A Two-stage Approach for Detecting Spammers in Online Social Networks

Bandar Alghamdi[1,2] ,Yue Xu[1] and Jason Watson[1]

[1] Faculty of Science and Engineering, Queensland University of Technology Australia Brisbane City 4000, Australia.
bandar.alghamdi@hdr.qut.edu.au
{yue.xu, ja.watson}@qut.edu.au
[2] Institute of Public Administration Riyadh City 11141, Saudi Arabia
alghamdib@ipa.edu.sa

***Abstract*— The phenomenon of evolving behavior by spammers in social networks has received consistent attention from different researchers to combat this challenge. Twitter is an example of a micro-blogging where spammers take place and change their spamming strategies and behavioral patterns to evade detection. Several approaches have been put forward to fight this problem, nevertheless they lost their effectiveness. The main limitation of existing methods they use unified features to characterize spammers' behavior without considering the fact that spammers behave differently, and this results in distinct patterns and features. In this research project, we approach the challenge of spammer's behavior by utilizing the level of focused interest patterns of users to uncover the differences between spammers and legitimate user. We propose quantity methods using three topical features: topic entropy, standard deviation of topic distributions, and Local Outlier Standard Score (LOSS) to measure the change in user's interest and then determine whether the user has a focused-interest or a diverse-interest. We develop a framework by combining unsupervised and supervised learning to differentiate between spammers and legitimate users. The results of this experiment show that our proposed approach can effectively differentiate between spammers and legitimate users regarding the level of focused interest. It also demonstrates the similarity level between the explicit user's interest and implicit tweets content. Compared with other detection methods, our method has better performance. To the best of our knowledge, our study is the first to provide a generic and efficient framework to represent user-focused interest level that can handle the problem of the evolving behavior of spammers.**

***Index Terms*— Spam, Behavior, Spammers detection, User interest, Online social networks, Machine learning**

I. INTRODUCTION

Recent developments in the field of online social networks have led to the integration of OSNs into nearly all aspects of everyday activity; however, spammers take advantage of these services for malicious purposes. With the increase in the influence of OSNs among users, a large platform has been established that spammers use to spread spam messages [2]. In Twitter, Spam tweets refer to unsolicited tweets containing malicious links that direct victims to external sites containing malware downloads, phishing scams, drug sales, etc.[1] . Spammers utilize different methods in spreading spam content, either using compromised accounts with already established reputations and exploiting the inherent trust of these accounts to spread malicious messages [2, 3] or creating fake accounts that appear to be legitimate to mimic legitimate user behavior by posting spam content and normal content [4].

Existing approaches address the detection of spam and malicious content on social networks through the use of language patterns and content-based metadata [5, 6]. Some works employ the user's profile in detecting compromised and fake accounts [3, 7]. Some recent additions to the literature have offered valuable findings about spammers' behavior, using networked communities or developing a hybrid approach for spam detection using multiple views [8-11]. Further studies have discussed communities and cooperative spammers [10, 12] for spam detection. Although most of the aforementioned detection methods detect spammers, a major limitation is that they characterize spammers' behavior with unified features, without considering the fact that spammers behave differently, and this results in distinct patterns and features for different spammers with different purposes.

Topic-based features proposed by Liu et al [13] discriminate human-like spammers from legitimate users based on user's content interest, which is represented by the user's topic distribution. Liu used the same set of features to classify users as spam or legitimate. However, using only one set of features is insufficient to differentiate spam users from legitimate users because both spam and legitimate users can have focused or diverse information interests in terms of information content. Users with a wide scope of interests are called diverse users, and users with limited scope of interest are called focused users. The study in [13] indicates that spam users can have either very focused interests or diverse interests. However, their research was not specifically designed to analyse this overlap or to characterize spammers with focused interests to separate them from legitimate users whose interests are focused too. Likewise, spam and legitimate users with multiple topics of interest cannot be classified by their approach.

Before describing our study in detail, we will provide the motivation behind our work and the assumptions used in our approach. As mentioned earlier, evolving behaviors by

spammers on online social networks continue to be a big challenge. Most existing approaches characterize users on the basis of features that are used commonly for all spammers, whereas spammers change their spamming strategies and behave differently, which requires us to consider this difference. Our study was conducted under two assumptions.

Assumption1: Spammers can behave differently, and this results in distinct patterns and features that need to be considered. The assumption of there being different behavior models for spammers has drawn attention recently. Some approaches have been proposed for addressing the difference [14, 15] where the users' features are noticed across different groups of users. Their study indicates that some users interact with others with less mentions, whereas the users in another group use more hashtags. In our study, we assume that this pattern used by spammers must be reflected in some features that may be good for a certain type of spammer yet that is not applicable to another type. We extract different features to represent users in two different groups, focused user group and diverse user group. In each group, more effective features allow a more accurate classifier by applying classification techniques

Assumption 2: The integration of both content and profile features is effective to properly understanding users' behavior and interest through combining implicit and explicit information. We assume that there is a need to combine the relevant features of the user's profile and content messages from the user's interest perspective in order to obtain a comprehensive understanding of spammer behavior. Therefore, in this paper, we propose a feature which takes account of users' self-descriptions which explicitly reflect the user's interest and the relation to their tweets. This can be used as a unique feature for identifying spammers that mimic legitimate user's behavior.

In this paper, we propose to take user information interests as a key factor for spammer detection since the engagement of users in any activity is driven by their interests. In online social networks, users tend to post messages that are interesting to them. However, since spammers intend to propagate spam messages or malicious URLs, their interests change frequently so that they can exploit any event that is trending or that has active users. Therefore, users' information interest alone is insufficient for identifying spammers from legitimate users given the fact that spammers could be focused users or diverse users in terms of information interest. This has motivated us to deeply understand users' behavior in terms of topics of interest in order to separate users into different groups so that we can use different features to build a classifier for each group in order to classify spammers more accurately. In order to separately analyze users with different behaviors, we propose to split users into two different groups by using clustering techniques in terms of the scope of their content interest. To this end, we propose a novel two-stage approach to detect spammers in online social networks.

The purpose of the first stage is to separate users into two different groups: Focused-Interest who have focused information interests, and Diverse-Users who have diverse information interests. Three topic-based features are proposed to assess the focuses level of user's interest: topic entropy, standard deviation of topic distributions and Local Outlier Standard Score (LOSS) [13]. Based on the topical features, we uncover a clear distinction between focused-interest group and diverse-interest group using clustering algorithm. In the second stage, different sets of features are proposed for characterizing the users in the focused cluster and users in the diverse cluster separately. Based on the features for each cluster, a separate classifier can be generated. With this approach, spammers with focused interests and diverse interests can be more accurately classified.

The main contribution of this study can be summarized as follows:

- We propose a novel two-stage approach for detecting spammers. In the first stage, users are grouped into two clusters based on the content diversity of their posts, i.e., focused cluster and diverse cluster. In the second stage, a classification technique is used to classify spammers from legitimate users for each of the clusters.
- Based on the level of content diversity, we represent users in one cluster with the features that are different from the features used for the users in the other cluster. Therefore, the classification accuracy can be greatly improved
- We propose a new feature to represent users for differentiating spammers from legitimate users. The new feature measures the consistency between a user's self-stated interest and the content of the user's posts.

