Survey on spam review detection using review mining and machine learning on social media

Padmini C
Department of Computer Science & Engg
Presidency University,Bangalore
padmini.c@presidencyuniversity.in

Dr. Mary Cherian
Professor,Department of Computer Science & Engg
Dr.Ambedkar Institute of Technology,Bangalore
thamasha2005@yahoo.c

Abstract— In the existing era, online social networks are the most popular and rapidly growing media on the Internet. Individuals of all gatherings invest the vast amount of their time on social communication sites like Facebook, Twitter, WhatsApp, LinkedIn, etc. for reading news, discussing events, sharing their views and posting messages about products & services. Users spend most of the time on these famous social networking websites. These interests have rouse thrust to illegitimate customers, this attracted the spammer who engages in fraudulent activity in opposition to social network users. Fraudulent activity is taken into consideration to cause greater harm than any other form of cyber-crime. This violation must be ratified beforehand than any individual is informed of the fake profile creation. Different Algorithm and techniques have been proposed for the finding of spam reviews, spam profiles, spam emails, and spam messages. In this investigation, we have done wide research on existing strategies for spam review identification.
Keywords—social network, Spam, Spammer, cyber-crime

I. Introduction

As the Internet keeps on developing in its size and significance, the amount and impact of online surveys continually increased. Reviews can affect a solitary individual to a sweeping scope of enormous endeavors, yet they are critical in the area of online business and administrations. Where reviews, comments, and opinions concerning items and services are regularly most helpful for a purchaser, to decide on a choice on whether to get them. Online surveys might be done for various reasons on social networking services (like e.g., Instagram, Facebook, Twitter, LinkedIn), e-commerce systems (e.g., eBay, flipkart, Amazon) review websites may request their clients to furnish opinion about the items, services or product they have purchased to know whether they are satisfied or not, with an objective to improve and upgrade their organizations. Online reviews are useful and it has dazed trust for both vendors and purchasers. Before placing an online request, people peek online reviews. Most of the time, any choice is subjective to online reviews or opinions must be made mindfully as these reviews may be faked for their advantage or expansion. There is a need that any decision reliant on online studies must be made carefully. A few enterprises encouraged business visionaries or spammers to create great surveys about their item or to form horrendous reviews about their opponent's things or organizations. Survey spam are those fake reviews that can have an incredible effect because of their significance on commercial business.

Spams are categorized into a different form such as-

- Email spam
- Message spam
- Cell phone spam
- Video spam
- Online game spam
- Social media spam

On all, the social networking spam has become very popular as the number of users using social network have increased exponentially.
In the past few years, a different strategy has been proposed to illuminate the issue of spam review. Here, we have done a far survey of existing research on spam and spammer discovery using various methodologies.
This paper is portioned into four sections, section II contains related work on finding feature selection on spam detection. Section III depicts the strategies utilized for spam survey discovery in social media networks. Section IV elaborates comparative study of their different approaches. Section V is the conclusion and future enhancement.

II. Existing research

The existing research work on spam review detection can be categorized into two different features

- First one is based on review
- Second is on reviewer characteristics

Review centric
Text mining is a technique utilized in the review analysis, which uses various methodologies like a Bag of words approach, Term reappearance, Linguistic Inquiry and Word Count (LIWC), Part of Speech (POS) and so on., where the event of single words or little gatherings of words and its recurrence are the highlights utilized. Most of the research has discovered that one of these methodologies is not adequate to prepare a classifier with satisfactory execution in review spam identification. Hence, there is a need for extra techniques to include designing (extraction) to

separate a progressively enlightening list of capabilities that will improve the exhibition of review spam recognition.

A. Bag of words approach

Bag of words approach is a methodology that utilizes single or little gatherings of words from the given content as its characteristic. These characteristics are called n-grams. It chooses words from a given succession to make bags of words, by taking one or a few adjoining words from a given text. These mean a unigram, bigram, and trigram (n = 1, 2, and 3) individually. These highlights had utilized to train the model. [3]

B. Term frequency

The Term frequency is used not to check only for the presence or absence of any specific word. But It also considers the frequencies of each word that occurred in a given textual review. The bag of words approach has additionally enhanced with the term-frequency approach. A dataset that utilizes the term frequencies is like that of the bag of words dataset. So, it includes the count of occurrences of a term in the review [4].

