Detection of Spam Content in Web Pages Survey

A. Malathi
Assistant Professor, PG & Research Department of Computer Science, Government Arts College (Autonomous), Coimbatore, Tamil Nadu, India

B. Amutha
M.Phil. Research Scholar, Department of Computer Science, Government Arts College (Autonomous), Coimbatore Tamil Nadu, India

Abstract – Now a day, a big part of human beings depends upon available content material in social media in their decision (e.g.: reviews and comments on a subject or product). Online Social Networks (OSNs) comprises two users namely, Spammers and Non-Spammers. Spammers comment unwanted information about a product on Online Social Networks. Writing fake reviews will promote or demote some particular products. When find the spammer, customer can select the good products. In this paper study is on problem of spam detection in Online Social web pages we are going to implement the new algorithm technique is hybrid buying sequence and spam detection using in our proposed system. This hbssd method is use to analyze the customer buying behavior and opinion details, then detect the spam content.

Index Terms – Network Spam Detection, Fake Spammer, Spam Review, OSNs.

1. INTRODUCTION

The web has greatly enhanced the way people perform certain activities [e.g. shopping], find information and interact with others**.** More people or customer requires about before product reviews have spending the money on the particular selected product also. Today many people read/write reviews on merchant sites, blogs, forums, and social media before/after them purchase or services. More and more users prefer would make their purchase decisions based on these online review. Products (with a large percentage of positive reviews tend to attract more customers than products without large percentage) of positive reviews based on the reason for the profit or fame or proud. Imposters have tried to cheat the online review to deliberately mislead potential customers. Positive opinions often mean profits and fames for business and individuals. Negative reviews products in order to damage their reputations. This product reviews are true or fake, this is not identified by the customer or user. This survey study is on analysis the fake comments by spammers.

2. REVIEW OF LITERATURE

This paper [1] author has presented the spam detection Net Spam based on a metapath concept, this concept is same as a new graph-based method to label reviews but it is not based on a rank-based approach. In this paper main concept of Net spam based on a metapath is included by a novel spam detection framework and this framework is evaluated by using two real-world labeled datasets of Yelp and Amazon websites. This concept is to view that calculated weights by using this metapath concept. This concept is very effective in identifying spam reviews and, then leads to get a better performance and it includes additional feature that even without a train set, and the Net Spam is calculates the importance of each feature and it yields better performance in the features' addition process. That is its performance better than existing system, (previous work), with only a minimum number of features.

Authors in [2] presented Language Modeling Approach for Consumer Review Spam Detection. Propose a novel probabilistic language model to estimate the similarity between any pairs of reviews in terms of the likelihood of a review generating the contents of another review. This approach based technique is applied in the experiment and it shows the results by using a large data set. This data set shows that the probabilistic language modeling based computational model. It is very effective for the detection of untruthful review.

In this paper [3] authors proposed that, new holistic framework called SpEagle that exploits both relational data (user–review–product graph) and metadata (behavioral and text data) collectively, so this collections of data used to detect the suspicious users and reviews its same as products targeted by spam.

In this paper [4] author has presented a fast and effective framework called "FRAUDEAGLE". This framework is used to spotting fraudsters and then, fake reviews in online review datasets is carried out within second. This methods is give more advantages are: (i) it is used to network effect among reviewers and products, but the previous system is only focus on review text or behavioral analysis, (ii) it consists of two main steps (a) scoring the users and reviews for the fraud detection (b) grouping for the visualization and sensemaking (iii) it is operated only for unsupervised fashion, but it requiring no labeled data, while still it incorporating side information available (iv) it is a scalable large data sets and its run time generating sequentially with network size. This framework is very effective on creating or generating and a real datasets, so the FRAUDEAGLE is successfully directly fraud-bots in a large online application review database.

In this paper [5] authors introduce that, Collective PU learning framework for fake review detection. This paper introduced concept based proposed system is to provide a supervised learning algorithm MHCC (Multi-typed Heterogeneous Collective Classification). In this algorithm is used for the heterogeneous network of reviews, users and IPs and then extended it to CPU model. This model is more appropriate for the PU learning problem. This problem come with reason are the labels of reviews is very high precision but unknown recall, and then this concept with the labeled data provided by review the hosting website. The hosting website Dianping is conducted more experiments and to show that combining the collective classification. Another result produce the PU learning, the proposed system that is, CPU model includes more importance advantages the previous system like, state-of-the-art baseline algorithms. It is not performs outside, but also more importantly, and then it detects a large number of potential fake reviews hidden in the unlabeled set. It shows that the power of PU learning in solving the problem.