The remainder of this paper is organized as follows: Section 2 introduces rated work. Section 3 explains the proposed method: the novel application of focused and diverse user's interest based on topical features. Section 4 discusses the experiment, evaluation and results. Section 5 contains a discussion. Section 6 finishes this paper by presenting a conclusion and future work.

II. Related Work

Many studies have been conducted to investigate spammers' behavior in online social networks, and researchers have shown an increased interest in this regard. A survey of potential solutions and challenges on spam detection in online social network has been proposed by [16]. Previous works have focused on characterizing spammers' behavior using different features and approaches [6, 8, 12, 15, 17].

A considerable amount of literature has been published on spam detection using content-based features [5, 14]. The statistical analysis of language, such as linguistics evolution, self-similarity and vocabulary, are the primary features used for spam detection. Although they perform well in detecting spam tweets, their limitation relies in the fact that content features alone cannot be used to properly analyse spammer behavior. Spammers mainly utilize the tactics of mixing normal tweets

and posting heterogeneous tweets. Therefore, the inclusion of other features than content features would result in higher accuracy in detection and would provide extensive range in understanding spammer behavior. Some works consider only user's profile [18] to detect spam users, without considering content features due to the idea that this is a fast and effective way. They capture users' behavior and identify certain patterns from the profile to detect spammers and compromised accounts. Alternatively, [19] combined profile features with some content features to identify suspended accounts and spam campaigns. [7] has reported that spam and non-spam profiles overlap, which can make it a challenge to identify spam users across a network. However, certain characteristics are noticeable among spam profiles, including young accounts, tweets with a higher succession rate, tweets with greater status and tweets that contain spam words. However, these studies were limited to characterizing spammer behaviors in regard to a few aspects, and they showed a lack of classification accuracy as their approaches are not sophisticated.

It has conclusively been shown that combining content and profile features provides a comprehensive understanding of spammer behavior [8, 11, 17]. They determine that there is a strong and consistent correlation between the profile and content for all suspicious accounts. Such a combination shows the fundamental characteristics of spammers from different views and provides a different level of detection rates. [8] proposed dynamic metrics to measure the change in user activities and to identify abnormal behavior with a combination of some user profile features. [6] presented a detailed analysis of 14 million tweets with a focus on hashtags and tweet content. They observed that spam detection at tweet-level can be made more accurate by combining user- level. In our present study, we extend this combination by considering content features and user demographic data with a focus on the user's interests.

The social graph is one of the most widely used approaches for spam detection [8, 10]. In social networks, users are connected with each other to form network communities that share similar characteristics, such as interests, location or past common history. Analysing the underlying structure of the network community provides insight in detecting the outlier or spammer that drifts from the community or that behaves abnormally. Despite the efficacy of this method, analysing community networks requires effort and time, and spammers work cooperatively to form communities that are a challenge to identify through the network graph approach [12]. In addition, spammers change their behavior and strategies to evade detection [20], which makes this technique not very effective.

To meet the challenge of the evolving behavior of spammers, subsequent approaches have been proposed using topics features to detect spammers. [21] introduced word-, topic- and user-based features, using the Labeled Latent Dirichlet Allocation (L-LDA) model to model discriminated topics and words to detect spam comments in YouTube comments. Another study by [13] performed an experiment using the standard Latent Ditichlet Allocation (LDA) approach to measure the degree of change in user's interest to detect human-like spammers. After generating a number of topic probabilities for each user, they calculate two topical features: Local Outlier, which captures the user's interest, and Global Outlier, which reveals user's interest in comparison with the interests of other users. The results of this study indicate that spam users either concentrate on certain topics or have interests in some topics. Similarly, legitimate users mainly focus on limited topics. The main limitation of this study, however, is that they did not provide a separation between legitimate and spam users who have focused on different topics or who have focused on certain topics; also, the topical features proposed by this study showed a low detection rate when we applied them without the integration of profile features.

Alternatively, [12] proposed a distinctive approach using retweeting behavior to discover anomalous topics among trending topics on Twitter. Their aim was to detect cooperative spammers who hijacked topics by analysing the change in the topology of characteristics of their retweeting networks. Another sophisticated approached offered by [10] to detect malicious messages is by inspecting the way in which the messages spread on online social networks. Nilizadeh et al [10] identified different communities that share similar topics of interest and inspected the dissemination path to predict the pattern of posting within and outside of the community in order to detect malicious messages. They argued that each community has normal messages between members within that community, reflected by intra-community communication and inter-community exchanges between structural communities, and malicious messages do not match these normal message patterns. Nevertheless, this approach is scalable and successfully detects spam messages; it involves multiple phases that make it complicated. The study would have been less complicated and more efficient if it had considered user's interest rather than community interest.

The limitations of the above approaches are that they do not address the critical issue, as they still characterize spammers with unified features, whereas spammers behave differently and should be represented by different features too. To address this problem, we propose a metric to describe changing patterns in user's interest and develop a detection method for spammers. In our present study, we extend the combination of content and profile features by considering user's interest as a novel way that is different from previous works.

III. Proposed Method

We propose a two-stage approach: unsupervised learning stage, and supervised learning stage. The components of our approach are shown in Fig. 1. Stage one has the following components :1)- Modelling users' interest, 2) topic-based features extraction, 3) clustering of focused and diverse users. Stage two consists of: 4) feature extraction for representing users in each cluster, 5) classification of spam and legitimate

users. The idea behind this approach is to be able to model focused and diverse users who usually behave differently. The crucial part of this approach is to extract appropriate features for clustering and for classification, especially the different features for the users in the two different clusters. We propose to model users' information interests using topics generated from users' posts by using topic modelling techniques. In the following sections, we explain each component in detail.

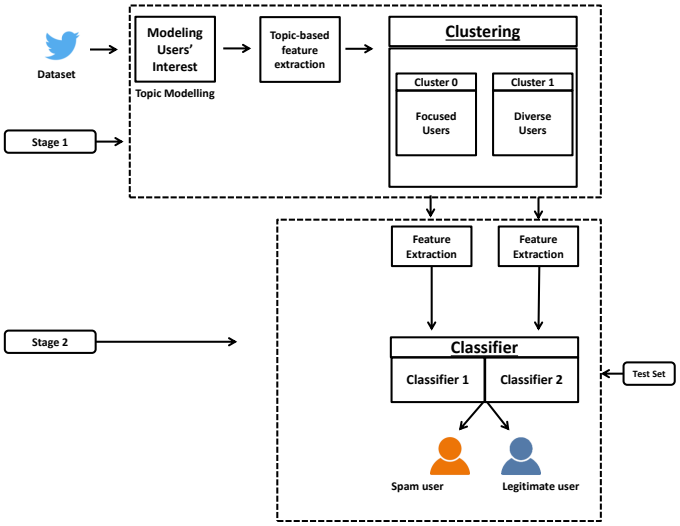

Fig. 1: Proposed Framework: Spam user detection through user's interests, employing unsupervised and supervised machine learning to classify spam users.