C. LIWC output and POS tag frequencies

Linguistic Inquiry and Word Count (LIWC) is a textual examination or investigation software tool. The customers use this tool to guide and develop their word dictionary to analyze the feature of the language of their particular interest and frequencies.

Parts of Speech (POS): It is a method that goes through the content and identifies grammatical features to each word as Noun, verb, adverb, adjective, preposition... It labels and highlights the words with a grammatical form dependent on the definition inside the sentence in which it has occurred, where this labelled highlight assists in accomplishing expected outcomes over the bags of words approach. The users use this tool to build their dictionaries and to analyze their dimensions of specific interests and frequencies [5].

D. Stylometric

It is an elective technique that uses the term recurrence strategy for feature extraction, which utilizes character-based and word-based lexical highlights or syntactic highlights. Lexical highlights utilize the sort of words and characters, exceptional characters, capitalized or lowercase characters, and the word length is the sign for spam identification. Model "a"," the "," @are". Where the syntactic component helps in speaking to the reviewer style of composing the survey [*6*].

E. Semantic

The semantic language model utilizes the fundamental importance or the ideas of the words used in their survey to distinguish spam reviews. It manages the basic idea or significance of those words to make semantic language models for recognizing untruthful reviews. The analysis is that changing of words should not impact the meaning of the sentence in the review. Like "I love you" to "I like you" wouldn't have any influence on the review, as they have the same meaning. [7].

Reviewer centric

The behavioral feature is another feature that helps in understanding the behavior of the spammer. It gives a different dimension in identifying spammer rather than only considering the spam reviews. This approach also has given efficiency. As there are many approaches in existing research to find the spammer by observing their behavior, but its not easy to find the spammer in a single review. So, there are few approaches to find spammer is like having created multiple users-Id for the same user, group of spammer writes more reviews for the same brand, or they are using the same email-Id and writing their fake views on different bands these behavioral footprints help us to understand the spammer better [2]. Another technique to understand the spammer's behavior is by their relation with other reviewers using a graph theory-based approach that has shown promising results. The behavioral features used in many researches is like the length of the review, date, and time they have written the reviews, the rating that they have given, a reviewer I'd, etc., by using some of the features like these they can identify the spammer and label all their reviews written by them as spam.

Some of the features may not be available for all the reviews of different sources and thus restricts to use these characteristics for detecting spam [8].

A. Maximum number of reviews

The spammers write more reviews on many products on any given day using the same or different email-id than compared to a genuine user, by taking this feature into account it helped in detecting spammer.

B. Percentage of positive reviews

When compared reviewer who gives positive feedback for most of the product on any given day this abnormal state of the positive review might be an indication of a spam review.

C. Review length

It is observed that the length of the review gives a significant sign for the indication of a spammer, genuine review length will not be more than 135 words. Whereas 92% of spammers reviews length will be more than 200 words.

D. Reviewer deviation

Reviews rating given by spammer will far veer off from the genuine reviewer this distinguishing behaviour of the spammer will help in identifying their reviews.

E. Maximum contents similarity

The similarity in the content of review across the website is a strong indication of a spammer. This is an advanced feature that gives strong support to the text analysis.

III. Techniques used to detect spam review in a social network

The analysts have made use of different methods on online social organizations to recognize spam, such as identifying the spammers, fake profile attribute, honey profiles, creation and analysis of message, investigation of contextual information, investigating their companion's relationship, or system structures and some more. These methodologies can be arranged into various strategies

- Honey profiles
- URL /black listing-based approach
- Clustering
- Supervised and unsupervised machine learning

