Two-stage Unsupervised Approach for Combating Social Spammers

Darshika Koggalahewa
School of Electrical Engineering and Computer Science
Queensland University of Technology
Brisbane, Queensland, Australia
darshikaniranjan.koggalahewa@hdr.qut.edu.au

Yue Xu
School of Electrical Engineering and Computer Science
Queensland University of Technology
Brisbane, Queensland, Australia
yue.xu@qut.edu.au

Ernest Foo
School of Information and Communication Technology
Griffith University
Brisbane, Queensland, Australia
e.foo@griffith.edu.au

***Abstract*— Spammers use Online Social Networks (OSNs) as a popular platform for spreading malicious content and links. The nature of OSNs allows the spammers to bypass the combating techniques by changing their behaviours. Classification based approaches are the most common technique for spam detection. "Data labelling", "spam drift", "imbalanced datasets" and "data fabrication" are the most common limitations of classification techniques that hinder the accuracy of spam detection. The paper presents a two-stage fully unsupervised approach using a user's peer acceptance within OSN to distinguish spammers from genuine users. User's common shared interest over multiple topics and the mentioning behaviour are used to derive the peer acceptance. The contribution of the paper is a pure unsupervised method to detect spammers based on users' peer acceptance without labelled datasets. Our unsupervised approach is able to achieve 95.9% accuracy without the need for labelling.**

Keywords—Classification, Spam Detection, Peer Acceptance, Content Interest, Mutual Mentioning

I. INTRODUCTION

Online Social Networks (OSNs) have become an essential tool due to the increasing popularity of social media. OSNs are often vulnerable to a variety of cybercrimes and malicious attacks. Spam can be defined as unsolicited messages sent across electronic media. There are different types of spam such as email spam, instant messaging spam, webspam. "Social spam" in OSNs manifests in different ways such as malicious links, bulk messages, insults, hate speech, fraudulent reviews, fake friend requests, and other varieties. Classification-based methods such as SVM and Random forest are the most commonly used methods for spam detection [1]. These classification methods use specific user and OSN features for spam detection. The variation of research efforts includes a change of features or the use of combined features to improve the accuracy of the classifiers [2], [3]. The classifiers are constructed based on a labelled training dataset with extracted features [3]. However, the generation of labelled training datasets is time-consuming and prone to error. Moreover, spammers change their behaviour over time. Classifiers trained on older data find it difficult to detect new forms of attacks. Therefore, successful classifiers developed for spam detection have problems with data labelling, spam drift, imbalanced datasets, and data fabrication [3], [9]. These open issues along with the highly dynamic and diversified nature of both OSNs and spammers raise a question about the suitability and sustainability of classification-based techniques for future spam detection. Unsupervised solutions do not require labelled training datasets and act as an alternative method for spam detection.

Homophily theory is used as a baseline for our approach. Homophily theory suggests that people with similar interests are likely to connect [7]. Weng et al. and Cardoso et al. [8], [9] investigated that two users should have the same information interest in their shared contents when they mutually follow each other in a social network. The tweet content should be related and thus be similar to the same common topic shared by two users. The higher content similarity of two users is an indication of their similar interest in terms of the topic. The common content interest over a certain area (topic), indicates that users have an agreement on the shared content. If this common interest propagates across the majority of their content usage, we can confirm that these two users have similar interests and they are accepting each other. Following the theory of homophily, it is evident that these two users are possible to connect. By applying the same theory across all the users in the community, it is possible to determine how many users connect each other with reference to their content interest. If the majority of the users are connected to a user via content interest, we determined that they are accepting the user as a reliable user.

It is evident from the literature that if an individual is accepted by their peers in the same social community, the individual will not be considered as suspicious or a malicious user. The other users' acceptability indicates that this user is reliable in the community. A user must show acceptable behaviours within the community to be accepted by peers. Lower peer acceptability is an indication of not belong to a social network and perhaps malicious. Therefore, the peer acceptability of a user in the social network could be considered as a useful measure to detect malicious users.

The paper introduces peer acceptance to represent the readiness of community members to approve an individual's reliability in a social community. Peer acceptance is defined as "the degree to which a child or adolescent is socially accepted by peers in Psychology and sociology. It includes the level of peer popularity and the ease with which a child or adolescent can initiate and maintain satisfactory peer relationships". OSN relationships such as connections or friendships can be used to derive peer acceptance. [16], [17] Nevertheless, it is essential to seek alternative approaches when direct OSN relationships are not present. We introduce a novel approach to determine the peer acceptability of a user in online social networks based on user content and behaviour.

OSN user posts contain various content with various topics. User posts can be considered as a reflection of a user's information interest. Usually, users insert a topic relevant to the content that they included in the post. Hence, the post content of all community members for a certain topic should

