An Approach for Detecting Fake Reviewers in Social Media

Dr. Nivas Mohideen Jinna
Assistant Professor
College of Administrative & Financial Sciences, AMA International University, Salmabad, Kingdom of Bahrain

Abstract

As the majority of the general population require audit about an item before spending their cash on the item. So individuals run over different audits in the site however these surveys are bonafide or counterfeit isn't distinguished by the client. In some survey sites some great audits are included by the item organization individuals itself with the end goal to make with the end goal to deliver false positive item surveys. They give great surveys for some, extraordinary items produced by their very own firm. Client won't have the capacity to see if the survey is veritable or counterfeit. In this paper the followings are talk about, to discover counterfeit reviews(Spam) made by posting counterfeit remarks about an item by distinguishing the Rate Filter, User Filter, IP address Filter alongside survey posting designs. To discover the audit is phony or real, It will discover the IP address of the client if the framework watch counterfeit survey send by a similar IP Address numerous multiple times. This framework causes the client to discover remedy audit of the item.

Keywords: IP Address, Spam

I. INTRODUCTION

Item and administration audits assume an imperative job in settling on buy choices. In current occasions, when we are looked with numerous decisions, the supposition based audits enable us to limit the alternatives and settle on choices dependent on our necessities. This is particularly obvious on the web, where the audits are effortlessly open. A few organizations where audit based choices are exceptionally conspicuous are Amazon, TripAdviser, Yelp, and AirBnB, to give some examples. From a business perspective, positive surveys can result in huge budgetary advantages. This likewise gives chances to trickiness, where counterfeit audits can be created to collect positive assessment about an item, or to offensiveness some business. To guarantee validity of the surveys posted on a stage, it is critical to utilize a solid distinguishing model. In this paper, a few techniques for recognizing counterfeit audits are examined. The models talked about here fall into three classifications: Rate based classification, User based classification, and IP based classification.

A. User Based Classification

The user-based model asserts that a spamming user displays an abnormal behavior, and it is possible to classify users as spammers and non-spammers. The user information can be extracted from their public profiles. The relevant features include:

1) Content Matching: Spammers, often write their reviews with same template and they prefer not to waste their time to write an original review. In result, they have similar reviews.
2) Burstiness Calculation: Spammers, usually write their spam reviews in short period of time for two reasons: first, because they want to impact readers and other users, and second because they are temporal users, they have to write as much as reviews they can in short time.
3) Negative Ratio: Spammers tend to write reviews which defame businesses which are competitor with the ones they have contract with, this can be done with destructive reviews, or with rating those businesses with low scores. Hence, ratio of their scores tends to be low.

A standard learning algorithm, such as SVM or Random Forests, on these features can create a classification model for fake reviewers and non-fake reviewers.

Other than these important features, there are some other features that can be extracted from the user's profile, which can be used in detecting fake reviews.

- Number of reviews: A spammer is likely to create a lot of reviews, and this can be used to identify fake reviewers. Most of the users create not more than 1 review per day.
- Average review length: As mentioned earlier, a spammer is not going to invest much time in creating his reviews (especially when you are being paid by number of the reviews you write) and is more likely to create shorter reviews.
- Number of positive votes: Most of the fake reviews tend to be extremely positive. A high percent of strong positive votes indicated abnormal behavior. Non-fake reviewers have varying rating levels.

- Geographical Information: A user who is reviewing location-based products (for example, businesses on Yelp) at two or more locations in a day is surely exhibiting suspicious behavior. The credit card companies use this kind of information to track down scams.
- Activity: On social sites (for example, Yelp, Foursquare, and more), the account activity can also be an indicator of abnormal behavior. Users with a friend base and who post share check-ins on Facebook and Twitter are mostly genuine. In fact, linking your other accounts is a positive indicator.
- Useful votes: Yelp also allows its users to vote on a review, and the number of people of 'useful' votes for a review can also be used to classify spammers and non-spammers.

B. IP Based Classification

This method is used to find out fake reviews made by posting fake comments about a product by identifying the IP address along with review posting patterns. To find out the review is fake or genuine, system will find out the IP address of the user if the system observe fake review send by the same IP Address many a times. This system helps the user to find out correct review of the product.

C. Rate Based Classification

This approach to classify fake and non-fake reviews is very similar to the ideas used in spam classification.

- Time Frame: Spammers try to write their reviews asap, in order to keep their review in the top reviews which other users visit them sooner.
- Rate Deviation: Spammers, also tend to promote businesses they have contract with, so they rate these businesses with high scores. In result, there is high diversity in their given scores to different businesses which is the reason they have high variance and deviation.