A. *Unsupervised Learning Stage*

1) Modelling user's interests based on LDA topic models: Latent Dirichlet Allocation (LDA) was first introduced by Blei [22] as an example of a topic model. Each document d_i is represented as a bag of words $W = \{w_1, w_{2.....}w_M\}$, and M is the number of words. Each word is attributable to one of the document's latent topics $Z = \{z_1, z_{2.....}z_k\}$, and k is the number of topics. φ_j is a multinomial distribution over words for topic z_j, $\varphi_j =< p(w_1|z_j), \dots\dots, p(w_M|z_j) >$, $\sum_{i=1}^{M} p(w_i|z_j) = 1$. φ_j is called the topic representation for topic z_j. θ_i is another multinomial distribution over topics for document d_i. $\theta_i =< p(z_1|d_i), p(z_2|d_i), \cdots, p(z_k|d_i) >$, and $p(z_j|d_i)$) indicates the proportion of topic z_j in document d_i. θ_i is called the topic distribution for document d_i.

We considered each user's tweets as one document. The document collection contains all users' tweets. The user's information interest is reflected in the tweet content, and we need to model the user's interest using LDA. So, we apply the LDA Topic Model to generate k topics for each user and get the topic probabilities for each single user. From these topic distribution values, we extract three topic-based features, which are discussed in next section, to measure the user's interest in order to distinguish between users who have focused interests and users who have diverse interests.

2) Topic-based features to depicting users' interest focus level: The separation of users based on interest concentration is motivated by the observation that users with focused interests should have different features from those who have diverse interests. In this paper, we propose to cluster users into two groups based on their content interest described by topic distribution generated from their tweets. Clustering in this research project is different from previous studies [8, 23], as previous studies used clustering techniques to group spammers into a cluster with similar spamming behavior, whereas in this research we utilize clustering to identify focused users and diverse users, and then we extract features that are more representative for each cluster for spammer detection. The following section details three topic-based features for clustering users as focused and diverse: Topic Entropy, Standard Deviation of Topic Distribution, and Local Outlier Standard Score.

We mentioned previously that user's interest is a reliable feature that is difficult for spammers to evade and that can therefore be used for detection. After generating topics from users' documents by using LDA, we used topic entropy to measure the diversity of topics for each user, using the following equation:

$$\mathrm{H(u)} = -\sum_{i=1}^{k} p(z_i|u) \log_2 p(z_i|u) \quad (1)$$

$p(z_i|u)$ is the topic distribution for user u. A user with low topic entropy is more likely concentrated meaning that the user is interested on limited topics, while a user with higher entropy is more likely to have wide interest spreading somewhat evenly over many topics. $H(u)$ can help us later in the clustering stage to get users with different levels of focus. The example in Table 1 shows two users with different values of entropy and topic distributions, where $k = 5$. The two users are from the Honeypot dataset [24] which is a popularly used labelled spam dataset for spam detection research. User 1 has higher entropy values comparing with User 2, and the topic distribution for this user shows that this user looks interested on many topics. User 2, however, **has** very lower entropy value, and the topic distribution shows that this user has very strictly focused on Topic 0, which can clearly indicate by the uneven topic probability distribution.

Standard deviation of topic distribution is also a good indicator for differentiating focused and diverse users in

TABLE I: Two users have different topic distributions, entropy and standard deviation.

	Topic Entropy	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Std
User 1	**1.604808**	0.203534	0.199607	0.187173	0.176047	0.23363	0.021694
User 2	0.015792	0.998186	0.000454	0.000454	0.000454	0.00045	**0.446199**

addition to the topic entropy. Using the standard deviation, we have a 'standard' way of knowing how spread out the topics are from the mean of a given user. This demonstrates the degree of change in topics for a particular user. From Table 1 we see that User 2 has a higher degree of deviation (Std) than User 1.

Local outlier standard score was first proposed by [13] to discriminate human-like spammers from legitimate users using topic distribution. Liu used this feature for the purpose of classifying spammers and legitimate users, but we use it for clustering users into focused and diverse groups. This feature measures the degree of interest of user in respect to a certain topic, using the following equation:

$$\mu(u_i) = \frac{\sum_{j=1}^{k} p(z_j|u_i)}{k}$$

$$LOSS(u_{il}) = \frac{x_{il} - \mu(x_i)}{\sqrt{\sum_{j=1}^{k}\left(x_{ij} - \mu(x_i)\right)^2}} \quad (2)$$

Where, $\mu(u_i)$ is the average interesting degree for all topics for a certain user. If we extract k topics for each user, we will end up with a vector of k features for each user, $LOSS(u_{i1}), \ldots\ldots, LOSS(u_{ik})$. In our experiment discussed in Section 4, we generate 5 topics for each user, we will have a vector of 5 LOSS features for each user.

In this paper, we propose to use topic entropy, standard deviation of topics distributions, and the vector of LOSS values as the features to cluster users. Next section will discuss the clustering process to cluster users. Since clustering is a typical type of unsupervised learning technique, this clustering process is considered as the unsupervised learning stage in the proposed spam detection approach.

3) User clusters with focused interest and diverse interest: The purpose of using topic-based features in this research is to model user's interest and then use this to identify two groups of users in terms of their interest concentration: focused user (who mainly is interested in a few topics) and diverse user (who have a wide range of interests).

A common opinion is that spammers do not have clear information interest and thus their tweets involve a wide range of topics meaning that they show diverse information interest. Some existing classification-based detection methods [6, 17, 25] use the number of hashtags as a feature to classify spammers from legitimate users because it is considered that spammers use more hashtags than legitimate users. However, in reality, some spammers could be focused such as content polluters for promoting some specific commercial product. To get an idea of the importance of level of interest concentration for the purpose of spam detection, we provide in Table 2 an example of two spam users having different levels of interest concentration. For example, User 1 shows one topic of interest, which is "American Football" across all tweets, and a link associated with each tweet. User 2 on the other hand, has diver interest with @mention and hashtags in most of the tweets. Both users are spammers, but they show different features and behavior, and if we consider, for example, the number of @mention or the number of hashtags as features to differentiate them, we will have an error of misclassification, because these features may not be applicable to classify User 1 as a spammer.

It would be ineffective to look for unified features of spammers to detect smart spammers, and it would be more useful to analyze them from the interest-level perspective to extract the most effective features to detect spammers properly.

TABLE II: Example of two spam users with different messages and interest.