978-1-7281-2547-3/20/$31.00 ©2020 Crown

2020 IEEE Symposium Series on Computational Intelligence (SSCI)
December 1-4, 2020, Canberra, Australia

be relevant to the topic and thus similar to each other. The peer acceptability of a user in a topic can be derived by considering the similarity between a user's post content with the peers' post content. Users' pairwise interest similarity based on the users' post content in terms of each topic is developed by following the above idea. User's peer acceptance is generated based on similarity over multiple topics. Normal users should have consistent peer acceptance across the topics that they are interested in. Spammers insert hashtags in their posts to show that they are interested in some topics. Usually, these hashtags are irrelevant to the content and they don't show a similar interest with other users in the same topic in terms of their post content. We used peer acceptability of a user to determine whether a user is spam or genuine. We consider the consistency of peer acceptability at two levels in our spam detection approach, peer pairwise level and community level. Peer acceptability among two individual users is considered at the pairwise level. Pairwise peer acceptance should be consistent over multiple topics, thus the post content of two users should be similar to each other and relevant to all the topics to which the posts belong. Usually, a genuine user must be accepted by the majority of community members. The community-level determines the peer acceptability of a user over all the other users in the community. Higher post content similarity of a user in each topic in the community is used to determine the global acceptance of a user. It is evident from the literature in sociology and the psychology that users with behaviour inconsistencies will be rejected by the peers in the same community [4], [5]. Some genuine users may not post their content often. Their shared content is low and they may not have sufficient content to generate the peer acceptability. It might generate a negative impact on common shared interest-based peer acceptance. Spammers post similar content as well as near-duplicate tweets among spam groups as well as other users in OSN [18]. The similar content will generate higher content similarity over shared content. Both these concerns will negatively affect our spam detection method.

This paper introduces an unsupervised two-stage spam detection approach. Stage I generates a peer acceptance-based spam detection approach. It suffers from the two limitations mentioned in the previous paragraph. In stage II, we modify the stage I approach with explicit mentions to address these two problems. The spam detection is based on peer acceptability of a user derived through a user's common shared interest over multiple topics. Compared to traditional classification-based spam detection approaches, our method does not require a labelled dataset. User's information interest is represented using a three-dimensional user profile based on user, topic and content. The content dimension is represented using frequent words used for each topic. First, we calculate the information interest of two users. Using the calculated information interest, pairwise acceptability is calculated. Using the pairwise acceptability over all the users, peer acceptance of a user is calculated.

People can tag other people in their posts. Spammers often tag other people to increase the reachability of their posts. Genuine users use tagging as an indication of agreement of participation in their conversations. Twitter uses the mentioning with @username for tagging purposes. If two people tag each other in their posts, it is an indication that they have mutually agreed on the content they shared. Hence, we use mutual mention as a factor to increase the pairwise peer acceptance. Having a higher number of mentions in their post is an indication of suspicious behaviour. We considered the average mentioning across multiple topics and generate a Local Mentioning Score (LMS) as a reduction feature for overall peer acceptability.

Our new algorithm is tested with two datasets. The Fake Project [15] dataset is publicly available and the other dataset is crawled and prepared by ourselves for the experiment. Existing classification-based spam detection approaches are used as baseline systems to test the approach. It is evident from the evaluation results that our peer acceptance-based unsupervised method is a potential alternative for traditional classification-based spam detection approaches when no labelled training datasets are available. Even though it does not beat the classifier results, it is much closer to classification-based approaches than other unsupervised attempts. The contribution of the paper can be summarized as follows.

1. Development of a novel peer acceptance approach using the user's common shared interest and explicit mentions with other users.
2. A pure unsupervised spam detection approach that does not require labelled training datasets.

Section 2 of the research paper describes the existing contributions related to classification-based spam detection approaches. The experimental approach and the evaluation results are presented in section 3. Section 4 explains the experimental results along with the observations obtained from the research. A discussion on strengths and weaknesses is described in section 5. Section 6 concludes the research work.

II. State of the Art

OSNs are full of social spams. There are different types of social spams such as malicious links, bulk messages, insults and hate speech, fraudulent reviews and fake friends. There is a significant growth in social spam attacks which include phishing scams, publishing malicious content and promotion campaigns over the last two decades [1], [2]. It is common for spammers to change their spam behaviours to disguise themselves as genuine users. Among all other OSN platforms, Twitter users are highly vulnerable to spam activities [10]. Traditional Twitter spam detection approaches employ machine learning algorithms to label malicious tweets by analyzing the data used in the platform. Spammers frequently use embedded malicious URLs to redirect twitter users to malicious websites. They broadcast a series of malicious tweets inviting the users to follow their links. Twitter usually use various features such as content-based, account-based or graph-based features to apply classifiers to combat spammers. It is essential to have a labelled dataset for feature engineering and the classification process. URL based features are the other common type of strategy to detect spammers. Thomas et al. [10] introduced a spam detection approach by analyzing the product ratio of shortened URL availability in spam and non-spam tweets.

Current trends and the development of text mining approaches allow the content to be used for spam detection processes. Nevertheless, the representation of content in a single dimension is challenging and requires a sparse multidimensional representation. Yang et al. [11] developed an approach to use the correlation of two user accounts by using tweet contents. Chu et.al [12] extended their work by

considering the content similarity as a feature for the classification. It is evident from the literature that, people heavily use account and URL based features rather than content-based features for spam detection [12]. Content similarity can be used as an effective feature for spam detection. Semantic similarity of the post content will enhance the performance of features based on content similarities. There are various approaches used to quantify the similarity. Content-based, ontology-based, feature-based, and hybrid approaches use more than one technique for semantic similarity measurement [14]. Li et al. [13] used WordNet to calculate semantic similarity based on "hyponymy". Homophily theory suggests that "contact between similar people occurs at a higher rate than among dissimilar people" [7]. "The pervasive fact of homophily means that cultural, behavioural, genetic or material information that flows through networks will tend to be localized" [6].

