Survey on Network Based Spam Detection Framework for Identifying Trusted Reviews on Social Media

Chaitanya Kale, Prof. Manish Rai
Department of Computer Science & Engineering, RKDF College of Engineering, Bhopal, India
Asst. Professor, Department of Computer Science & Engineering, RKDF College of Engineering, Bhopal, India

Abstract- Major Society of people using internet trust the contents of net. The liability that anyone can take off a survey give a brilliant chance to spammers to compose spam surveys about hotels and services for various interests. Recognizing these spammers and the spam content is a widely debated issue of research and in spite of the fact that an impressive number of studies have been done as of late towards this end, yet so far the procedures set forth still scarcely distinguish spam reviews, and none of them demonstrate the significance of each extracted feature type. In this application, use a novel structure, named NetSpam, which proposes spam features for demonstrating hotel review datasets as heterogeneous information networks to design spam review detection method into a classification issue in such networks. Utilizing the significance of spam features helps us to acquire better outcomes regarding different metrics on review datasets. The outcomes represent that NetSpam results with the previous methods and encompassed by four categories of features; involving review-behavioral, user-behavioral, review linguistic, user-linguistic, the first type of features performs better than the other categories. The contribution work is when user will search query it will display all top products as well as there is recommendation of the product.

Keywords- Social Media, Social Network, Spammer, Spam Review, Fake Review, Heterogeneous Information Networks, Sentiment Analysis, Semantic Analysis

1. INTRODUCTION

Social Media portals play an important role in information propagation. Today a lot of people rely on the written reviews of other users in the selection of products and services. Additionally written reviews help service providers to improve the quality of their products and services. The reviews therefore play an important role in success of a business. While positive reviews can provide boost to a business, negative reviews can highly affect credibility and cause economic losses. Since anyone can leave comments as review, provides a tempting opportunity for spammers to write spam reviews which mislead users' choices. A lot of techniques have been used to identify spam reviews based on

linguistic patterns, behavioral patterns. Graph based algorithms are also used to identify spammers. However many aspects are still unsolved. The general concept of NetSpam framework is to build a retrieved review dataset as a Heterogeneous Information Network (HIN) and to convert the problem of spam detection into a classification problem. In particular, convert review dataset as a HIN in which reviews are connected through different features. A weighting algorithm is then employed to calculate each feature's importance. These weights are then used to calculate the very last labels for reviews using both unsupervised and semi-supervised procedures.

NetSpam is able to find features' importance relying on metapath definition and based on values calculated for each review. NetSpam improves the accuracy and reduces time complexity. It highly depends to the number of features used to identify spam reviews. Thus using features with more weights will resulted in detecting spam reviews easier with lesser time complexity.

2. LITERATURE SURVEY

The pair wise features are first explicitly utilized to detect group colluders in online product review spam campaigns, which can reveal collusions in spam campaigns from a more fine-grained perspective. A novel detecting framework [1] named Fraud Informer is proposed to cooperate with the pair wise features which are intuitive and unsupervised. Advantages are: Pair wise features can be more robust model for correlating colluders to manipulate perceived reputations of the targets for their best interests to rank all the reviewers in the website globally so that top-ranked ones are more likely to be colluders. Disadvantage is difficult problem to automate. The paper [2] proposes to build a network of reviewers appearing in different bursts and model reviewers and their co-occurrence in bursts as a Markov Random Field (MRF) and apply the Loopy Belief Propagation (LBP) method to induce whether a reviewer is a spammer or not in the graph. A novel assessment method to evaluate the detected spammers automatically using supervised classification of their reviews. Advantages are: High accuracy, the proposed method is effective. To detect review spammers in review bursts. To detect spammers automatically. Disadvantage is: a generic framework is not used for detect spammers.

