Advanced Mechanism for Finding Spammers in Social Media

Nivas Mohideen Jinna, PhD
Assistant Professor
CAFS, AMA International University, Salmabad, Kingdom of Bahrain

Abhishek Lal
MCA Student
KVM CE & IT, Cherthala
Kerala, India

Aswin Raju
MCA Student
KVM CE & IT, Cherthala
Kerala, India

ABSTRACT
As most of the people require review about a product before spending their money on the product. So people come across various reviews in the website but these reviews are genuine or fake is not identified by the user. In some review websites some good reviews are added by the product company people itself in order to make in order to produce false positive product reviews. They give good reviews for many different products manufactured by their own firm. User will not be able to find out whether the review is genuine or fake. In this paper we are discuss to find out fake reviews(Spam) made by posting fake comments about a product by identifying the Rate Filter, User Filter, IP address Filter along with review posting patterns. To find out the review is fake or genuine, we 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.

Keywords
IP Address, Spam.

1. INTRODUCTION
Product and service reviews play an important role in making purchase decisions. In current times, when we are faced with many choices, the opinion-based reviews help us narrow down the options and make decisions based on our needs. This is especially true online, where the reviews are easily accessible. Some companies where review-based decisions are very prominent are Amazon, TripAdviser, Yelp, and AirBnB, to name a few. From a business point of view, positive reviews can result in significant financial benefits. This also provides opportunities for deception, where fake reviews can be generated to garner positive opinion about a product, or to disrepute some business. To ensure credibility of the reviews posted on a platform, it is important to use a strong detecting model. In this paper, we'll talk about some methods for detecting fake reviews. The models discussed here fall into three categories: Rate Filter, User Filter, and IP Address Filter.

1.1 User Filter
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:

- **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.
- **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.
- **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.

1.2 IP Address Filter
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.

1.3 Rate Filter

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.

2. 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. We 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 we 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).

3. PROPOSED WORK

This system will find out fake reviews made by the social media optimization team by identifying the IP address. User will login to the system using his user id and password and will view various products and will give review about the product. 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 at times it will inform the admin to remove that review from the system. This system uses data mining methodology. This system helps the user to find out correct review of the product.

Spam Feature	**User Filter**	**Rate Filter**	**IP Address Filter**
Behavioral based Features	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. $x_{BST}(i) = \begin{cases} 0 & (L_i - F_i) \notin (0,\tau) \\ 1 & \frac{L_i - F_i}{\tau} \quad (L_i - F_i) \in (0,\tau) \end{cases}$ where Li − Fi describes days between last and first review for τ = 28. Users with calculated value greater than 0.5 take value 1 and others take 0. 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. Users with average rate equal to 2 or 1 take 1 and others take 0.	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. $x_{ETF}(i) = \begin{cases} 0 & (T_i - F_i) \notin (0,\delta) \\ 1 & \frac{T_i - F_i}{\delta} \quad (T_i - F_i) \in (0,\delta) \end{cases}$ where L i − Fi denotes days specified written review and first written review for a specific business. We have also δ = 7. Users with calculated value greater than 0.5 takes value 1 and others take 0. 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. $x_{DEV}(i) = \begin{cases} 0 & otherwise \\ 1 & \frac{r_{ij} - avg_{r \in E_{ij}} r(e)}{4} \geq \beta 1 \end{cases}$ where β1 is some threshold determined by recursive minimal entropy partitioning. Reviews are close to each other based on their calculated value, take same values (in [0,1)).	In IP Address filter, Find the IP Address of reviewer and find the interval between each review time from same IP Address. If the time is < 15 takes value as 1 And other 0
Linguistic based Features	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. Users have close calculated values take same values (in [0,1)).	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. Reviews are close to each other based on their calculated value, take same values (in [0,1)).	

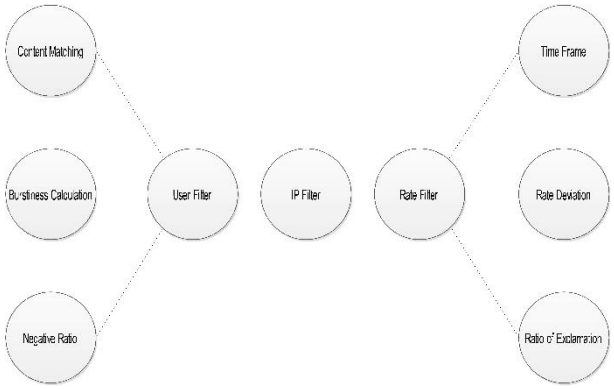

Fig. Architecture Diagram

ADVANTAGES

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

4. COMPARITIVE ANALYSIS AND SUGGESTIONS

When developing a new review spam detection framework, it is important to understand what approaches and techniques have been used in prior studies. In previous sections, we presented an overview of machine learning techniques that have been used in the review spam domain and some of the important results of these studies. As this domain is young, relatively few studies on machine learning techniques and review spam detection have been conducted.

Based on our survey, most of the previous studies have focused on supervised learning techniques. However, in order to use supervised learning, one must have a labeled dataset, which can be difficult (if not impossible) to acquire in the area of review spam. From the literature we discussed, it can be observed that most of the available datasets used in the previous studies are synthetically created, most likely due to the lack of review spam examples and the difficulty of labeling them. Building and evaluating classifiers based on these synthetic datasets can be problematic, as it has been observed that they are not necessarily representative of real world review spam. For example, when using the same framework to evaluate the artificial AMT dataset used in and Yelp's filtered reviews dataset, the extracted features and results differed greatly, especially when using n-gram text features. Comparing classification performance across these datasets shows that when evaluated on the synthetic review dataset, the classifier achieved an accuracy of 87 %, but while using Yelp's reviews only achieved 65 % accuracy. This 22 % drop in accuracy implies that synthetically created reviews have different distinguishing features than real-life fake reviews, and that the reviews produced by AMT do not accurately reflect real world spam reviews.

5. CONCLUSION

This study introduces framework based on a metapath concept as well as a new graph-based method to label reviews relying on a rank-based labeling approach. The performance of the proposed framework is evaluated by using two real-world labeled datasets of Yelp and Amazon websites. Our observations show that calculated weights by using this metapath concept can be very effective in identifying spam reviews and leads to a better performance. In addition, we found that even without a train set, this framework can calculate the importance of each feature and it yields better performance in the features' addition process, and performs better than previous works, with only a small number of features. Moreover, after defining four main categories for features our observations show that the reviews behavioural category performs better than other categories, in terms of AP, AUC as well as in the calculated weights. The results also confirm that using different supervisions, similar to the semi-supervised method, have no noticeable effect on determining most of the weighted features, just as in different datasets.

IP Address tracking gives more precious and optimal result.

6. 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] [34]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] [35]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.