III. Approach

This paper develops a fully unsupervised approach to detect spammers based on peer acceptance calculated through user's information interest. Our approach follows the idea that "Genuine users tend to share similar content for the same topics and spammers are less likely to maintain a common shared interest with other users". People have various topics of interests in different social, cultural and demographic avenues. The posts published in OSNs contain topics (hashtags, post headings, titles etc.) to highlight the content. Spammers include some frequent hashtags in their posts to increase the reachability. Hence spam topics could be irrelevant to their post content. We consider such behaviour of having irrelevant content for a certain topic as suspicious. Continuation over multiple topics could be a sign of potential spamming. Further, if a user's interest is dissimilar to other users over multiple topics, we could assume that the user does not belong and could be a potential spammer. In addition to that, people tag/mention each other in their posts. Spammers also use mentions to increase reachability. Genuine people may not mention other people unless they are known or linked to each other. It is fair to assume that mentioning unknown people is suspicious. Our approach uses these two mentioning behaviours to generate the peer acceptability of a person. Our approach examines two user groups separately.

- G1: Users without mentioning in their posts
- G2: Users with mentioning in their posts

Our approach begins with separating the users into the above groups. For G1, Peer Acceptance (PA) between users is generated based on a common shared interest. For G2, Common shared interest-based PA is combined with mentioning based PA.

1) Peer Acceptance based on the common shared interest

The goal of this work is to detect suspicious users in online social networks based on the users' content interest similarities. It is conceivable that shared content interest among users can be utilized to measure the acceptance of a user as perceived by others and extend it to detect spammers in OSNs. The content of a certain topic can be determined by a set of representative words extracted from all the posts of that topic. A user's content interest to a topic can be estimated by using the similarity between the topic content and the content of the user's posts on that topic. If a user's content interest is always similar to another user's interest across their shared topics, this user is more likely to be accepted by the other user.

2) Peer Acceptance based on Explicit Mentioning.

The peer acceptance-based spam detection results contain higher false negatives where a set of spammers are identified as genuine users with higher overall peer acceptance. This occurs as a result of less shared content among pairs of users. We use mutual mentioning (MM) to handle this problem. We assume that, if a pair of users mutually mention themselves in their posts, they are strongly accepting each other. We used MM as a factor to increase the pairwise peer acceptance. Analysis of stage I results revealed that the reason for many false positives is due to higher content similarity in spam user groups. Spammers post similar content or near duplicate tweets as groups for a set of topics. This will increase the content similarity for spammers over such topics. As a result, their overall peer acceptance is high. This will label the spam user as a genuine user. To address this problem, we analyzed the mentioning behaviour of spammers. Compare to genuine users' spammers have a higher number of mentions across multiple topics. If a user continuously mentioned some other users in many posts over multiple topics, it is suspicious. Based on this hypothesis, we developed a Local Mentioning Score (LMS) for each user. A higher LMS is an indication of spam behaviour. The Peer Acceptance model and the spam detection approach introduced in stage I is modified with Mutual Mentions and Local Mentioning Score for better performance. Section A and B discuss a Peer Acceptance based spam detection and section C and D describe the enhanced spam detection approach.

3) Using Peer acceptance for spam detection

User content interests are essential for determining the overall peer acceptance of a user in a social network. For malicious users, the purpose of joining a social network is not for sharing information with peers in the community. Therefore, it is unusual for them to show similar content interest with other users in the same topic across all topics. Malicious users are not to be accepted by most members of the community. Based on this idea, the peer acceptance calculated using common shared interest can be used to detect spammers. The next section contains a detailed description of the peer acceptance approach with formal definitions.

A. Users' Peer Acceptance (PA)

In this paper, Twitter posts (i.e., tweets) which include topics (i.e., hashtags) and words, as well as the users who publish the posts, are used. In order to derive the peer acceptance between users, we first describe a method to represent users' interest based on their Twitter posts. Let $U = \{u_1, u_2, \dots, u_m\}$ be a set of users in a social network and $T = \{t_1, t_2, \dots, t_n\}$ be a set of hashtags used by users in U, the topics in T are frequent topics extracted from all the tweets posted by users in U based on the percentage of tweets which contain each topic.

Assume P_i to be a document constructed by concatenating all the tweets posted by user $u_i \in U$. For the words in P_i, based on their tf-idf values on the collection $\{P_1, \dots P_m\}$, the top frequent words, denoted as W_i, are

selected for user u_i. Overall, $W = \bigcup_{u_i \in U} W_i$ contains all the frequent words for all users.

1) Representation of user's content interest

A user's information interest can be reflected in the user's posts. A user's interest can be calculated using the content of the user's posts over multiple topics. A tensor $\mathcal{CI} \in \mathbb{N}^{|U|\times|T|\times|W|}$ is constructed to represent the user profile, where $\mathcal{CI}(u_i, t_j, w_k)$ is the term frequency of w_k in user u_i's tweets in topic t_j. $\overrightarrow{\mathcal{CI}}(u_i, t_j) \in \mathbb{N}^{|W|}$ is the vector to represent user u_i's interest in the topic t_j.

Representation of topics: The content of a topic *t* can be represented by the average interest of all users in that topic (i.e., the centroid vector for topic *t*), which can be derived from vectors $\overrightarrow{\mathcal{CI}}(u_i, t), i = 1, 2, \dots, m$ as follows:

$$\vec{T}_t = \frac{1}{m}\sum_{u_i \in U} \overrightarrow{\mathcal{CI}}(u_i, \mathrm{t}) \quad (1)$$

Users can mention a set of topics in their tweets. Nevertheless, they may be interested in the topics with different degrees. Hence it is essential to find the set of topics that a user is most interested in. We assume that the user's topics which are closer to the topic representation $\vec{T}_t$ can be considered as the topics of interest to the user, while the topics that are far from the topic representation $\vec{T}_t$ may not be of interest to the user.