In [3] paper, the challenges are: The detection of fraudulent behaviors, determining the trustworthiness of review sites, since some may have strategies that enable misbehavior, and creating effective review aggregation solutions. The TrueView score, in three different variants, as a proof of concept that the synthesis of multi-site views can provide important and usable information to the end user. Advantages are: develop novel features capable of finding cross-site discrepancies effectively, a hotel identity-matching method with 93% accuracy. Enable the site owner to detect misbehaving hotels. Enable the end user to trusted reviews. Disadvantage is difficult problem to automate. In [4] paper describes unsupervised anomaly detection techniques over user behavior to distinguish probably bad behavior from normal behavior. To find diverse attacker schemes fake, compromised, and colluding Facebook identities with no a priori labeling while maintaining low false positive rates. Anomaly detection technique to forcefully identify anomalous likes on Facebook ads.

Achieves a detection rate of over 66% (covering more than 94% of misbehavior) with less than 0.3% false positives. The attacker is trying to drain the budget of some advertiser by clicking on ads of that advertiser.

In [5] paper, a grouped classification algorithm called Multi-typed Heterogeneous Collective Classification (MHCC) and then extends it to Collective Positive and Unlabeled learning (CPU).The proposed models can markedly increase the F1 scores of strong baselines in both PU and non-PU learning environment. Advantages are: Proposed models can markedly increase the F1 scores of strong baselines in both PU and non-PU learning settings. Models only use language self-contained features; they can be smoothly generalized to other languages. Identifies a large number of implied fake reviews hidden in the unlabeled set. Fake reviews hiding in the unlabeled reviews that Dianping's algorithm did not capture. The ad-hoc labels of users and IPs used in MHCC may not be very specific as they are computed from labels of neighboring reviews. The paper [6] elaborates two distinct methods of reducing feature subset size in the review spam domain. The methods include filter-based feature rankers and word frequency based feature selection. Advantages are: The first method is to simply select the words which appear most often in the text. Second method can use filter based feature rankers (i.e. Chi-Squared) to rank features and then select the top ranked features. Disadvantages are: There is not a one size fits all approach that is always better.

In [7] paper, presenting an efficient and effective technique to identify review spammers by incorporating social relations based on two assumptions that people are more likely to consider reviews from the ones connected with them as trustworthy, and review spammers are much less likely to keep a large relationship network with regular users. Advantages are: The proposed trust-based prediction achieves a higher accuracy than standard CF method. To overcome the sparsity problem and compute the overall trustworthiness score for every user in the system, which is used as the spamicity indicator. Disadvantages are: Review dataset required. The paper [8] proposes to detect fake reviews for a product by using the text and rating property from a review. In short, the proposed system (ICF++) will measure the honesty value of a review, the trustiness value of the reviewers and the reliability value of a product. Advantages are: Accuracy is better than ICF method. Precision is maximizing. Disadvantages are: Process need to be optimized.

The paper [9] provides an overview of existing challenges in a range of problem domains associated with online social networks that can be addressed using anomaly detection. It provides an overview of existing techniques for anomaly detection, and the manner in which these have been applied to social network analysis. Advantages are: Detection of anomalies used to identify illegal activities. Disadvantages are: Need to improve the use of anomaly detection techniques in SNA. The paper [10] proposes a new holistic technique referred to as SpEagle that utilizes clues from all metadata (text, timestamp, and rating) as well as relational information (network), and harness them collectively below a unified framework to spot suspicious users and reviews, as well as products focused via spam. SpEagle employs a review-

network-based classification task which accepts prior knowledge on the class distribution of the nodes, estimated from metadata. Advantages are: It enables seamless integration of labeled data when available. It is extremely efficient.

Survey the prominent machine learning techniques that have been proposed to solve the problem of review spam detection and the performance of different approaches for classification and detection of review spam [11]. The best part of current studies has focused on supervised learning techniques, which require labeled data, a scarcity in terms of online review spam. Advantages are: Higher Performance. Disadvantages are: Required labeled data. The paper [12] help to detect spam profiles even when they do not contact a honey-profile. The discontinuous behavior of user profile is detected and based on that the profile is implemented to identify the spammer. Advantages are: It improves the security. It detects spammers on Twitter which based on the machine learning algorithm. Disadvantages are: Mainly require the historical information to build the social graph.

