Perspectives in Communication, Embedded-Systems and Signal-Processing (PiCES)
ISSN: 2566-932X, Vol. 2, Issue 7, October 2018
Proceedings of Knowledge Discovery in Information Technology and Communication Engineering (KITE-2018), 4th May 2018

Detection of Fake Reviewers in Online Social Media

Ayur Raj[1], Abdul Shakib[2], Arshiya Kausar[3], Nidhi Thakur[4]

[1.] 8th sem, Dept of CSE, Bridavan college,ayurraj10@gmail.com

[2.] 8th sem, Dept of CSE, Brindvan college,abdulshakib090@gmail.com

[3.] 8th sem, Dept of CSE, Brindavan college,arshiya.kausar96@gmail.com

[4.] 8th sem, Dept of CSE, Brindavan college,sanaya4567@gmail.com

Abstract: 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). 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 a hot topic of research and although a considerable number of studies have been done recently 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 NetSpam, 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 help us to obtain better results in terms of different metrics experimented on real-world review datasets from Yelp and Amazon websites. The results show that NetSpam 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.

Keywords: Social Media, Social Network, Spammer, Spam Review, Fake Review, Heterogeneous Information Networks

I. INTRODUCTION

Online Social Media portals play an influential role in information propagation which is considered as an important source for producers in their advertising campaigns as well as for customers in selecting products and services. In the past years, people rely a lot on the written reviews in their decision-making processes, and 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 comments as review, provides a tempting opportunity for spammers to write fake reviews designed to mislead users' opinion. These misleading reviews are then multiplied by the sharing function of social media and propagation over the web. The reviews written to change users' perception of how good a product or a service are considered as spam [11], and are often written in exchange for money. 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 [30], [31]. These techniques can be classified into different categories; some using linguistic patterns in text [2], [3], [4], which are mostly based on bigram, and unigram, others are based on behavioral patterns that rely on features extracted from patterns in users' behavior which are mostly meta data based [34], [6], [7], [8], [9], and even some techniques using graphs and graph-based algorithms and classifiers [10], [11], [12]. Despite this great deal of efforts, many aspects have been missed or remained unsolved. One of them is a classifier that can calculate feature weights that show each feature's level of importance in determining spam reviews. In summary, our main contributions are as follows: (i) We propose NetSpam framework that is a novel network based approach which models review networks as heterogeneous information networks. The classification step uses different metapath types which are innovative in the spam detection domain. (ii) 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. Previous works [12], [20] also aimed to address the importance of features mainly in term of obtained accuracy, but not as a build-in function in their framework (i.e., their approach is dependent to ground truth for determining each feature importance). (iii) NetSpam improves the accuracy compared to the state of-the art in terms of time complexity, which highly depends to the number of features used to identify a spam review; hence, using features with more weights will resulted in detecting fake reviews easier with less time complexity.

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 we first present an overview of some of the concepts and definitions in heterogeneous information networks [23], [22], [24].

A. Definitions

a) 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 ∈ V and each link e ∈ 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.

b) Definition 2 (Network Schema)

Given a heterogeneous information network G = (V,E), a network schema T = (A,R) is a metapath with the object type mapping τ : V → A and link mapping φ : E → R, which is a graph defined over object type A, with links as relations from R. The schema describes the metastructure of a given network (i.e., how many node types there are and where the possible links exist).

c) Definition 3 (Metapath)

As mentioned above, there are no edges between two nodes of the same type, but there are paths. Given a heterogeneous information network G = (V,E), a metapath P is defined by a sequence of relations in the network schema T = (A,R), denoted in the form A1(R1)A2(R2)...(R(l−1))Al, which defines a composite relation P = R1oR2o...oR(l−1) between two nodes, where o is the composition operator on relations. For convenience, a metapath can be represented by a sequence of node types when there is no ambiguity, i.e., P = A1A2...Al. The metapath extends the concept of link types to path types and describes the different relations among node types through indirect links, i.e. paths, and also implies diverse semantics.

d) Definition 4 (Classification problem in heterogeneous information networks)

Given a heterogeneous information network G = (V,E), suppose V 0 is a subset of V that contains nodes of the target type(i.e., the type of nodes to be classified). k denotes the number of the class, and for each class, say C1...Ck, we have some pre-labeled nodes in V 0 associated with a single user. The classification task is to predict the labels for all the unlabeled nodes in V 0.