User 1	**"**American Football. NFC North Winner, Divisional Markets. Detroit Lions is decimal odds of 18.5 to win. [_RUL]" "American Football. AFC North Winner, Divisional Markets. Cincinnati "Bengals is decimal odds of 4.1 to win. [_RUL]" "American Football. AFC South Winner, Divisional Markets. Houston Texans is decimal odds of 5.5 to win. [_RUL]" "American Football. Super Bowl Winner, NFL Season 2010-11. New "Orleans Saints is decimal odds of 13.0 to win. [_RUL]" "American Football. AFC Conference Winner, Conference Markets. Baltimore Ravens is decimal odds of 8.4 to win. [_RUL]"
User 2	"RT @beisick306: Custom @Lemarvelous23 #NTD #S-10 #calgarystampeders [_RUL]" "@yasminhasann6 I'm so Wavesy [_RUL]" "When u combine wine and dinner the new word is winner" "RT @DRUGRANGE:DRU - Don't Be Afraid Teaser [_RUL] via @YouTube" "RT @DRUGRANGE: #NowPlaying [_RUL]" "I don't want all these other apps to have snapchat, too much stuff"

Therefore, it is desirable for this research to establish a way to determine focused users and diver users first before classifying spammers from legitimate users. We want to examine the effects of using multiple topics generated from each user's tweets and then quantify the change in these topics with different degree assessments to demonstrate that such assessments may reveal an important difference between focused-interest and diverse-interest users.

Based on our observation we find that any of the above three topic-based features alone is not sufficient in identifying focused and diverse users when we have somewhat low variance between topic probabilities. Therefore, we combine them together to construct a unified feature vector for each user. Formally, for each user u_i, let k be the number of topics, we can calculate a total of $N = k + 2$ topical features which include k $LOSS$ features $LOSS(u_{i1}), \ldots\ldots, LOSS(u_{ik})$, topic entropy $H(u_i)$ and standard deviation $Std(u_i)$. Each user is represented by a N-dimensional feature vector $V_i =< v_{i,1} \ldots\ldots v_{i,N} >$.

From the tweet dataset Honeypot [24], we generate a topic model with k = 5, then generate the topical features based on the topic model for each user in the dataset. By applying a clustering method, we generate two clusters based on the topical features. The results in Figure 2 indicate that both clusters contain spammers and non-spammers. Table 3 shows the average values of each feature over the users in each cluster.

TABLE III: Average values of each feature over the users in each cluster (K=5).

Attribute	Cluster0	Cluster1
Topic 0 LOSS	0.1126	0.1774
Topic1 LOSS	0.7051	0.2255
Topic 2 LOSS	0.0617	0.1644
Topic 3 LOSS	0.0247	0.2092
Topic 4 LOSS	-0.2888	-0.3005
Std of topics dis	0.2987	0.2321
Topic Entropy	0.8468	1.0584

Based on the average feature values, we can decide that the users in Cluster 0 are more focused than the users in Cluster 1. This is because Cluster 0's topic LOSS distribution is much uneven than that of Cluster 1, and Cluster 0's topic entropy is less than that of cluster 1, where Standard deviation of cluster 0 is higher than that of cluster 1. All these comparisons indicate that Cluster 0 contains focused users while Cluster 1 contains diverse users. Table 3 shows the overall average values for each feature of the clustering output. The concentricity and diversity of the two clusters are further discussed in the next subsection.

The size of a cluster is the number of users in the cluster. According to our observation of the clusters, we have 2540 users in Cluster 0, with a total of 2263 legitimate users and 287 spam users, which shows that the distribution of legitimate and spam users is unbalanced with 94% of the users in Cluster 0 being legitimate and only 6% being spammers. Even the number of spammers is low, but it also shows that spammers can be of focused.

In contrast, Cluster 1 is a balanced cluster, which composes of 2891 spam users and 3611 legitimate users as showed in Fig. 2, which indicates that a diverse user could be a legitimate user or a spammer with similar probability. The result in Cluster 1 shows that legitimate users tend to have diverse interests. For the spam users in this cluster, they are compromised accounts or fake accounts that randomly try to mimic legitimate account behavior to avoid detection by Twitter. As this study set out with the aim of assessing the importance of focused level, we want to distinguish between spammer and legitimate users for both groups. We hope that we can represent each group with divergent features and then utilize this to classify spammers and legitimate users.

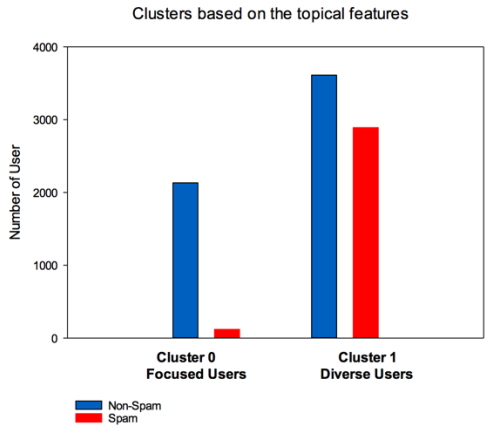

Fig. 2: Two clusters of users based on topical features.

Topic entropy can help to determine how focused a user is on different topics. In order to make this feature easier to understand, we presented early two distinct users in Table 1. We can see that User 1 has high entropy values, and the document topic distributions for this user show that this user does not have focused topics, whereas User 2 has very low entropy value, and document topic distributions show that this user has a very strict focus on Topic 0. For the two clusters generated from the Honeypot dataset, as showed in Table 3, the average value of entropy for Cluster 0 is 0.8468 and 1.0584 for Cluster 1, which indicates that the users in Cluster 0 are more focused than the users in Cluster 1.

The standard deviation of topic distribution shows how much the topic of a given user differs from the mean value for the other topics for the user. This standard deviation with topic entropy can measure the change pattern of user's topics. Fig. 3

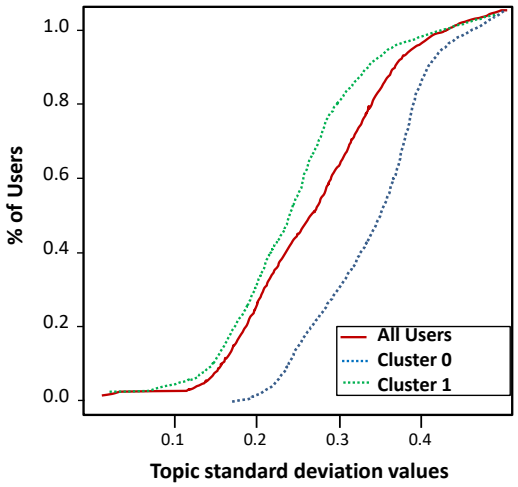

Fig. 1: Cumulative distribution function of standard deviation of topic distribution for each cluster

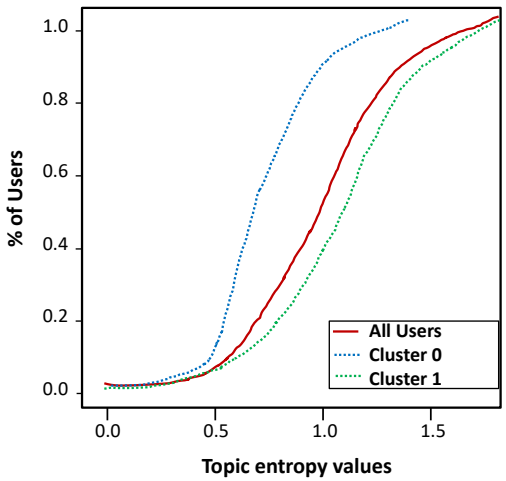

Fig. 2: Cumulative distribution function of topic entropy for each cluster

and Fig. 4 show the cumulative distribution function (CDF) of standard deviation and topic entropy for both Cluster 0 and Cluster 1. From the values in both figures, we can see that for the same percentage of users, the topic standard deviation of Cluster 1 is always smaller than that of Cluster 0, and the topic entropy of Cluster 1 is always larger than that of Cluster 0. Therefore, Cluster 1 contains diverse users and Cluster 0 contains focused users. The focused users have higher standard deviation values than those who are diverse as Fig 3 shows.