Proposed system [13] analyzes how spammers who target social networking sites perform. To collect the information about spamming activity, system created a large set of "honey-profiles" on three large social networking websites. Advantages are: The deployment of social Honey pots for harvesting deceptive spam profiles from social Networking. Statistical analysis of these spam's profiles. Disadvantages are: Mainly Time consuming and resource consuming for the system. The paper [14] proposed *Social Spam Guard*, a scalable and online social media spam detection system based on data mining for social network security. GAD clustering algorithm for large scale clustering and integrate it with the designed active learning algorithm Advantages are: Automatically harvesting spam activities in social network by monitoring social sensors with popular user bases; Introducing both image and text content features and social network features to indicate spam activities; Integrating with our GAD clustering algorithm to handle large scale data; Introducing a scalable active learning technique to detect existing spams with constrained human efforts, and carry-out online active learning to detect spams in real-time.

There are two methods for incorporating social context in the quality prediction: either as features, or as regularization constraints, based on a set of hypotheses. The method [15] proposes quite generalizable and applicable for quality (or attribute) estimation of other types of user-generated content. Advantages are: Improves the accuracy of review quality prediction. The resulting forecaster is accessible even when social context is unavailable. Disadvantages are: A portal may lack an explicit trust network.

3. OPEN ISSUES

Online Social Media websites play a main role in information propagation which is considered as an important source for producers in their advertising operations as well as for customers in selecting products and services. People mostly believe on the written reviews in their decision-making processes, and positive/negative reviews encouraging/discouraging them in their selection of products and services. These reviews that reason have emerge as an important issue in fulfillment of a business even as positive opinions can carry

blessings for an employer, bad evaluations can probably effect credibility and motive monetary losses. The critiques written to change customers' perception of ways top a product or a service are taken into consideration as spam, and are regularly written in trade for money.

Disadvantages:

- There is no information filtering concept in online social network.
- People believe on the written reviews in their decision-making processes, and positive/negative reviews encouraging/ discouraging them in their selection of products and services.
- Anyone create registration and gives comments as reviews for spammers to write fake reviews designed to misguide users' opinion.
- Less accuracy.
- More time complexity.

4. SYSTEM OVERVIEW

A novel proposed framework is to representative a given review dataset as a Heterogeneous Information Network (HIN) and to solve the issue of spam detection into a HIN classification issue. In particular, to show the review dataset as a HIN in which reviews are connected through different node types (such as features and users). A weighting algorithm is then employed to calculate each feature's importance (or weight). These weights are applied to calculate the final labels for reviews using both unsupervised and supervised methods. Based on our observations, defining two views for features (review-user and behavioral-linguistic), the classified features as review behavioral have more weights and yield better performance on spotting spam reviews in both semi-supervised and unsupervised approaches. The feature weights can be added or removed for labeling and hence time complexity can be scaled for a specific level of accuracy. Categorizing features in four major categories (review-behavioral, user-behavioral, review-linguistic, user-linguistic), helps us to understand how much each category of features is contributed to spam detection.