approach

- Incremental learning

REF	SOCIAL NETWORK	TECHNIQUES	FEATURES USED	PROS	CONS
Webb et al.., 2008 [9]	Sinaweibo	Honey profile	Honey pot receives friend request, time stamp of the request received	This techniques found about 52.76% of fake profile	It didn't focus on later effect after spam friend request.
Lee et al.., 2011[10]	Twitter	Honey profile	Created 60 honey profile and sent random message to track follow-follower ratio	Developed expectation maximization algorithm for cluster analysis and detected spam very fast	It didn't work well with false positive and negative spam which required deep investigation
cao& cover lee. et al.., 2014 [11]	Facebook	URL/blacklist based techniques	Observed number of clicks on URL link and their intention	Feature based on click helped to attain the accuracy of 0.859	It covered only bitly shortened URL and ignoring other URL shortening series.
Alghamdi et al.., 2016 [12]	Facebook	URL/blacklist based techniques	URL and other social networking features like domain name, host name etc...	Identified two independent class like URL and social network feature gave a new way for spam detection	The research lacked in experimental evaluation to claim and strengthen their work
Zeng et al.., 2015[13]	Sinaweibo	Machine learning - ML	user based and content based	Developed prototypesoftware used SVMclassifiergavebetter performance accuracy of 95.7%	measuring ELM requires a trained classification model which is time consuming.
P.T Nguyen and Takeda et al..,[14]	Twitter	Algorithm using Online learning	Network feature, user -profile, activity and contentbased feature	Evaluating Performance of 16 online learning algorithm.	steadily adjust the learning model with each and every input and acclimate to the changing pattern of spammers over additional time
Song et al.., [15]	Youtube	Classification using sofia-mltoolbox from google	Word, topic and user based feature	Incremental model helped in achieving best performance accuracy of 91.7%	Considered only YouTube comments for target study
Ala'M et al.., 2017[16]	Twitter	Machine learning classification and statistical analysis	Suspicious words, default image ,text-to-link ratio, FFR, repeated words, tweet-time pattern, no of tweets	Proposed approach helped to catch populated search magnified to conduct spam detection which even google safe browsing failed to detect	Technique is highly dependent on the requirement of URL and domain name in the content
Zheng et al.., 2015[17]	Sinaweibo	Machine learning	Content based and user based	Developed prototype software used SVM classifier gave bet performance attaining 95.7% accuracy	The requirement for the use of the measure, for example, ELM as preparing grouping model was tedious
Chen et al.., 2017[18]	Twitter	Lfun(incremental learning)	12 features(Account, age, no of followers, hashtag, URLs	By using incremental learning in Lfun helped identifying auto labelled and human labelled tweets	Didn't use review of more than 2 days for their experiment and no old tweets were deleted
Shehnepoor et ,al.., 2017[19]	Yelp and Amazon	A Semi-supervised and unsupervised learning	Review-behavioral, review linguistic, user-behavioral review linguistic	labelling and ranking different behavioural feature helpedNetspamto handle spammer well	Applied only to reviewer website need to think to apply to another social website also.
Soliman and Girdzijauskas- et al.., 2017[20]	Twitter	A Graph based unsupervised approach learning	User based feature derived from content and graph based feature	Proposed novel graph based technique attained high accuracy 92.3%	social network is huge which makes it difficult to expand and gather network features
Isa Inuwa-Dutse et al.., (2018)[21]	Twitter	A Content-based, network-basedlearzning and Twitter specific memes	User Profile Features (UPF), Account Information Features (AIF), Pairwise engagement features	The blend of high-quality features and features learnt in a solo way essentially improved baseline conduct.	The qualification between authentic human clients versus real social bots just as human spammers versus social bot spammers should be researched further

IV. COMPARIBILTY OF DIFFERENT SPAM DETECTION TECHNIQUES

Techniques	Pros	cons
Honey Profile	Spammer are trapped in Real time. Help to dissect diverse real-time examples are followed to analyse spammers	Constrained arrangement of honey profiles could be sent because of uncontrolled weight and high upkeep cost Difficult to deploy on a large scale
URL/Blacklist	To disseminate the information of the spammer URL act highly efficient. Blacklist URL is published by Popular service providers	Time lag problem Detection after the click and spread
Classification	Methods assists with building up a computerized model for the well-suited grouping of spam substance and spammers	Obtaining and maintaining the required labelled data is very difficult Typically suffers from experience the ill effects of class issue problem
IL-Incremental Learning	Encompassing timely behavioural changes and dynamism continuous update in training data improves accuracy	Monitoring to remove oldest data and Requires continuous timely update of data Challenging to Dynamically label the data

CONCLUSION AND FUTURE ENHANCEMENT

We have done exhaustive research on existing techniques for spam review detection using various methodologies. In this paper, we have also discussed different features used in each machine learning technique and along with the comparative study. We observe that there is no such single technique and distinctive element that can give a clue for grouping of surveys as genuine or phony.
Another fascinating element for future work is to consider the impact of the ongoing increment in the word length of tweets on spamming action. Instinctively, computerizing spam revelation is troublesome on lengthier tweets.