For LOSS feature, we use this feature to measure the degree

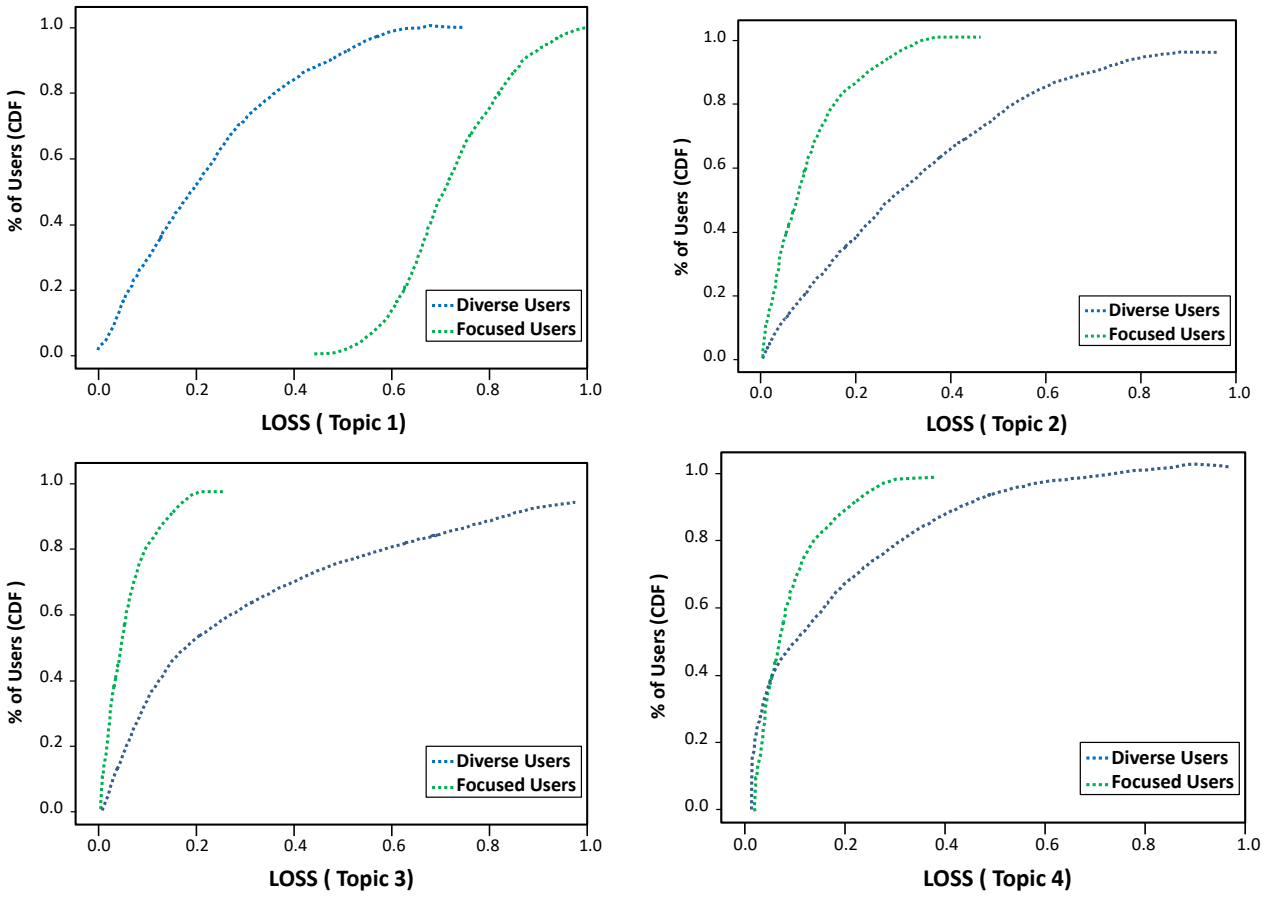

Fig. 3: Cumulative distribution function (CDF) for LOSS features grouped by cluster type.

of user's interest with respect to the 5 topics that we generated for each user. Fig 5 shows cumulative distribution function (CDF) for LOSS features for each cluster. It is clear that LOSS is also able to discriminate the two type of users, focused and diverse interest.

A. *Supervised learning stage*

While clustering provides a division of the observed topics and divides users based on focused and diverse interests, classification will be used to identify spammers in each of the clusters. As mentioned above, users in both groups, i.e., focused-interest and diverse-interest, can be spammers or legitimate users. For differentiating spammers from legitimate users, in this paper we propose new features based on users' self-description which describes what they are interested and their post content. We also use other existing features to represent each group.

1) *Features for classification:* As described in previous sections, the first stage of our proposed approach is to divide users into two clusters, focused-interest and diverse-interest, based on users' topic distribution. In the classification stage, we proposed a novel feature to differentiate spammers from legitimate users based on users' self-description and their posts content. This feature represents the consistence between a user's self-description and his/her posts content. For the classification, we also use some of the existing features proposed by different researches, which are listed in Table 4.

In social networks such as Twitter, users are allowed to describe their interest in the description statement contained in their profile. The users' descriptions provide explicit information about their interests, whereas the users' tweets reveal their interest implicitly. As the aim of this study is to detect spammers, we try to understand spammers' behavior through explicit and implicit behavior from user's interest point of view. The comparison between a user's description and his/her tweets can help in understanding their behavior. Therefore, the user can be assessed by analyzing the user's description in relation to the tweets that they have posted. The description in a user's profile in Twitter is called a Bio, which gives the users up to 160 text characters to tell the others about themselves, plus 30 bonus characters for location as well as an opportunity to give a backlink to their own site. This field contains a few simple sentences describing the user's interest to give people a first impression about a user, and many attackers use it to attract more followers. All existing studies so far, however, failed to take the content of Bio into consideration for detecting spammers, which we cover in this research.

The basic assumption about the user's description is that a user's self-description is generally consistent with the content of the user's posts. It is worth to mention that this assumption doesn't mean that a user's self-description should match every single tweet he/she posts. For example, if a user's description states that she/he is a "specialist in children with special needs", this user should have a large number of tweets that relate to the topic of "children with special needs", but every single tweet does not have to necessarily match with the description.

