviewer and the reader in the social relationship graph. A practical challenge along this direction is that the dataset could be extremely sparse since each reviewer only reviews limited number of products. The limitation of the dataset could be addressed by actively integrating the review social networks with common social networks (e.g., the merging of Google Local Review and Google+) and other online information sources (e.g., personal Websites or online blogs).

V. Conclusion

This paper discusses the issues, challenges, and opportunities of online review spam detection. We overview the existing detection methods based on supervised machine learning and data mining. A few challenges on reliable data acquisition and behavior correlation have been analyzed along with a discussion on the new opportunities from integrating current review systems with new applications/services to increase the credibility of reviews to a more comprehensive behavior analysis.

VI. Acknowledgement

This work is partially supported by KU NFGRF 2302283.

References

[1] "The company behind the brand: In reputation we trust," Weber Shandwick's online survey, 2012.

[2] "2011 Cone online influence trend tracker," http://www.coneinc.com/2011coneonlineinfluencetrendtracker, 2011.

[3] Y. Liu and Y. L. Sun, "Anomaly detection in feedback-based reputation systems through temporal and correlation analysis," in *Proceedings of the 2010 IEEE Second International Conference on Social Computing*.

[4] N. Jindal, B. Liu, and E.-P. Lim, "Finding unusual review patterns using unexpected rules," in *Proceedings of the 19th ACM international conference on Information and knowledge management*.

[5] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding deceptive opinion spam by any stretch of the imagination," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*.

[6] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Review graph based online store review spammer detection," in *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*.

[7] ——, "Identify online store review spammers via social review graph," *ACM Trans. Intell. Syst. Technol.*

[8] S. Xie, G. Wang, S. Lin, and P. S. Yu, "Review spam detection via time series pattern discovery," in *Proceedings of the 21st international conference companion on World Wide Web*.

[9] E. Gilbert and K. Karahalios, "Understanding deja reviewers," in *Proceedings of the 2010 ACM conference on Computer supported cooperative work*.

[10] N. Jindal and B. Liu, "Review spam detection," in *Proceedings of the 16th international conference on World Wide Web*.

[11] A. Mukherjee, B. Liu, J. Wang, N. Glance, and N. Jindal, "Detecting group review spam," in *Proceedings of the 20th international conference companion on World wide web*.

[12] Y. Liu, Y. Yang, and Y. Sun, "Detection of collusion behaviors in online reputation systems," in *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, oct. 2008, pp. 1368 –1372.

[13] M. Ott, C. Cardie, and J. Hancock, "Estimating the prevalence of deception in online review communities," in *Proceedings of the 21st international conference on World Wide Web*.

[14] N. Jindal and B. Liu, "Opinion spam and analysis," in *Proceedings of the international conference on Web search and web data mining*.

[15] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting product review spammers using rating behaviors," in *Proceedings of the 19th ACM international conference on Information and knowledge management*.

[16] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," in *Proceedings of the 21st international conference on World Wide Web*.

[17] J. W. Byers, M. Mitzenmacher, and G. Zervas, "Daily deals: prediction, social diffusion, and reputational ramifications," in *Proceedings of the fifth ACM international conference on Web search and data mining*.

[18] V. Lenders, E. Koukoumidis, P. Zhang, and M. Martonosi, "Location-based trust for mobile user-generated content: applications, challenges and implementations," in *Proceedings of the 9th workshop on Mobile computing systems and applications*.

[19] Y. Zhang, Z. Li, and W. Trappe, "Evaluation of localization attacks on power-modulated challenge–response systems," *Trans. Info. For. Sec.*

[20] Z. Zhu and G. Cao, "Applaus: A privacy-preserving location proof updating system for location-based services," in *INFOCOM, 2011 Proceedings IEEE*, april 2011, pp. 1889 –1897.

[21] J. He, "A social network-based recommender system," Ph.D. dissertation, 2010.

[22] I. Guy, N. Zwerdling, I. Ronen, D. Carmel, and E. Uziel, "Social media recommendation based on people and tags," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*.

[23] Y. Xie, Y. Cheng, D. Honbo, K. Zhang, A. Agrawal, and A. Choudhary, "Crowdsourcing recommendations from social sentiment," in *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*.

[24] C. Hu, C. Zhang, T. Wang, and Q. Li, "An adaptive recommendation system in social media," in *System Science (HICSS), 2012 45th Hawaii International Conference on*, jan. 2012, pp. 1759 –1767.