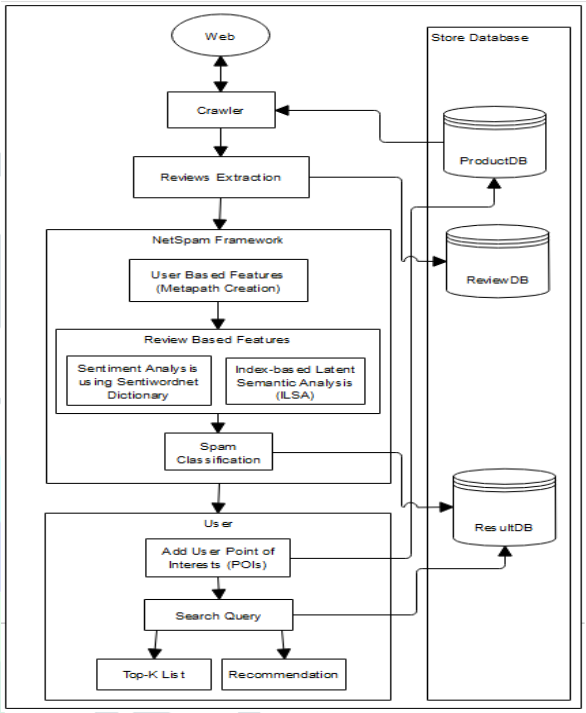

Fig.1 Proposed System Architecture

1. NetSpam framework that is a novel network based approach which models review networks as heterogeneous information networks.
2. A new weighting method for spam features is proposed to determine the relative importance of each feature and shows how effective each of features are in identifying spams from normal reviews.
3. NetSpam framework increases the accuracy as opposed to the state-of-the art in terms of time complexity, which distinctly relies upon to the variety of capabilities used to perceive an unsolicited spam evaluation.

The general concept of our proposed framework is to model a given review dataset as a Heterogeneous Information Network and to map the problem of spam detection into a HIN classification problem. In particular, model review dataset as in which reviews are connected through different node types. The fig. 2 shows the flowchart of NetSpam framework.

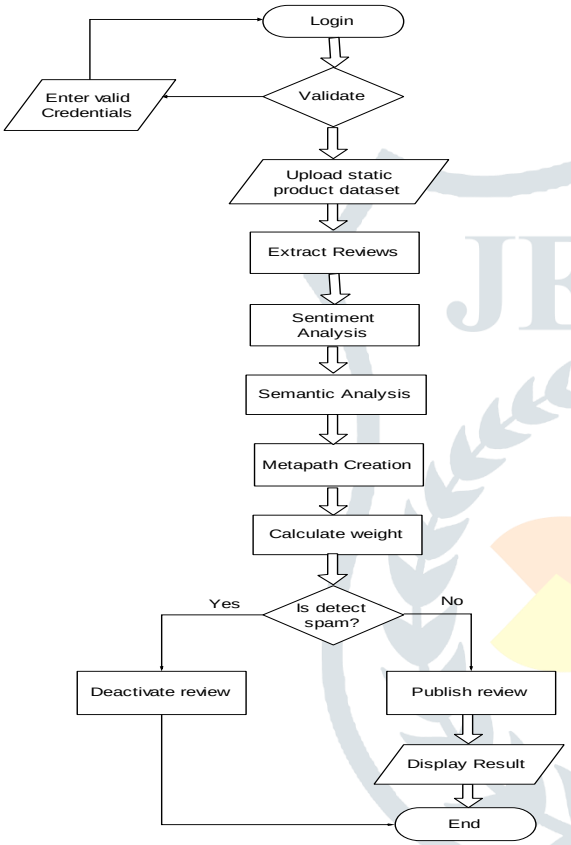

Fig. 2 Flowchart of NetSpam Framework

A weighting algorithm is then employed to calculate each feature's importance. These weights are applied to calculate the final labels for reviews using both unsupervised and supervised techniques. Based on the observations defining two views for features.

Advantages:

1. To identify spam and spammers as well as different type of analysis on this topic.
2. Written reviews also help service providers to enhance the quality of their products and services.
3. To identify the spam user using positive and negative reviews in online social media.
4. To display only trusted reviews to the users.

5. FEATURES

User-Behavioral (UB) based features:

Burstiness: 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.

$$x_{BST}(i) = \begin{cases} 0 & (L_i - F_i) \notin (0, \tau) \\ 1 - \frac{L_t - F_t}{\tau} & (L_i - F_i) \in (0, \tau) \end{cases} \quad (1)$$

Where,

$L_i - F_i$ describes days between last and first review for $\tau = 28$.

Users with calculated value greater than 0.5 take value 1 and others take 0.

User-Linguistic (UL) based features :

Average Content Similarity, Maximum Content Similarity: 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]).