We propose a new feature that is the similarity between a user's self-description (i.e., user Bio) and the user's posts or messages such as tweets. We propose this feature to find the relationship between users' explicit statement in their profile and implicit behavior in their tweet content. This feature reveals how consistent the users' self-stated interest is with the interest showed in their tweets. Spammers often create a fake account or use a compromised account and they try to mimic legitimate behavior to avoid detection. However, because spammers aim to spread unsolicited or harmful messages to as many users as they can, very often the content of the messages does not match their self-description. The primary hypothesis of this feature was described earlier in Assumption 2 that the integration of both post and profile features are effective to properly understand users' behavior. We believe that this behavior of inconsistent interest must be reflected in the evolution pattern of the tweets content and could help detecting spammers.

Cosine similarity can be used to measure the similarity between a user's self-description and the tweet content represented as vectors. Given two vectors and the cosine similarity is calculated as follows, where A and B are vectors representing the user's self-description and the tweet content:

$$Sim(A,B) = cos(\theta) = \frac{A * B}{\|A\|\|B\|}$$

We generated a vector for the self-description of each user, and a vector that represents each of the user's tweets. We used term frequency–inverse document frequency TF-IDF values to produce these vectors, where each tweet and each user's self-description is treated as a document. We then calculated the similarities between a user's description and each of the user's tweets and got the average similarity to represent the interest consistency of this user. Usually most of a legitimate user's tweets are relevant to his/her interest described in his/her profile. However, this type of behavior is not always pretested in spammer behavior because most spammers do not have clear information interest, or they use compromised accounts. For this feature, legitimate users showed that averagely the content of their posts has relatively higher similarity to their self-description in comparison with that of spammers. The following existing features proposed in [6, 8, 13, 17] are also chosen in the classification stage.

2) Number of unique words: It has been proved that legitimate normal accounts are more innovative in their use of language, while spammers may repeat themselves more often, since they usually have a specific agenda or target to achieve [14]. The unique words feature can reflect the innovative pattern of using language for a particular user, but it is important to note that this is not applicable to all spammers. In our dataset,

we have found that a number of legitimate users in Cluster 0 (focused users) somehow exhibit similar use of the same word and do not post a large number of unique words. This is why we cluster users into two different groups based on the topical features to characterize each cluster from topics point of view. The number of unique words is more effective for diverse users than for focused one, because users are interested in different topics and the use of unique words can differentiate spammers from non-spam users. For the focused user, however, this feature is not applicable, as all users in this group have almost the same behavior of using words and posting similar tweets most of the time. Although users in the diverse-interest group have different interests and normally use new words, spammers have limitations in the use of unique words, which is a significant feature when distinguishing spammers from legitimate users in this group.

2) Average count of "@username" per tweet: The insertion of @username is essentially used to deliver the tweet to the username's account, even if the user has no relationship with the intended target. This is very common behavior by spammers and has been examined in previous studies [24, 26]. Users with this type of behavior use @username in order to attract new followers or to harm the user. In all cases, the use of @username is found to be a good feature for diverse- users, to discriminate spam from non-spam users. The reason for this is that users with focused-interest generally exhibit similar behavior of posting similar content and posting @username very often. [9] in their research, categorized harvested spammers into different groups such as duplicate group and promoters group, both of which make use of @ very often in their tweets. However, we found that most focused users (both spam and non-spam) have somehow similar behavior in this regard with relatively small notable differences between them, and this feature is more useful for diverse-users than focused-users.

3) Number of links: This feature is very similar to the use of @username for both groups. Among focused users, spammers and non-spammers post a large number of links to target users. If we consider this feature as a unified feature for spam detection, we would have misclassification of legitimate users who have focused-interest. We have noticed that there are a number of bots among focused users that post the same content with links that take users to disreputable web pages such as phishing sites or drag sellers. In contrast, diverse-interest users vary in term of using links, and spammers show a higher usage of links in their tweet than normal users do, and this feature is more applicable in diverse- users. We may also attribute the high usage of links among diverse-interest users to the idea that spammers exploit reputable accounts and seek to harm existing followers that already have a trusted relationship with the owner of the account.

4) Number of Following and followers: Number of following as feature can be used in both clusters, focused-interest and diverse-interest. The number of Following is abused by spammers to gain access to many targeted users. This behavior is a common characteristic of spammers and has been extensively used for spammer detection by [9, 10, 15, 27] and [11]. It is worth mentioning that the number of followers as a feature is not effective for in diverse users, whereas this feature has a high contribution to the focused users. As one of the contributions from this paper is to demonstrate that both groups, focused and diverse, are characterized by different features, the number of followers is not a significant feature in diverse-interest users. Spam users in the diverse group are hidden as legitimate accounts that have a good reputation and that do not seek to have more followers in order to appear as legitimate accounts. Interestingly, this feature is significant in focused users, as spammers tend to appear as legitimate users, and they use third parties to get more followers as Lee [9] reported that the number of following and followers fluctuated significantly over the time of the spam users. They lose or gain followers quickly, and this can be reflected in our findings that most of this type of behavior is among focused users, not diverse users.

In addition to this feature, three other features closely linked with the following and follower features are also significant for both groups, which are standard deviation of following, ratio of following and followers, and change rate of following [9]. These are temporal features that show how often a user follows others. In this paper, we confirm that these features are applicable to both focused and diverse users, with the absence of followers as a single feature for diverse users.

In summary, the approach of having two different clusters based on the level of focus interest suggested that characterizing spammers with distinct features for users in different clusters is the key point of detecting spammers with higher detection rate. The effectiveness of using different features for different clusters is demonstrated with some example comparisons showed in Fig 6. The top two figures show the comparison of the feature @username for the focused and diverse clusters using the cumulative distribution function (CDF) of the features representing focused users in left figure and diverse users in the right. We can see from Fig. 6_(a) that, for focused users, the feature value of spammers is lower than that of normal users for some of the users, but higher for some other users, indicating that the feature @username is not consistent over all users and thus is not effective for focused users. However, for diverse users, the feature value of spammers is consistently smaller than that of normal users, indicating that this feature can be used to differentiate spammers from normal users for diverse users.

The unique word feature show differences to the two clusters in the amount of unique words used by both groups. Although this feature shows that normal users are more innovative in their use of language [28] comparing with spam users in both clusters, it is more effective with diverse users. With the nature of diversity in the user's interest for diverse groups, there will be more unique words in their content, however spammers in this group still show less ability to use new words. For the focused cluster the number of spammers is much smaller than that of normal users, which might be why the value of normal users is almost the same as the average since normal users dominate. The figure shows that the unique word feature can be used for both clusters, but it is more effective for diverse group than focused group. The CDF curves of these sample features have proven the assumption that the focus level of user's interest