By creating the linguistic n-gram features and using a supervised learning algorithm such as Naive Bayes or SVM, one can construct the classification model. This approach, of course, relies on the assumption that the fake and non-fake reviews consist of words with significantly different frequencies. In case the spammers had a little knowledge of the product, or they didn't have a genuine interest in writing the reviews (for example, the cheaply paid spammers), there are more chances of them creating reviews linguistically

- Ratio of Exclamation '!': First, studies show that spammers use second personal pronouns much more than first personal pronouns. In addition, spammers put '!' in their sentences as much as they can to increase impression on users and highlight their reviews among other ones.

We don't have any reason to believe that the spammer won't be careful enough to create reviews linguistically similar to the genuine ones, or have strong inclinations to write fake opinions. In that case, the pure text-based models won't be successful. We will need to incorporate more information.

Other than these important features, there are some other features that can be extracted from the rate of products, which can be used in detecting fake reviews.

- Length of the review: Even if a spammer tried to use words similar to real reviews, he probably didn't spend much time in writing the review. Thus, length of the fake-review is smaller than the other reviews of the same product. Lack of domain knowledge also increases the chances of a shorter review. Also, it could have happened that the spammer tried to overdo his job and wrote a longer review.
- Deviation from the average rating: There is a high probability for the spamming review to deviate from the general consensus rating for the product or the service.

II. LITERATURE SURVEY

In the last decade, a great number of research studies focus on the problem of spotting spammers and spam reviews. However, since the problem is non-trivial and challenging, it remains far from fully solved. I can summarize our discussion about previous studies in following categories.

A. Linguistic-based Methods

This approach extract linguistic-based features to find spam reviews. Feng et al. use unigram, bigram and their composition. Other studies use other features like pairwise features (features between two reviews; e.g. content similarity), percentage of CAPITAL words in a reviews for finding spam reviews. Lai et al. in use a probabilistic language modeling to spot spam. This study demonstrates that 2% of reviews written on business websites are actually spam.

B. Behavior-based Methods

Approaches in this group almost use reviews metadata to extract features; those which are normal pattern of a reviewer behaviors. Feng et al. in focus on distribution of spammers rating on different products and traces them. In Jindal et. al extract 36 behavioral features and use a supervised method to find spammers on Amazon and indicates behavioral features show spammers' identity better than linguistic ones. Xue et al. in use rate deviation of a specific user and use a trust-aware model to find the relationship

between users for calculating final spamicity score. Minnich et al. in use temporal and location features of users to find unusual behavior of spammers. Li et al. in use some basic features (e.g polarity of reviews) and then run a HNC (Heterogeneous Network Classifier) to find final labels on Dianpings dataset. Mukherjee et al. in almost engage behavioral features like rate deviation, extremity and etc. Xie et al. in also use a temporal pattern (time window) to find singleton reviews (reviews written just once) on Amazon. Luca et al. in use behavioral features to show increasing competition between companies leads to very large expansion of spam reviews on products. Crawford et al. in indicates using different classification approach need different number of features to attain desired performance and propose approaches which use fewer features to attain that performance and hence recommend to improve their performance while they use fewer features which leads them to have better complexity. With this perspective our framework is arguable. This study shows using different approaches in classification yield different performance in terms of different metrics.

C. *Graph-based Methods*

Studies in this group aim to make a graph between users, reviews and items and use connections in the graph and also some network-based algorithms to rank or label reviews (as spam or genuine) and users (as spammer or honest). Akoglu et al. in use a network-based algorithm known as LBP (Loopy Belief Propagation) in linearly scalable iterations related to number of edges to find final probabilities for different components in network. Fei et al. in also use same algorithm (LBP), and utilize burstiness of each review to find spammers and spam reviews on Amazon. Li et al. in build a graph of users, reviews, users IP and indicates users with same IP have same labels, for example if a user with multiple different account and same IP writes some reviews, they are supposed to have same label. Wang et al. in also create a network of users, reviews and items and use basic assumptions (for example a reviewer is more trustworthy if he/she writes more honest reviews) and label reviews.Wahyuni in proposes a hybrid method for spam detection using an algorithm called ICF++ which is an extension to ICF of in which just review rating are used to find spam detection. This work use also sentiment analysis to achieve better accuracy in particular. Deeper analysis on literature show that behavioral features work better than linguistic ones in term of accuracy they yield. There is a good explanation for that; in general, spammers tend to hide their identity for security reasons. Therefore they are hardly recognized by reviews they write about products, but their behavior is still unusual, no matter what language they are writing. In result, researchers combined both feature types to increase accuracy of spam detection. The fact that adding each feature is a time consuming process, this is where feature importance is useful. Based on our knowledge, there is no previous method which engage importance of features in the classification step. By using these weights, on one hand I involve features importance in calculating final labels and hence accuracy of spam detection increase, gradually. On the other hand we can determine which feature can provide better performance in term of their involvement in connecting spam reviews (in proposed network).