Whether a user is interested in a topic or not is determined based on the similarity between the topic centroid vector and the user's content vector for the topic. A user's topic set is defined as follows, where σ is a threshold of the minimum similarity.

$$\mathcal{UT}(u) = \{t | t \in T, sim(\overrightarrow{\mathcal{CI}}(u, t), \vec{T}_t) \geq \sigma\} \quad (2)$$

2) Peer Acceptance based on content interest

It is reasonable that users in a social network do not have the same interest as other users but can partially share their interest. By 'peer acceptance' we want to estimate how much of a user's information interest is shared with another user across their shared topics. The basic idea is that the more a user shares his/her interest with another user, the more the user accepts the other user in terms of information interest. Let $PA(u_i, u_j)$ denote the peer acceptability of user u_i to u_j, where u_i is the acceptee of the relationship and u_j represents the acceptor. We consider that both users' content interest for their common topics can be used to determine the peer acceptance of the two users. We define the peer acceptance of u_i to u_j as the ratio between the common interest of the acceptor u_j and the acceptee u_i across their shared topics and the acceptor u_j 's interest for all the topics of u_j. Equation (3) is defined to calculate the peer acceptance of u_i to u_j. In equation (3) the numerator is the content interest of user $\boldsymbol{u_j}$ over the common topics of both users weighted by the similarity of the two users' interests, while the denominator is the content interest of user $\boldsymbol{u_j}$ for all the topics of user $\boldsymbol{u_j}$.

$$PA(u_i, u_j) = \frac{\sum_{t_k \in \mathcal{UT}(u_i) \cap \mathcal{UT}(u_j)} (sim(T_k, \mathcal{CI}(u_j, t_k)) * sim(\mathcal{CI}(u_i, t_k), \mathcal{CI}(u_j, t_k)))}{\sum_{t_k \in \mathcal{UT}(u_j)} sim(T_k, \mathcal{CI}(u_j, t_k))} \quad (3)$$

The Algorithm *Peer acceptance* describes the steps to generate peer acceptance between pairs of users based on their content interest over multiple topics.

If $PA(u_i, u_j)$ is larger than a pre-specified threshold, it is considered that the user $\boldsymbol{u_j}$ accepts user $\boldsymbol{u_i}$**.** This threshold is an experimental threshold which is calculated by considering the average peer acceptance for all the users in the network. The peer acceptance of a user in a social network can be measured by the percentage of users who accept the user in the same network. If the majority of the members in a network accept a certain user, we could say that the user is a reliable person. The threshold is determined with the assumption that the majority of the users need to accept the user as a reliable user. If the peer acceptance of $\boldsymbol{u_i}$ to $\boldsymbol{u_j}$ is above the threshold we consider that $\boldsymbol{u_j}$ accepts $\boldsymbol{u_i}$. The overall peer acceptance of a user is determined by the percentage of users who accepts the user.

Algorithm 1: Peer acceptance

Compute pairwise peer acceptance for all users

Input: User content profile $\mathcal{CI}$,

Topic content representation $T_j, j = 1, \dots n$,

Representative topics for each user $\mathcal{UT}(u_i)\ i = 1, \dots, m$

Output: Peer Acceptance matrix $PA \in \mathbb{R}^{|U|\times|U|}$

1. For each user u_a in U
2. For each user u_b in U
3. $PA(u_a, u_b)$ =0 // initializing the Peer Acceptance
4. $\mathbb{C} = \mathcal{UT}(u_a) \cap \mathcal{UT}(u_b)$//common topics
5. $S_n = 0\ , S_d = 0$
6. For each $t_k \in \mathbb{C}$
7. $S_n = S_n + (sim(T_k, \mathcal{CI}(u_b, t_k)) * sim(\mathcal{CI}(u_a, t_k), \mathcal{CI}(u_b, t_k))$
8.
9. For each $t_j \in \mathcal{UT}(u_b)$
10. $S_d = S_d + sim(T_j, \mathcal{CI}(u_b, t_j))$
11. $PA(u_a, u_b) = \frac{S_n}{S_d}$
12. Return PA

3) Overall Peer Acceptance of a user

The overall peer acceptability of a user denoted as $\overline{PA(u_i)}$, is calculated as follows.

$$\overline{PA(u_i)} = \frac{\sum_{j=1}^{m} PA(u_i, u_j)}{m} \quad (4)$$

m is the total number of users who have shared interest with u_i . For users in G1, the equation (4) is used to calculate the peer acceptability of the user.

B. Explicit mentioning for peer acceptance

1) Nature of less shared content and posting behaviours

The peer acceptance of two users is calculated based on the two users' interest in their common topics. The more the topics they share, the more accurate the peer acceptance calculated using equation (3). In social networks, it is normal that some users are not interested in many topics. These users are not necessarily spammers. However, they could be mistakenly considered as not acceptable by peers due to not sharing many topics with peers or having unique interests. Analysis of these users indicates that they use very few topics and other users generally use a larger number of topics. The spam detection results of section A contains a higher false-positive rate where a set of genuine users were identified as spammers due to lower peer acceptance. This occurs due to lower usage of shared content among such genuine users.