References

1. Mukherjee A, Kumar A,Liu B,Wang J, Hsu M, Castellanos M, Ghosh R,"spotting opinion spammers using behavioural footprint", In: proceedings of the 19th international conference on knowledge discovery and Data mining-ACM SIGKDD,Chicago,2013.
2. Fei G, Mukherjee A, Liu B, Hsu M, Castellanos M, Ghosh R "Exploiting Burstiness in reviews for review spammer detection". In: international conference on web and social media-ICWSM,2013
3. Jindal N, Liu B, Lim EP, "Finding unusual review patterns using unexpected rules". In. Proceedings of the 19th international conference on Information and knowledge management-ACM,(pp. 1549–1552), Canada, 2010.
4. Li J, Ott M, Cardie C, Hovy E, "Towards a general rule for identifying deceptive opinion spam". In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics-ACL, (pp.1566–1576), Baltimore, Maryland,2014.
5. Xie S, Wang G, Lin S, Yu PS, "Review spam detection via temporal pattern discovery". In: Proceedings of the 18th international conference on Knowledge discovery and data mining-ACM, SIGKDD (pp. 823–831). ACM, Beijing, China,2012.
6. Shojaee S, Murad MAA, Bin Azman A, Sharef NM, Nadali S, "Detecting deceptive reviews using lexical and syntactic features". In: Intelligent Systems Design and Applications -ISDA, 13th International Conference. (pp. 53–58), Serdang,Malaysia,2013
7. Lau RY, Liao SY, Kwok RCW, Xu K, Xia Y, Li Y, "Text mining and probabilistic language modeling for online review spam detecting",In:Association for Computing Machinery-ACM.Trans Manage InfSyst 2(4):1–30.
8. Fei G, Mukherjee A, Liu B, Hsu M, Castellanos M, Ghosh R "Exploiting Burstiness in reviews for review spammer detection". In: Thirteenth International Conference on Web and Social Media -ICWSM (PP.175–184),2013
9. Webb, S., Caverlee, J., Pu, C.," Social honeypots: making friends with a spammer near you". In: The Computerized Engine Application System-CEAS, (pp. 1–10),2008
10. Lee, K., Caverlee, J., Cheng, Z., Sui, D.Z., "Content-driven detection of campaigns in social media". In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management-ACM, (pp. 551–556),2011.
11. Cao, C., Caverlee, J, "Behavioral detection of spam URL sharing: posting patterns versus click patterns". In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE, (pp. 138–141),2014.
12. Alghamdi, B., Watson, J., Xu, Y, "Toward detecting malicious links in online social networks through user behavior". In: ACM International Conference on Web Intelligence Workshops-(WIW). IEEE, (pp. 5–8),2016
13. Zheng, X., Zeng, Z., Chen, Z., Yu, Y., Rong, C.,"Detecting spammers on social networks". In: Neurocomputing 159 (1), (PP.27–34),2015.
14. P. T. Nguyen, H. Takeda, "Online Learning for Social Spammer

Detection on Twitter", In: arXiv preprint arXiv:1605.04374.
15. Song, L., Lau, R.Y.K., Kwok, R.C.-W., Mirkovski, K., Dou, " Who are the spoilers in social media marketing? Incremental learning of latent semantics for social spam detection".In: Electron. Commer. Res. 17 (1), (pp.51–81),2017
16. Ala'M, A.-Z., Paris, H., et al. "Spam profile detection in social networks based on public features". In: 8th International Conference on Information and Communication Systems (ICICS). IEEE, (pp. 130–135),2017
17. Vidas, T., Owusu, E., Wang, S., Zeng, C., Cranor, L.F., Christin, N. "QRishing: the susceptibility of smartphone users to QR code phishing attacks". In: International Conference on Financial Cryptography and Data Security. Springer, (pp. 52–69),2013.
18. Chen, C., Wang, Y., Zhang, J., Xiang, Y., Zhou, W., Min, G. "Statistical features-based real-time detection of drifted Twitter spam". In: IEEE Trans. Inf. Forensics Secur. 12 (4), 914–925,2017
19. Shehnepoor, S., Salehi, M., Farahbakhsh, R., Crespi, N,"NetSpam: a network-based spam detection framework for reviews in online social media". In: IEEE Trans. Inf. Forensics Secur. 12 (7), 1585–1595,2017
20. Soliman, A., Girdzijauskas, S."AdaGraph: adaptive graph-based algorithms for spam detection in social networks". In: International Conference on Networked Systems. Springer, (pp. 338–354),2017
21. Isa Inuwa-Dutse, Mark Liptrott, IoannisKorkontzelos , " Detection of spam-posting accounts on Twitter ".In: International journal. Springer, (pp. 338–354),2018.