ISSN2394-3777 (Print)
ISSN2394-3785 (Online)
Available online atwww.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 5, Special Issue 12, April 2018

E-Spam: Spam Detection Framework for Reviews in Online Social Media

Suranya Das M S #1, Devi Gopal T *2
#Computer Science And Engineering
Gurudeva Institute of Science And technology(GISAT)
Kottayam, kerala, india
Suranyadasms888@gmail.com
devilinish@gmail.com

***Abstract* — We know that online social media portal plays influential rule in the information propagation around the people of the world. that is, big part of the people make the decision about their product or topic based on the available content in the social media(reviews, feedback) The possibility that anybody can leave a review provide a golden opportunity for spammers to write spam reviews about products and services for different interests. Identifying these spammers and the spam content is the one of challenges in the research world. number of studies have been done toward this end, but so far the methodologies put forth still barely detect spam reviews, and none of them show the importance of each extracted feature type. In this study, we propose a novel framework, named ESpam, which utilizes spam features for modeling review datasets as heterogeneous information networks to map spam detection procedure into a classification problem in such networks. Using the importance of spam features ,to obtain better results we can use reviews from real word datasets (Amazon, Yelp). The results show that ESpam outperforms the existing methods and among four categories of features; including review-behavioral, user-behavioral, review linguistic, user-linguistic, the first type of features performs better than the other categories.**

Index Terms—Social Media, Social Network, Spammer, Spam Review, Fake Review, Heterogeneous Information Networks

1 INTRODUCTION

Nowadays, a big part of people rely on available content in social media in their decisions (e.g. reviews and feedback on a topic or product) which is considered as an important source for producers in their advertising campaigns as well as for customers in selecting products and services. The past few years people make the decision about their product and services based on these written reviews. positive/negative reviews encouraging/discouraging them in their selection of products and services. In addition, written reviews also help service providers to enhance the quality of their products and services. These reviews thus have become an important factor in success of a business while positive reviews can bring benefits for a company, negative reviews can potentially impact credibility and cause economic losses. The fact that

anyone with any identity can leave comment as review, provides a tempting opportunity for spammers to write fake reviews designed to mislead users' opinion. so that user could not able to find the good one and spam. 20% of the reviews in the Yelp website are actually spam reviews. On the other hand, a considerable amount of literature has been published on the techniques used to identify spam and spammers as well as different type of analysis on this topic. These techniques can be classified into different categories; some using linguistic patterns in text which are mostly based on bigram, and unigram, others are based on behavioral patterns that rely on features extracted from patterns in users' behavior .but the problem remain unsolved. The general concept of proposed framework is to model a given review dataset as a Heterogeneous Information Network (HIN) and to map the problem of spam detection into a HIN classification problem. In particular, we model 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 utilized to calculate the final labels for reviews using both unsupervised and supervised approaches. proposed ESpam framework that is a novel network based approach which models review networks as heterogeneous information networks. The classification step uses different meta path types which are innovative in the spam detection domain.

II. PRELIMINARIES

As mentioned earlier, we model the problem as a heterogeneous network where nodes are either real components in a dataset (such as reviews, users and products) or spam features. To better understand the proposed framework here first present an overview of some of the concepts and definitions in heterogeneous information networks .

A. *Definitions*

Definition 1 (Heterogeneous Information Network). Suppose we have r(> 1) types of nodes and s(> 1) types of relation links between the nodes, then a heterogeneous information network is defined as a graph G = (V;E) where each node v 2 V and each link e 2 E belongs to one particular node type and link type respectively. If two links belong to the same type, the types of starting node and ending node of those links are the same.

Definition 2 (Network Schema). Given a heterogeneous information network G = (V;E), a network schema T = (A;R) is a meta path with the object type mapping: V ! A and link mapping _ : E ! R, which is a graph defined over object type A, with links as relations from R. The schema describes the meta structure of a given network (i.e., how many node types there are and where the possible links exist).