2) Higher content similarity of Spam groups

It is evident from the experiment that the spammers post similar content or near duplicate tweets as groups for the same set of topics. This is considered as a common behaviour for spammers. At the same time, spammers embed trending topics with their posts and broadcast the same or closely related message as a group. In terms of numbers, there are not many such instances. Besides many spammers exploit globally popular words in their tweets. Some of these collective spamming activities include "similar tweets promoting the same product/service that is posted by many spammers" and some aggressive content promotions to draw the attention of ordinary users. The literature reveals that the spammers are posting spam content among spam accounts to remain hidden from spam detection approaches [12] and there are plenty of spam campaigns to get the attention of social network users [17].[19]. On average, over 35 % of selected tweets from all two datasets are near-duplicate for spam users while genuine users have less than 5% of the near duplicated tweet content. It is evident from the experiment that the majority of these spam accounts are created for promotional activities and genuine users are less likely to post near-duplicate tweets. This spam behaviour generates a high content similarity in this set of users. In stage I of our approach, if users have not connected and shared many topics with genuine users and if they share the majority of their content with the spammers, their pairwise peer acceptability is high in most of the cases. As a result, their average peer acceptability is also high. The final detection would identify these users as "Genuine" users and it will affect the overall performance of the detection.

To address the two problems mentioned above we used the mentioning behaviour of OSN users as an additional feature. Peer acceptance-based spam detection approach is extended with two levels of mentioning. At first, we consider mutual mentioning at pairwise peer acceptance level as a solution to address the low availability of shared content in genuine users. We use mutual mentioning as a factor to increase the pairwise peer acceptance. Second, we developed a local mentioning score (LMS) across all topics as a solution to the high content similarity of spam groups. Finally, the model introduced in stage A is enhanced by embedding mutual mentioning and LMS.

3) Tagging behaviour in Online Social Networks

Tagging refers to using a social handle or username of a person or business in a post or photo. Tagging and mentioning of people in posts are common among OSN users. When we tag people or content in a post or photo, we are identifying them and essentially "linking" them in our post. The most common way to tag in twitter is, to use the **@** symbol and then the user's profile name. Using the mentioning or tagging signifies to someone that we are talking about them, It is common to mention other people. Both spammers and genuine users follow the same approach. Using mentions is a common behaviour in spammer posts and having many mentions can be suspicious behaviour. Nevertheless, mentioning each other in their posts or mutual mentioning is an indication that both parties are known/linked to each other and there is an agreement of the content that they are sharing. In OSNs any person can mention other people in their posts. But we may not mention them back if we are not certain about the person or the content. Continuous availability of such mutual mentions is an indication that there exists higher acceptability or strong link between the pair of users. Spammers often include mentionings in their posts.

Literature suggests that the "number of tags present in a normal user's profile is less when compared to that of a spammer's profile. Spammers, exploiting the tagging feature, tag a large number of users and pages"[18]. On Twitter, we observed that normal users do not use a large number of @mentions when compared to spammers. This is because normal users communicate with a small group of users who are active on Twitter, whereas spammers mostly @mention every Twitter contact regardless of a user's activity status [18]. In other terms, one-way mentioning is a potential spamming behaviour while two-way mentioning is a potential sign of the peer acceptability in OSNs.

Based on the nature of two different mentioning behaviours of spam and genuine users we formed the following two hypotheses for our approach.

H1: *If two users mutually mention each other in their posts over a certain topic we can assume that both users have a common acceptance of that topic, and they accept each other's interest.*

H2: *If a user explicitly inserts more mentions in their posts and if this mentioning behaviour continues across multiple topics, it is suspicious and hence the overall peer acceptability of the user should be decreased.*

By following the above two hypotheses, we developed the Acceptance Factor for Peer Acceptance through a common shared interest (AFPA). At the overall peer acceptance level, we employ the Local Mentioning score (LMS) by following the second hypothesis.

4) Acceptance Factor for Peer Acceptance (AFPA)

By 'peer acceptance' we want to estimate how much of a user's information interest is shared with another user across their shared topics. We need to define some other features

which can strengthen our idea of peer acceptance based on a common shared interest. The basic idea is that, if two users mutually mention each other, the two users more likely to accept each other in terms of information interest. So $PA(u_i, u_j)$ should be increased. We define the Pairwise Acceptance Factor for Peer Acceptance of u_i to u_j, PAFPA (u_i, u_j), as the ratio between mentions of the two users. Let $Mention\ (u_i, u_j\ t_k)$ be the percentage of $u_j's$ posts in topic t_k that mentioned u_i, $PAFPA(u_i, u_j)$ is defined below.

$$PAFPA(u_i, u_j) = \sum_{t_k \in \mathcal{UT}(u_i) \cap \mathcal{UT}(u_j)} \frac{Mention\ (u_i, u_j\ t_k)}{Mention\ (u_j, u_i, t_k)} \quad (5)$$

Since the mutual mentioning is an indication of stronger common shared interest, we consider the mutual mentions for the content only used in common topics. Therefore, the Acceptance Factor for Peer Acceptance (AFPA) of user u_i is the average of PAFPA of u_i to all other users who have mutual mentions with u_i. m_i is the total number of users who have mutual mentions with u_i.

$$AFPA(u_i) = \frac{\sum_{j=1}^{m_i} AFPA\ (u_i, u_j)}{m_i} \quad (6)$$

5) Local Mentioning Score (LMS)

It is evident from the literature that the spammers include mentions to increase their reachability. Having a higher number of mentions across multiple topics is an indication of suspicious behaviour. For a genuine user, they should have a balanced number of mutual mentions and they may not have unnecessary mentions made by themselves. We define the LMS by following this idea.