III. PROPOSED WORK

This framework will discover counterfeit surveys made by the internet based life streamlining group by recognizing the IP address. Client will login to the framework utilizing his client id and secret key and will see different items and will give audit about the item. To discover the audit is phony or veritable, framework will discover the IP address of the client if the framework watch counterfeit survey send by a similar IP Address numerous on occasion it will illuminate the administrator to expel that audit from the framework. This framework utilizes information mining strategy. This framework encourages the client to discover remedy audit of the item.

A. *Advantages*

- User gets genuine reviews about the product.
- User can post their own review about the product.
- User can spend money on valuable products.

IV. COMPARATIVE ANALYSIS AND SUGGESTIONS

When building up another audit spam location structure, it is vital to comprehend what methodologies and procedures have been utilized in earlier examinations. In past areas, I introduced a diagram of machine learning procedures that have been utilized in the survey spam space and a portion of the imperative aftereffects of these examinations. As this space is youthful, moderately few examinations on machine learning procedures and survey spam location have been led.

In light of our overview, the vast majority of the past investigations have concentrated on managed learning systems. Be that as it may, with the end goal to utilize managed learning, one must have a named dataset, which can be troublesome (if certainly feasible) to gain in the territory of survey spam. From the writing talked about, it very well may be seen that the greater part of the accessible datasets utilized in the past examinations are artificially made, probably because of the absence of audit spam models and the trouble of naming them. Building and assessing classifiers dependent on these engineered datasets can be risky, as it has been seen that they are not really illustrative of true audit spam. For instance, when utilizing a similar system to assess the fake AMT dataset utilized in and Yelp's sifted audits dataset, the extricated highlights and results varied enormously, particularly when utilizing n-gram content highlights. Contrasting order execution over these datasets demonstrates that when assessed on the

engineered audit dataset, the classifier accomplished an exactness of 87%, however while utilizing Yelp's surveys just accomplished 65 % precision. This 22% drop in precision infers that artificially made audits have distinctive highlights than genuine phony surveys, and that the audits created by AMT don't precisely reflect true spam surveys.

V. CONCLUSION

This investigation presents structure dependent on a metapath idea and another diagram based technique to mark surveys depending on a rank-based naming methodology. The execution of the proposed structure is assessed by utilizing two genuine marked datasets of Yelp and Amazon sites. Our perceptions demonstrate that figured weights by utilizing this metapath idea can be exceptionally compelling in recognizing spam audits and prompts a superior execution. Moreover, it is discovered that even without a train set, this system can compute the significance of each element and it yields better execution in the highlights' expansion procedure, and performs superior to past works, with just few highlights. In addition, in the wake of characterizing four principle classifications for highlights our perceptions demonstrate that the surveys social class performs superior to anything different classifications, as far as AP, AUC and also in the ascertained weights. The outcomes likewise affirm that utilizing diverse supervisions, like the semi-regulated strategy, have no recognizable impact on deciding a large portion of the weighted highlights, similarly as in various datasets. IP Address following gives all the more valuable and ideal outcome.

REFERENCES

[1] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Exploiting burstiness in reviews for review spammer detection. In ICWSM, 2013.
[2] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao. Spotting fake reviews via collective PU learning. In ICDM, 2014.
[3] L. Akoglu, R. Chandy, and C. Faloutsos. Opinion fraud detection in online reviews bynetwork effects. In ICWSM, 2013.
[4] S. Feng, R. Banerjee and Y. Choi. Syntactic stylometry for deception detection. Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers; ACL, 2012.
[5] G. Wang, S. Xie, B. Liu, and P. S. Yu. Review graph based online store review spammer detection. IEEE ICDM, 2011.
[6] C. L. Lai, K. Q. Xu, R. Lau, Y. Li, and L. Jing. Toward a Language Modeling Approach for Consumer Review Spam Detection. In Proceedings of the 7th international conference on e-Business Engineering. 2011. [34] N. Jindal and B. Liu. Opinion Spam and Analysis. In WSDM, 2008.
[7] Ott M, Choi Y, Cardie C, Hancock JT (2011) 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 (pp. 309–319). Association for Computational Linguistics.
[8] Morales A, Sun H, Yan X (2013) Synthetic review spamming and defense. In: Proceedings of the 22nd international conference on World Wide Web companion (pp. 155–156). International World Wide Web Conferences Steering Committee, Rio de Janeiro, Brazil.
[9] Mukherjee A, Venkataraman V, Liu B, Glance NS (2013) What yelp fake review filter might be doing? Boston, In ICWSM.