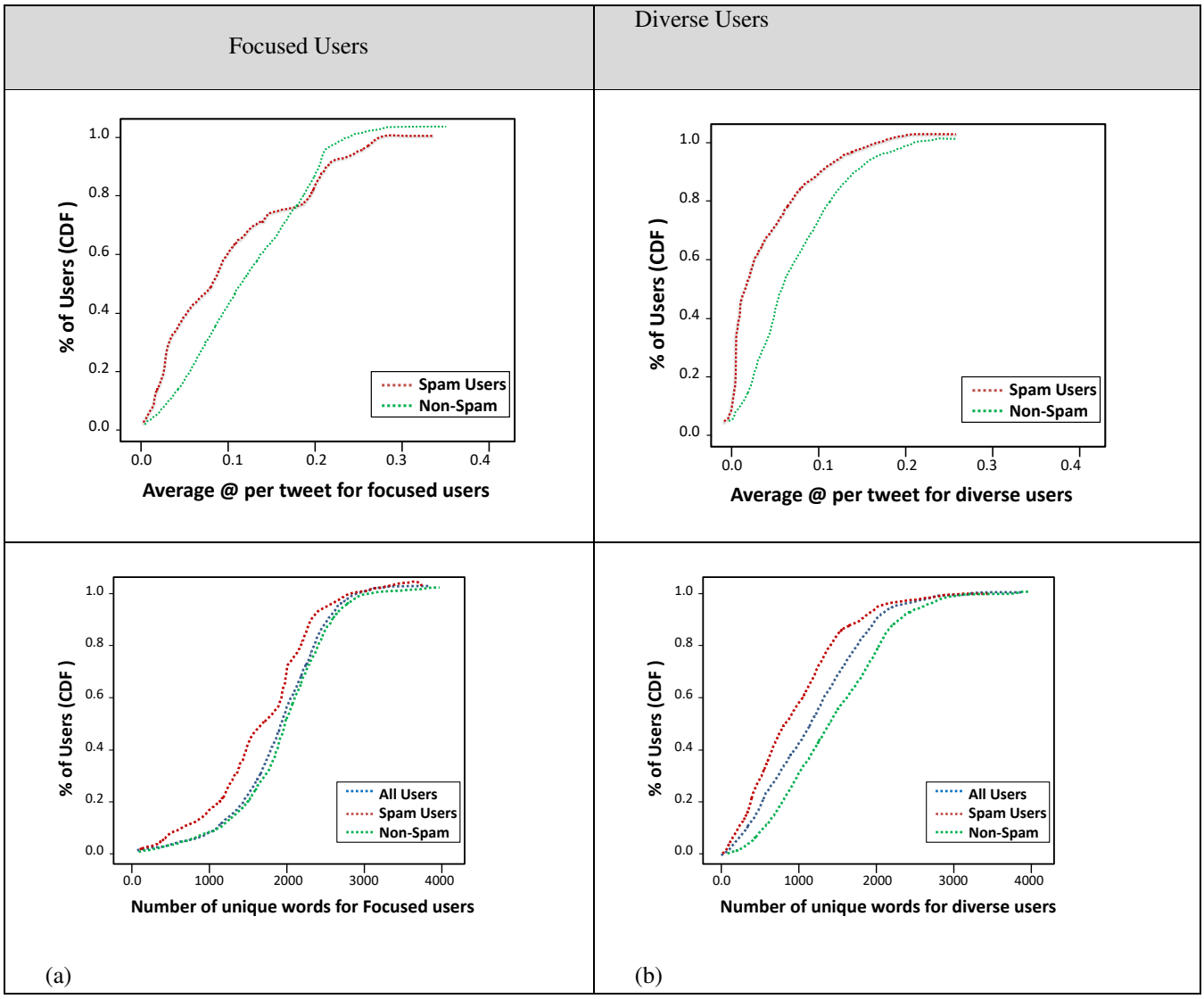

Fig. 4: The comparison between focused users and diverse users in terms of the content features

plays a central role in characterizing spammer behavior and content features are not effective in the focused group.

5) Feature selection: For our approach, an important task is to select the most effective features for each cluster. Here, we select features that best classify users using the correlation-based feature (CFS) algorithm [29]. Then a supervised machine learning module is adapted to train a classifier that is used to make a decision on each user in the testing dataset on whether the user is a spammer or a legitimate user. Table 4 shows the selected features for classification from existing features and our proposed features. We organized features into four categories: *content features*, such as unique words, number of links and number of @ signs; user *demographic features*, such as number of followers and number following; topical features such as LOSS features that has been used for clustering stage; and our proposed feature, the consistency of user interest. We assume that some features are not suitable for focused users while they are effective for diverse users. By using the CFS algorithm, the most effective features for each cluster are selected. This algorithm evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. The subsets of features that are highly correlated with the class while having low inter-correlation with class are preferred [29]. Irrelevant features should be ignored because they will have low correlation with the class. The CFS's feature subset evaluation function is as follows:

$$Ms = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \quad (3)$$

Where Ms is the correlation between the class and the features in feature subset s containing k features. $\overline{r_{cf}}$ is the mean feature-to-class correlation over the features in s (i.e., $f \in s$, c is a class), the higher the better. $\overline{r_{ff}}$ is the average inter-correlation between features in s, the lower the better. Overall, the higher the Ms, the better the feature set s is selected. We used this evolution algorithm to select the best features that have high correlation with the class and low inter-correlation between the features in each cluster. _ Since the aim of clustering is to divide users into different groups according to their content interest, the clustering stage was designed based on topical features derived from tweets content. As a result, we expect the features for each cluster, especially the content features, to be potentially different, and we also expect there might be overlaps in user demographic features.

Table 5 shows the chosen features for representing the users in each of the two clusters. From Table 5 we can see that both content features and demographic features are chosen to represent the diverse users in cluster 1, which means that both types of features are effective for diverse users. However, for the focused users in cluster 0, only user demographic features are chosen and none of the content features seems effective to be used to represent focused users. Both clusters contain spammers and legitimate users as well. This confirms our

TABLE IV: Features selected for each cluster using the CFS filtering algorithm.

Reference	Category	Features	Cluster 0 (focused users)	Cluster 1 (Diverse users)
[6] [8, 24, 17]	**Content features**	Num of hashtag	—	—
		Num unique word	_	✓
		Num links	—	✓
		Num unique links	—	—
		Num of at@	—	—
		Num of unique at@	—	—
		Aver links/tweet	—	—
		Aver unique link/tweet	—	—
		Aver at@ per/tweet	—	✓
		Aver unique at@/tweet	—	—
[17] [24]	**Demographic Features**	Num of followers	✓	—
		Num of followings	✓	✓
		len about me	—	—
		len username	—	—
		Std following	✓	✓
		Ratio following & followers	✓	✓
		Change rate of following	✓	✓
Our proposed features	**Consistency of User's interest**	Max of Similarity	—	—
		Ave of Similarity	—	✓
		STD of Similarity	—	—
		Min of Similarity	—	—
Our proposed features	**Topical Features**	Topic Entropy	**Used for clustering stage**	
		Std of Topic Distributions		
[13]		LOSS 0		
		LOSS 1		
		LOSS 2		
		LOSS 3		
		LOSS 4		

Assumption 1 that spammers behave differently, and this results in distinct patterns and features. In general, our results indicate that the variances of features between different types of spammers are existing and need to be considered. In Fig. 6, we showed samples of features that are effective for cluster 1 but it is not suitable for cluster 0 conversely. [9] and [14] point out that the strength of classification lies mainly in the choice of features, and we try to model this phenomenon utilizing the level of focused interest in our current research project in order to detect spammers, with higher rates of detection.