Review-Behavioral (RB) based features :

- Early Time Frame: Spammers try to write their reviews a.s.a.p., in order to keep their review in the top reviews which other users visit them sooner.

$$x_{ETF}(i) = \begin{cases} 0 & (L_i - F_i) \notin (0, \delta) \\ 1 - \frac{L_t - F_t}{\delta} & (L_i - F_i) \in (0, \delta) \end{cases}$$

(2)

Where,

$L_i - F_i$ 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 using threshold: 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.

$$x_{DEV}(i) = \begin{cases} 0 & Otherwise \\ 1 - \frac{r_{ij} - avg_{e \in E*j} r(e)}{4} > \beta_1 \end{cases} \quad (3)$$

Where,

β_1 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)).

Review-Linguistic (RL) based features:

Number of first Person Pronouns, Ratio of Exclamation Sentences containing '!': 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]).

6. CONCLUSION

The novel spam detection framework named NetSpam based on a metapath creation as well as new graph-based method for labeling reviews relying on a rank-based labeling approach. The calculated weights by utilizing this metapath concept can be very impressive in identifying spam reviews and spammers leads to a better performance. In extension, found that even without a train set, NetSpam can calculate the consequence of each feature and it yields better performance in the features' addition process, and performs better than existing works, with only a small number of features. Moreover, after defining four main categories for features our conclusions show that the reviews behavioral category performs better than other categories.

REFERENCES

[1] Ch. Xu and J. Zhang. Combating product review spam campaigns via multiple heterogeneous pairwise features. In SIAM International Conference on Data Mining, 2014.

[2] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh. Exploiting burstiness in reviews for review spammer detection. In ICWSM, 2013.

[3] A. j. Minnich, N. Chavoshi, A. Mueen, S. Luan, and M. Faloutsos. Trueview: Harnessing the power of multiple review sites. In ACM WWW, 2015.

[4] B. Viswanath, M. Ahmad Bashir, M. Crovella, S. Guah, K. P. Gummadi, B. Krishnamurthy, and A. Mislove. Towards detecting anomalous user behavior in online social networks. In USENIX, 2014.

[5] H. Li, Z. Chen, B. Liu, X. Wei, and J. Shao. Spotting fake reviews via collective PU learning. In ICDM, 2014.

[6] M. Crawford, T. M. Khoshgoftaar, and J. D. Prusa. Reducing Feature set Explosion to Faciliate Real-World Review Sapm Detection. In

Proceeding of 29th International Florida Artificial Intelligence Research Society Conference. 2016.

[7] H. Xue, F. Li, H. Seo, and R. Pluretti. Trust-Aware Review Spam Detection. IEEE Trustcom/ISPA. 2015.

[8] E. D. Wahyuni and A. Djunaidy. Fake Review Detection From a Product Review Using Modified Method of Iterative Computation Framework. In Proceeding MATEC Web of Conferences. 2016.

[9] R. Hassanzadeh. Anomaly Detection in Online Social Networks: Using Datamining Techniques and Fuzzy Logic. Queensland University of Technology, Nov. 2014.

[10] R. Shebuti and L. Akoglu. Collective opinion spam detection: bridging review networks and metadata. In ACM KDD, 2015.

[11] M. Crawford, T. D. Khoshgoftar, J. N. Prusa, A. Al. Ritcher, and H. Najada. Survey of Review Spam Detection Using Machine Learning Techniques. Journal of Big Data. 2015.

[12] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in Proc. 26th Annu. Comput. Sec. Appl. Conf., 2010, pp. 1–9.

[13] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots + machine learning," in Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2010, pp. 435–442.

[14] X. Jin, C. X. Lin, J. Luo, and J. Han, "Socialspamguard: A data mining based spam detection system for social media networks," PVLDB, vol. 4, no. 12, pp. 1458–1461, 2011.

[15] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi, "Exploiting social context for review quality prediction," in Proc. 19th Int. Conf. World Wide Web, 2010, pp. 691–700.