Authors in [6] propose to exploit bursts in detecting opinion spammers due to the similar nature of reviewers in a burst. A graph propagation method for identifying spammers was presented. A supervised learning is explains the difficult problem. This problem of evaluation is without real data or ground truth data. It which classifies and reviews is based on a different set of features from identifying spammers. This supervised learning is based on a novel evaluation method. In this proposed system show that the experimental results, but this experiment is use the Amazon.com that is, Amazon oriented reviews from the software domain. This domain shows the proposed systems effective, but which is not only solved its effectiveness objectively based on the supervised learning or classification. But it is also subjectively based on human expert evaluation. The real fact of that supervised learning/classification experimental results are consistent with human judgment or a decision also indicates the proposed system concept supervised learning, which is based on evaluation technique or method is justified.

Authors in [7] propose a graph based review spammer detection model this approach heterogeneous review graph to capture the relationships among reviewers, reviews and stores that the reviewers have reviewed. In this graph is reveals with the reason of spam and the iterative model, it is to identify the suspicious reviewers. This graph is explores that analyze how interactions between nodes, that is interaction between two or more nodes. This concept based experimental result is shows that this model is to find out the spamming activities with more accurately.

Authors in [8] propose collaborative setting model to discover fake reviewer groups. In this model initially uses a frequent itemset mining method. This method is to find a set of candidate groups. It was use the more behavioral models, and then this model is derived from the collusion phenomenon among fake reviewers. In this system use the relation models and this model based on the relationships among groups, and then, individual reviewers. This model is additionally the products they reviewed to detect fake reviewer groups.

Table 1.0. Comparison table

Paper No	Technique	Advantages	Disadvantage
1	Novel Spam Detection Framework	very effective in identifying spam reviews and leads to a better Performance.	This approach cannot apply for multilayer network.
2	Probabilistic language model.	model is effective for the detection of untruthful Reviews.	It does not always offer the speed. More difficulties apply this model.
3	SpEagle	predict suspicious users and reviews more accurately than other traditional method	Needs more parameters and deep study is necessary, so that can't able to apply big data set.
4	FRAUDEAGLE	This model automatically detect fraudulent users and fake reviews in online review	Making each link is associated with the review rating is time consuming process.
5.	PU learning framework	More importantly, detects a large number of potential fake reviews hidden in the	our models only use language dependent features, cannot can be generalized

		unlabeled set.	to any other languages
6.	Loopy Belief Propagation and Markov Random Field.	Visualize and spammers detection more accurate.	Need Effective classification algorithm.
7.	Heterogeneous review graph Model.	This model find Spamming activities with more accurately.	This model is expensive, both in terms of memory and compute time.
8	GSRank.	Effectively detect group spammers in product reviews	labeling individual fake reviews or reviewers is hard

3. DISCUSSION

A recently proposed all technique which have mentioned in comparison table not effectively find the online spam it has some limitations. However, the methods GSRank can only perform effective detect group spammers in product reviews compare other Technique. Some of the ranking algorithm such as spam rank, In-Degree Ranking, Page Rank etc. spam rank this filter out the fraudulent reviews based on similarity of review content. In-Degree Ranking algorithm ranks the post review according to popularity of the review page. Page ranking algorithm has probability distribution someone is post review randomly on the links that will group based on probability specific page. Every ranking algorithm proposed in results it's an ineffective. This concludes GSRank is a effective method to detect and group spammers so proposed GSRank consider our research work.

4. PROPOSED METHODOLOGIES

The existing framework requires more training dataset and time. This can be computationally demanding for large data sets. The proposed system provides a framework called Hybrid Buying and Sequence Spam Detection (H-BSSD), to effectively detect Spam Detection in Social Network. The proposed system aims at increasing the detection performance through the high dimensional data set.