B *Feature Types*

Review- Behavioral (RB) based features. This feature type is based on metadata and not the review text itself. The RB category contains two features; Early time frame (ETF) and Threshold rating deviation of review (DEV)

Review-Linguistic(RL)based features. Features in this are based on the review itself and extracted directly from text of the review. In this work two main features in RL category; the Ratio of 1st Personal Pronouns (PP1)and the Ratio of exclamation sentences containing '!' (RES) .

User-Behavioral (UB) based features. These are specific to each individual user and they are calculated per user, so we can use these features to generalize all of the reviews written by that specific user. This category has two main features; the Burstiness of reviews written by a single user, and the average of a users' negative ratio given to different businesses.

User-Linguistic (UL) based features. These features are extracted from the users' language and shows how users are describing their feeling or opinion about what they've experienced as a customer of a certain business. We can use this type of features to understand how a spammer communicates in terms of wording. There are two features engaged for our framework in this category; Average Content Similarity (ACS) and Maximum Content Similarity (MCS). These two features show how much two reviews written by two different users are similar to each other, as spammers tend to write very similar reviews by using template pre-written text.

III. RELATED WORKS

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 three following categories.

A *Linguistic-based Methods*

This approach extract linguistic-based features to find spam reviews. Feng et al. [7] use unigram, bigram and their composition. Other studies [1], [2], [8] use other features like pair wise features (features between two reviews; e.g. content similarity), percentage of CAPITAL words in a reviews for finding spam reviews. [1] N jindal's 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 [7] 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 [8] indicates behavioral features show spammers' identity better than linguistic ones. chandy et al. in [6] 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 [8] use temporal and location features of users to find unusual behavior of spammers. Li et al. in [9] 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 [16] almost engage behavioral features like rate deviation, extremity and etc. Xie et al. in [7] also use a temporal pattern (time window) to find singleton reviews (reviews written just once) on Amazon. zhang et al. in [1] use behavioral features to show increasing competition between companies leads to very large expansion of spam reviews on products.

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 [6] 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 [3] also use same algorithm (LBP), and utilize burstiness of each review to find spammers and spam reviews on Amazon. Li et al. in [5] 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.

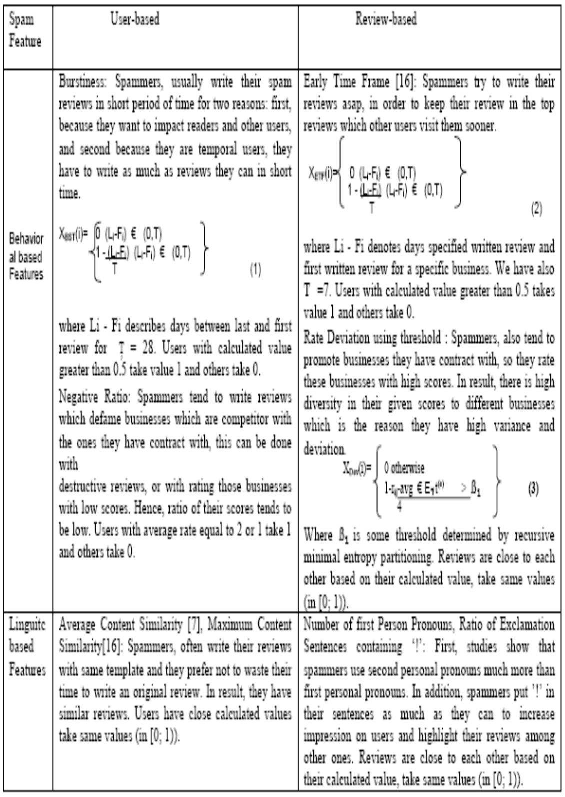
IV. E SPAM; THE PROPOSED SOLUTION

In this section, we provides details of the proposed Solution

A Prior Knowledge

The first step is computing prior knowledge, i.e. the initial probability of review u being spam which denoted as yu. The proposed framework works in two versions; semi-supervised learning and unsupervised learning. In the semi-supervised method, yu = 1 if review u is labeled as spam in the pre-labeled reviews, otherwise yu = 0. If the label of this review is unknown due the amount of supervision, we consider yu = 0 (i.e., we assume u as a non-spam review).