B. Feature Types

In this paper, we use an extended definition of the metapath concept as follows. A metapath is defined as a path between two nodes, which indicates the connection of two nodes through their shared features. When we talk about metadata, we refer to its general definition, which is data about data. In our case, the data is the written review, and by metadata we mean data about the reviews, including user who wrote the review, the business that the review is written for, rating value of the review, date of written review and finally its label as spam or genuine review. In particular, in this work features for users and reviews fall into the categories as follows (shown in Table I):

a) 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) [16].

b) Review-Linguistic (RL) based features

Features in this category are based on the review itself and extracted directly from text of the review. In this work we use two main features in RL category; the Ratio of 1st Personal Pronouns (PP1) and the Ratio of exclamation sentences containing '!' (RES) [6].

c) User-Behavioral (UB) based features

These features 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 [7], and the average of a users' negative ratio given to different businesses [20].

d) 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 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 [11].

III. NETSPAM; THE PROPOSED SOLUTION

In this section, we provides details of the proposed solution which is shown in Algorithm III.1.

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). In the unsupervised method, our prior knowledge is realized by using yu = (1/L)PL l=1 f(xlu) where f(xlu) is the probability of review u being spam according to feature l and L is the number of all the used features (for details, refer to [12]).

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 spam detection. This Schema are general definitions of metapaths 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. *Metapath Definition and Creation*

As mentioned in Section II-A, a metapath is defined by a sequence of relations in the network schema. Table II showsall the metapaths used in the proposed framework. As shown, the length of user-based metapaths is 4 and the length of review based metapaths is 2. For metapath creation, we define an extended version of the metapath concept considering different levels of spam certainty. In particular, two reviews are connected to each other if they share same value.

IV. ALGORITHM: NETSPAM()

```
Input : review-dataset,spam-feature-list,
Pre-labeled-reviews
Output : features-importance(W),
Spamicity-probability(Pr)
% u,v: review,$y_u$:spamicity probability of review u
% f($x_{/u}$): initial probability of review u being spam
% pl: metapath based on feature l, L: features number
% n: number of reviews connected to a review
% $m_{pl}$:the level of spam certainty
% $m_{pl}$: the metapath value
%Prior Knowledge
if semi-supervised mode
if u 2 pre-labeled-reviews
$y_u$ = label(u)
else
$y_u$ = 0
else % unsupervised mode
$y_u$ = 1/L
%Network Schema Definition
schema = defining schema based on spam-feature-list
% Metapath Definition and Creation
for $p_l$ schema
```

VI. CONCLUSION

This study introduces a novel spam detection framework namely NetSpam 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, NetSpam 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 reviewsbehavioral 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. For future work, metapath concept can be applied to other problems in this field. For example, similar framework can be used to find spammer communities. For finding community, reviews can be connected through group spammer features (such as the proposed feature in [29]) and reviews with highest similarity based on metapth concept are known as communities. In addition, utilizing the product features is an interesting future work on this study as we used features more related to spotting spammers and spam reviews. Moreover, while single networks has received considerable attention from various disciplines for over a decade, information diffusion and content sharing in multilayer networks is still a young research [37]. Addressing the problem of spam detection in such networks can be considered as a new research line in this field.

REFERENCES

[1] J. Donfro, A whopping 20 % of yelp reviews are fake. http://www.businessinsider.com/20-percent-of-yelp-reviews-fake-2013-9. Accessed: 2015-07-30.

[2] M. Ott, C. Cardie, and J. T. Hancock. Estimating the prevalence of deception in online review communities. In ACM WWW, 2012.