In this proposed system we should give the dataset first then hissed algorithm preprocessing the dataset details, and analyze the buyer behaving analysis with the user product buying details with their opinion about the product. Using the user comment the sequential analysis process will clustering the details and finding the accuracy of the result.

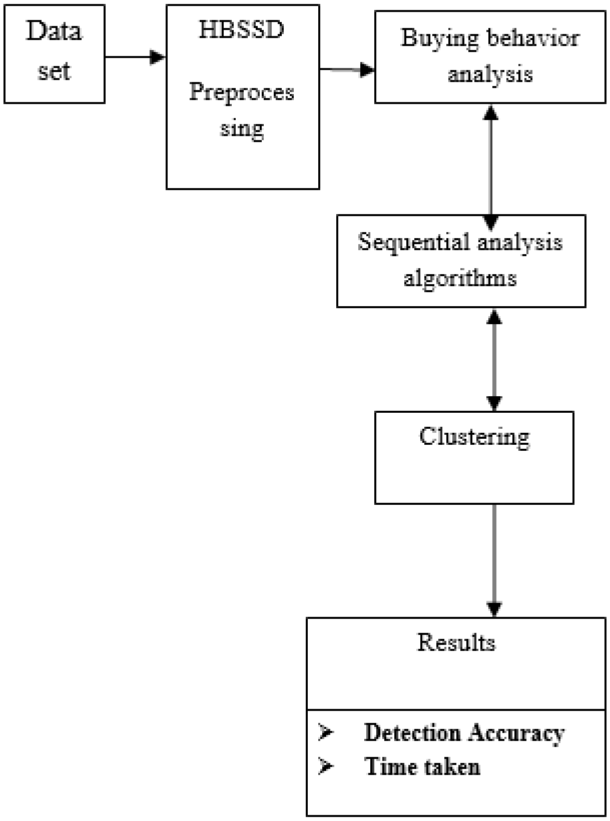

Figure 1.0 steps of H-BSSD

5. CONCLUSION

More number of people see or visit the various reviews in the website but these reviews are genuine or fake is not identified by user is the real fact. Sometimes the review websites is good reviews or best reviews so added by the product company people or gold person of my website itself in the order to make, but this order is to produce false that is negative product reviews. In this paper, the problem of finding Spam Detection in Online Social Network techniques is investigated. There are numerous researches from various domains are continuously working towards developing Spam Detection in Online Social Network. The aim of this survey was to summarize the recent researches and its demerits in Spam Detection. This paper gives the merits and demerits of the recent techniques and its capabilities are studied. This paper concludes that there is no effective prediction method not applies and concentrates on the Spam Detection with high accuracy. So, our proposed approaches should overcome all the above issues. Proposed Hybrid Buying and Sequence Spam Detection (H-BSSD) implementation will be to be done in order to Spam Detection with more accurate in an Online Social Network.

REFERENCES

[1] Shehnepoor, Saeedreza, et al. "NetSpam: a network-based spam detection framework for reviews in online social media." IEEE Transactions on Information Forensics and Security 12.7 (2017): 1585-1595.

[2] Lai, C. L., et al. "Toward a language modeling approach for consumer review spam detection." *e-Business Engineering (ICEBE), 2010 IEEE 7th International Conference on*. IEEE, 2010.

[3] Rayana, Shebuti, and Leman Akoglu. "Collective opinion spam detection: Bridging review networks and metadata." Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining. ACM, 2015.

[4] Akoglu, Leman, Rishi Chandy, and Christos Faloutsos. "Opinion Fraud Detection in Online Reviews by Network Effects." ICWSM 13 (2013): 2-11.

[5] Li, Huayi, et al. "Spotting fake reviews via collective positive-unlabeled learning." Data Mining (ICDM), 2014 IEEE International Conference on. IEEE, 2014.

[6] Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., & Ghosh, R. (2013). Exploiting Burstiness in Reviews for Review Spammer Detection. Icwsm, 13, 175-184.

[7] Wang, Guan, et al. "Review graph based online store review spammer detection." Data mining (icdm), 2011 ieee 11th international conference on. IEEE, 2011.

[8] Mukherjee, Arjun, Bing Liu, and Natalie Glance. "Spotting fake reviewer groups in consumer reviews." Proceedings of the 21st international conference on World Wide Web. ACM, 2012.