Let $mf\ (u_i, u_j)$ be the number of times that u_i was mentioned by u_j, then the relative mutual mentions of u_i in terms of u_j, denoted as MM(u_i, u_j), is defined as follows.

$$MM(u_i, u_j) = \frac{mf\ (u_i, u_j)}{mf(u_j, u_i)}.$$

The overall mutual mention of user u_i is calculated using the summation of $MM\ (u_i, u_j)$ over all the users.

$$OMM_i = \sum_{j=1}^{m} MM\ (u_i, u_j).$$

Finally, the LMS of user u_i is defined below

$$LMS(u_i) = \frac{OMM_i}{\mathrm{M}_i} \quad (7)$$

Where M_i is the total number of users mentioned by u_i,

$$\mathrm{M}_i = \sum_{j=1}^{m} mf\ (u_i, u_j)$$

LMS is less than 1 for spammers with less mutual mentions and it should be much smaller compared to the genuine users. For genuine users, it is expected to be 1 or greater than one since we expect more mutual mentions for genuine users. Spammers who made higher mentions to increase their reachability will be penalized from this feature. Finally, overall peer acceptance for users in G2 is calculated by using both common shared interest-based peer acceptance and mentioning based peer acceptance. Equation (8) defines the overall peer acceptability $\overline{PA'(u_\iota)}$ of user u_i in G2.

$$\overline{PA'(u_\iota)} = \overline{PA(u_\iota)} + AFPA(u_i) * LMS\ (u_i) \quad (8)$$

The $AFPA(u_i) * LMS$ indicates how strong a user is accepted through the mentions made by themselves and the mentions received from the peers. As stated above $LMS\ (u_i)$ is close to 1 for genuine users and the content acceptability that they have shown through their content interest will remain the same. For spammers, $LMS\ (u_i)$ is closer to 0 and their content acceptability will be penalized due to their unusual mentioning behaviour. Depending on the hypothesis, the $AFPA(u_i) * LMS$ must be higher for the genuine users and it should be 0 or much lower for the spam users. This will penalize the spam users with low mutual mentions and higher mentions by themselves. The method ensures that the spammers would have lower overall peer acceptability over all the other users.

C. Spam Detection using peer acceptance

For each group, we calculate how many users, are accepting a certain user in the OSN. To calculate the number of peers, who accepts the individual, we check whether pairwise peer acceptance is above the overall peer acceptability of the user. If it is above the overall peer acceptability, it is accepted and rejected otherwise. For the two groups G1 and G2, the overall peer acceptability thresholds are different. The experiment shows that the threshold for G1 is lower compare to G2.

Finally, from the total number of pairs above the threshold, we decide whether the user is a spammer or a genuine user. In that case, we use peer acceptance of a user to detect spammers in the social network. If peer acceptance of a user is above 40% we consider that the user is a genuine user and otherwise a spammer. 40% is an experimental threshold used for the detection approach. Initially, we assumed that at least 50% of the community must accept that user as a peer belongs to that community. Several experiments were conducted by decreasing and increasing the threshold with 2% intervals. Finally, it is observed that at least 40% of the users should accept a user for accurate identification. The acceptance is calculated for each pair and the average acceptability of the user is calculated using the pairwise acceptance. Based on peer acceptance, a user will be categorized as "Genuine" or "Spam".

IV. Experiment and Evaluation

In the proposed spam detection algorithm, we consider common shared interest over multiple topics among the users to calculate the peer acceptability of a user. We expand the peer acceptability by adding explicit mentioning at two levels. We used the user's information interest derived from the content similarity over different topics (hashtags) in tweets to represent the users. We also used mutual mentioning between pair of users and average mentioning distribution over topics for the experiment.

A. Dataset Description and criteria for evaluation

The publicly available datasets do not contain sufficient data related to mutual mentions. The Fake Project dataset [15] is a dataset of twitter spambots and genuine accounts. By satisfying the requirement of the presence of mutual mentions

and number of posts, 981 users were used for the experiment. To conduct the experiment, we have collected our own twitter dataset Peer Acceptance Data Collection (PADC). Both manual labelling and classification-based labelling is used to label the dataset. 1265 users were used for the experiment.

The following criteria are used to calculate the accuracy of the system. The true positive (TP) is the number of correctly identified spammers. True negative (TN) is the number of correctly identified legitimate users. False-positive (FP) is the number of wrongly identified spammers. False-negative (FN) is the number of wrongly identified legitimate users. The accuracy is calculated as (TP + TN)/ (TP + TN + FP + FN).

B. Impact of common shared interest

It is evident from the analysis that the peer acceptance between a pair of users is proportional to the number of common topics between the two users. Peer acceptance is high when the number of common topics is increased. The analysis is conducted for two user groups, spam and genuine separately. Fig. 1 depicts the correlation between the average pairwise acceptance and the number of common topics for genuine users. Though this observation is valid for genuine users, spam users did not follow the same pattern.