Table:1 Spam Calculation methods



Spam Feature	User-based	Review-based
Behavioral 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. $X_{BST}(i)= \begin{cases} 0 & (L_i-F_i) \in (0,T) \\ 1-\frac{(L_i-F_i)}{T} & (L_i-F_i) \in (0,T) \end{cases}$ (1) where Li - Fi describes days between last and first review for T = 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.	Early Time Frame [16]: Spammers try to write their reviews asap, in order to keep their review in the top reviews which other users visit them sooner. $X_{ETF}(i)= \begin{cases} 0 & (L_i-F_i) \in (0,T) \\ 1-\frac{(L_i-F_i)}{T} & (L_i-F_i) \in (0,T) \end{cases}$ (2) where Li - Fi denotes days specified written review and first written review for a specific business. We have also T =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. $X_{Dev}(i)= \begin{cases} 0 \text{ otherwise} \\ \frac{1-r_{ij}-avg \in E_1 r^{(i)}}{4} > \beta_1 \end{cases}$ (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)).
Linguitc based Features	Average Content Similarity [7], Maximum Content Similarity[16]: 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)).	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)).

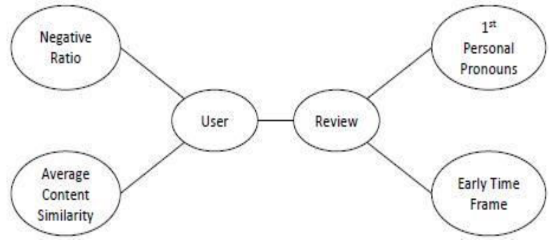


Fig.1 An Example for generating Network Schema generated based on a given spam features list; NR ACS,PPI and ETF

B Network Schema Definition

The next step is defining network schema based on a given list of spam features which determines the features engaged in a spam detection. This schema are general definitions of meta paths and show in general how different network components are connected. For example, if the list of features includes NR, ACS, PP1 and ETF, the output schema is as presented in Fig 1

C Spam Calculation

. This step explaining the calculation of spam and not spam review Table 1 gives the summery of calculation., so by using these calculation we made the decision about the review whether it is spam or not spam.

D Weight Calculation

Next we discuss about features weights and their involvement to determine spamicity. by considering each feature and calculate the overall weight of spam review among the features and we can draw the spamcity of review in the given dataset .

V. CONCLUSION AND FUTURE WORK

Our observation against this topic concluded over here so that herewith we focused on the rating and reviews from deferent reviewers to find the spam activity among them and by getting the better result we can go with real world datasets like (amazon or yelp).we detected the spam based on the user and review behavior and linguistic based features so that it gives better performance than the old researches. So the spammer reviews from the online social media are become blocked and gives a trusted activity for the user.

for future work this framework can applied over spammer community to find group of spammer. And also implement this framework on social media portals .

REFERENCES

[1] Ch. Xu and J. Zhang. Combating product review spam campaigns viamultiple heterogeneous pairwise features. In SIAM International Conference.

[2] F.LI.m huang y.yang and .zhu Learning to identify review spam proceedings of the 22 nd international joint conference on artificial intelligence.

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

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

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

[6] L. Akoglu, R. Chandy, and C. Faloutsos. Opinion fraud detection in online reviews bynetwork effects. In ICWSM, 2013.

[7] S. Feng, R. Banerjee and Y. Choi. Syntactic stylometry for deception detection. Proceedings of the 50th Annual Meeting of the Association

[8] P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In ACM CIKM, 2010.