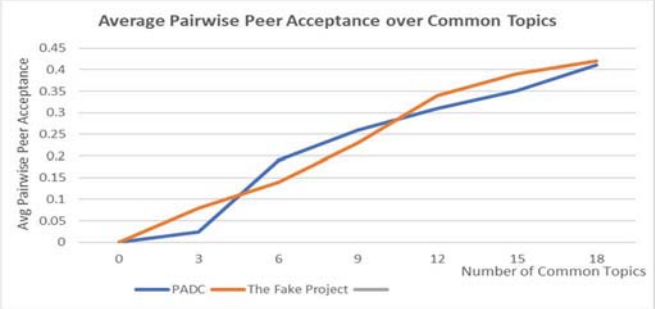

Fig. 1. Average pairwise acceptance distribution over common topics

Spammers have fewer common topics compared to genuine users. Some spam users have shown very high peer acceptance between a pair of users. It is revealed that their content for each topic was similar for most occasions. Spammers include frequent hashtags in their posts to increase the reachability. The other reason is these users can be from the same spam campaigns. They can also be "focused spammers". Some genuine users have not shared many contents with others. They can be isolated users in the OSN. In such scenarios, the acceptance is much closer to 0. Usually, spammers have more posts than the genuine users.

C. Observations from the datasets

It is noticed that spammers include frequent hashtags in their posts without relating to the content. The nature of malicious users targeting the trending topics confirms the findings from [11] and [19]. Some of the frequent hashtags are barely used by genuine users. In contrast, spammers use them quite often. Spammers embed frequent hashtags in their posts as a method of promotion. This supports our hypothesis that "spammers should not have a common shared interest with other users in the community compared to genuine users". We analyzed the content similarity distribution for the same topic in both genuine and spam users. For some instances, the content similarity of a pair of users is high for spammers. This is due to the same content usage by spammers. But when we consider the average content similarity over all the users, genuine user's content similarity is higher than the spammers and there is a clear difference between the similarities. Fig. 2 depicts the average content similarity between the user's content in a topic and the topic representation for each topic. This supports the fact that spammers' post content deviated from the content of legitimate users for any given topic.

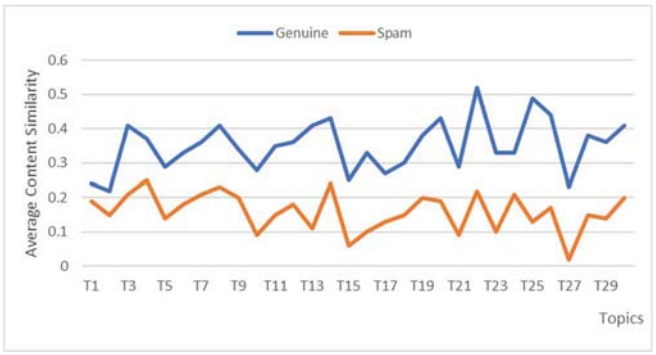

Fig. 2. Average Content similarity distribution variation for spam and genuine users.

D. Detection of Social Spammers using PAccept algorithm and overall based on users' peer acceptance

The overall peer acceptability of a user is calculated at two stages for two user groups G1 and G2. For G1, First, we execute the PAccept algorithm to derive the pairwise peer acceptance for each pair of users. Then based on the average peer acceptance, the acceptance between each pair of users is determined. For G2, we combine explicit mentioning when calculating the average peer acceptability of a user. We use mentioning behaviours to penalize the spammers. Finally, if a user is accepted by at least 40% of users, we consider the user as a legitimate user, and a spammer otherwise. The results are compared with the original label of the user in the datasets. Table 1 summarizes the accuracy of detection PA only and the PA + Explicit mentioning (EM). for both datasets. The results are encouraging and close to the best results of the traditional classification-based spam detection techniques referred to in the literature.

TABLE I Accuracy of spam detection for datasets

Dataset	Model	Accuracy
PADC	Our model: PA only	86.39
	Our model: PA+EM	94.62
	Baseline [18]	98.97
The Fake Project	Our model: PA only	87.59
	Our model: PA+EM	95.92
	Baseline [18]	98. 67

V. DISCUSSION

The foundation of our approach is that a genuine user's content is relevant to the topic. Their content should be consistent with other users who used the same topic and the posting behaviour is consistent across multiple topics. Spammers' behaviour is inconsistent over multiple topics. Spammers often insert content which is irrelevant to the topics of the posts. We use the behaviour inconsistency of common shared interest over multiple topics to generate our Peer Acceptance model. The Peer Acceptance model is enhanced with the explicit mentioning behaviour of the user for the users who have mentions. Depend on the peer acceptability of the user we perform the spam detection. Peer

Acceptance based spam detection approach was developed as an alternative to traditional classification-based approaches. The main contribution of our approach is it does not require labelled datasets. In terms of spam detection, this is more vital since spammers tend to change their behaviour frequently. The results are much closer to classifier-based results which is encouraging. The approach consists of two stages, where we considered the explicit mentioning behaviour of the users to enhance the peer acceptability. Results indicate that false positives are high, i.e., many genuine users were categorized as spammers. We identified the following reasons for this behaviour. Some users have fewer posts and they barely use topics in their posts. Hence their overall peer acceptance is low, which makes them unacceptable to peers. They might be isolated ones with low social interaction. But they do not have to be spammers. For such users, it is hard to determine a common shared interest. In terms of mentioning, Fake Project dataset does not have users with many mutual mentions.

The user content is represented using frequent words with their synonyms. The content representation could be further improved and it may improve the accuracy of the detection. The results contain some spam users with very high content similarity as well as higher peer acceptance relationships. This is due to spammers using similar or closely related content with the same topic. There were some cases where spammers have sent the same content with different topics in their posts. If they use few topics and connect only with spammers their overall peer acceptability is high.

User content interests are essential for determining the overall peer acceptance of a user in a social network. For malicious users, the purpose of joining a social network is not for sharing information with peers in the community. Therefore, it is unusual for them to show similar content interest with other users in the same topic across all topics. The peer acceptance of two users is calculated based on the two users' interest in their common topics. The more the topics they share, the more accurate the peer acceptance calculated using Equation 3. In social networks, it is normal that some users are not interested in many topics. These users are not necessarily spammers. However, they could be mistakenly considered as not acceptable by peers due to not sharing many topics with peers or having unique interests. Analysis of these users indicates that they use very few topics and they are focused only on these topics. Other users generally use a larger number of topics and their interest is diverse. Hence it would be biased to use the same peer acceptance threshold for both diverse and focused users. The future work includes the improvements for addressing the bias content interest.

VI. Conclusion

The paper introduces a novel unsupervised spam detection algorithm. It uses the user's common shared interest and the mentioning behaviour to derive the peer acceptability of a user in OSN. We developed a spam detection approach based on peer acceptance. The system does not require labelled training datasets. Experiments are conducted on two datasets and the results are encouraging and much closer to the classifier-based spam detection systems. The results could be improved with an enhanced user profile. The insufficient content availability in the datasets is another limitation that affects the model. Users with fewer topics and posts are not willing to share many contents with others. They are focused users. They might be penalized with the system. We handled that problem for a certain extent by using explicit mentions. Both tested datasets are from the twitter social network. Future works include addressing the limitations of user focusness and improvement of user profiles. The method should work well for some other OSNs like Facebook since they do not have character limitations.

References

[1] W. Hua and Y. Zhang, "Threshold and Associative Based Classification for Social Spam Profile Detection on Twitter", 2013 Ninth International Conference on Semantics, Knowledge and Grids, 2013.

[2] Q. Dang, Y. Zhou, F. Gao and Q. Sun, "Detecting cooperative and organized spammer groups in micro-blogging community", Data Mining and Knowledge Discovery, vol. 31, no. 3, pp. 573-605, 2016.

[3] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen and B. Zhao, "Detecting and characterizing social spam campaigns", Proceedings of the 10th annual conference on Internet measurement - IMC '10, 2010.

[4] W. Sherchan, S. Nepal and C. Paris, "A survey of trust in social networks", ACM Computing Surveys, vol. 45, no. 4, pp. 1-33, 2013.

[5] J. Lewis and A. Weigert, "Trust as a Social Reality", Social Forces, vol. 63, no. 4, p. 967, 1985

[6] Y. Kim, E. Jhee, J. Choe, J. Choi and Y. Shin, "A measurement model for trustworthiness of information on social network services", 2015 International Conference on Information Networking (ICOIN), 2015.

[7] M. McPherson, L. Smith-Lovin and J. Cook, "Birds of a Feather: Homophily in Social Networks", Annual Review of Sociology, vol. 27, no. 1, pp. 415-444, 2001.

[8] C. Cao and J. Caverlee, "Behavioral detection of spam URL sharing: Posting patterns versus click patterns", 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014), 2014.

[9] J. Weng, E. Lim, J. Jiang and Q. He, "TwitterRank", Proceedings of the third ACM international conference on Web search and data mining - WSDM '10, 2010.

[10] K. Thomas, C. Grier, D. Song and V. Paxson, "Suspended accounts in retrospect", Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference - IMC '11, 2011.

[11] C. Yang, R. Harkreader, J. Zhang, S. Shin and G. Gu, "Analyzing spammers' social networks for fun and profit", Proceedings of the 21st international conference on World Wide Web - WWW '12, 2012.

[12] Z. Chu, I. Widjaja and H. Wang, "Detecting Social Spam Campaigns on Twitter", Applied Cryptography and Network Security, pp. 455-472, 2012.

[13] C. Li, J. Yang and S. Park, "Text categorization algorithms using semantic approaches, corpus-based thesaurus and WordNet", Expert Systems with Applications, vol. 39, no. 1, pp. 765-772, 2012.

[14] G. Pirró, "A semantic similarity metric combining features and intrinsic information content", Data & Knowledge Engineering, vol. 68, no. 11, pp. 1289-1308, 2009.

[15] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi and M. Tesconi, "The Paradigm-Shift of Social Spambots", Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion, 2017.

[16] S. Asher, J. Parkhurst, S. Hymel and Williams, "Peer rejection and loneliness in childhood", S. R. Asher & J. D. Coie (Eds.), Cambridge studies in social and emotional development. Peer rejection in childhood, pp. 253-273, 1990.

[17] R. Selman, The growth of interpersonal understanding. San Diego (Ca.): Academic Press, 1980.

[18] I. Inuwa-Dutse, M. Liptrott and I. Korkontzelos, "Detection of spam-posting accounts on Twitter", Neurocomputing, vol. 315, pp. 496-511, 2018.

[19] K. Thomas, C. Grier, J. Ma, V. Paxson and D. Song, "Design and Evaluation of a Real-Time URL Spam Filtering Service", 2011 IEEE Symposium on Security and Privacy, 2011.