TABLE V: The chosen features for the two clusters.

	Cluster 0 (Focused interest)	Cluster 1 (Diverse interest)
Demographic Features	Number of Followings. Number of Followers. Std of Following. Ratio of Following and Followers. Change rate of following.	Number of Followings. Std of following. Ratio of Following and Followers. Change rate of following.
Content features	None	*Number of unique words.* *Average at@ per tweet.* *Number of links.* *Average value of cosine similarity.*

As we have established the importance of correlating demographic and content features for spammer detection, separating users based on the level of interest diversity supports this association. Nevertheless, the selection of the most discriminative features is necessary in order to detect spam users. Demographic features play a key role in characterizing spammers' behavior for focused users, while on the other hand, content features in line with demographic features are more suitable for diverse users to uncover spammer behavior.

6) Spam detection by classification: Using the selected features for the focused user cluster and the diverse user cluster, a separate classifier can be constructed by applying a classification algorithm for each of the clusters, as shown in Fig. 1. The two classifiers together form an overall classifier which can be used to classify new users into spammers or legitimate users. Based on the similarity between a new user's feature vector and the centroid of the focused cluster and the centroid of the diverse cluster, the user can be considered as a focused user if he/she is more similar to the focused cluster, a diverse user otherwise. Then the corresponding classifier will be used to determine whether the user is a spammer or a legitimate user.

IV. Experiment and Evaluation

In this section, we first describe the implementation of our detection approach. We then introduce the dataset and the ground truth for evaluation. For the evaluation, we conduct several empirical studies to reveal the difference between spammers and legitimate users in terms of topical evolution patterns and some existing features. The results are found to conform to our assumptions. Finally, we evaluate the performance of our spammer detection method using the standard metrics.

A. *Overview*

Fig. 1 illustrates the framework of our proposed method. After we extract topical features of users' tweets content, we cluster users into two different groups, focused and diverse. Then we apply feature selection to select the most effective features for each cluster, as described in Section 3.A.5. We use these features to train our supervised learning classifier. Note that we use Weka machine learning framework [30] to conduct the experiment and evaluation. We use default values for parameters of the chosen clustering and classification methods, and 10-fold cross-validation where the original sample dataset is divided into 10 sub-sample sets, and 10 training and testing steps are performed. For the training, nine sub-sample sets are used, and the remaining sub-sample set is used for testing. The final evaluation result is the average of the 10 testing results.

B. *Data Set*

We chose the Honeypot dataset [24], which uses 60 honeypot accounts in Twitter to attract spammers and crawl any account that follows them. The data was collected from December 30, 2009 to August 2, 2010. We used the profile IDs in this dataset to crawl users' descriptions for each profile. It is worth mentioning that the dataset was reduced due to the limited number of users with descriptions in their profile or limited tweets that were not enough to understand the user's interest. We ended up with 9750 spam users, and 7167 legitimate users.

Before directly conducting the experiment on the employed dataset, we performed pre-processing steps. This involved deleting accounts that had few tweets, because a sufficient number of tweets are necessary to extract information on the user's interest and topics. Each of the remaining users has at least 20 tweets. We removed punctuations, stopwords and non-ASCII words and applied stemming. The ultimate dataset contained 5875 spam users with a total of one million tweets, and 3178 non-spam users with 572,040 tweets.

C. *Clustering*

In the first step of the experiment, we considered each user's tweets as one document and generate 5 topics from user's tweets document collection, then calculated topical features based on the topic model described in Section 3.A.2. We built the feature vectors with 7 features for clustering the users in our dataset. We used K-mean algorithm [31] and clustered users into two different groups. K-means is an iterative technique using a centroid-based method that takes the number of instances around which the clusters are built. Instances are

assigned to clusters based on similarities or distances. To evaluate the quality of the clustering result and verify that our clustering process can effectively divide focused and diverse users into different clusters, we collected statistics on the fractions of topic distributions. Differences between focused users and diverse users are shown using the topical features.

In general, our results indicate that the standard deviation of user's' topic distributions is higher for focused users than for diverse users because focused users have uneven topic distributions, while their topic entropy is low because of the same reason. Fig. 7 shows the comparison using the Honeypot dataset. From Fig. 7, we can see that the average topic probabilities (i.e., topic distribution) for focused users indicated in blue are very different, e.g., the probability of topic 1 is close to 1 and the other 4 topics have very small probabilities. In contrast, the probabilities of the 5 topics for diverse users indicated in red are very similar, all around 0.2. The mean value of topic entropy for focused users is found to be approximately 0.8468, whereas for diverse users it is 1.0584.

The decision to categorize a user as focused or diverse needs an effort and cannot be quantified easily. However, our proposed topical features provide quantity values which indicate that the topic interests of diverse users are more uncertain than those who are focused users because diverse users' topic probabilities are evenly distributed as Fig 7 shows. This behavioral difference between the two types of users is clearly represented by the quantity values in the topical features.

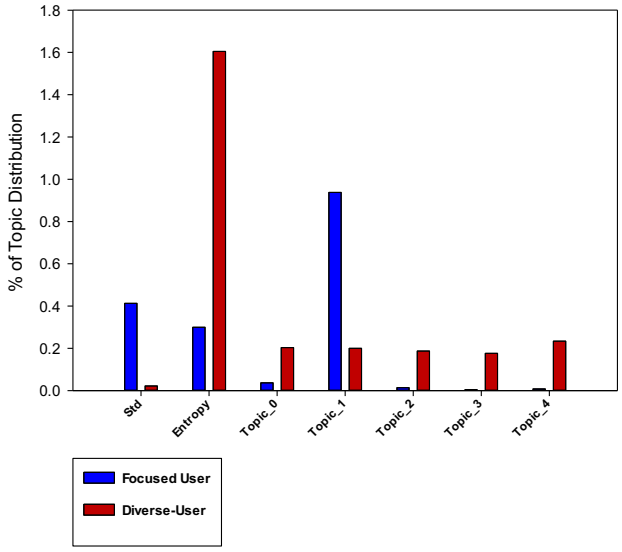

Fig. 7: Comparison between focused users and diverse users in terms of topic distribution, topic entropy and standard deviation of topics distributions.

In the clustering stage, three topical features are used: topic entropy, standard deviation of topic distribution and topic distribution. We observe that although users may have two topics of interest, their distributions can be entirely different and not close to each other. This type of user still has a concentration on one topic, with a notably higher value than for other topics. For example, user 2 in Table 6 has concentration around 71% on one particular topic, yet the user also shows interest in other topics, but with less concentration. Diverse users primarily have different topics of interest with different levels of focus, as user 4 shows in Table 6. We further calculated the topical change rate for each user obtained by topic distributions as follows:

$$topical\ change\ rate = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}|t_{i+1} - t_i|}$$

where n is the total number of topics, and ti is the topic distribution values. Most of focused users center on the vicinity of the average change value (i.e. 0.14), whereas diverse users are primarily distributed in a higher values (0.22) as boxplot shows in Fig 8.

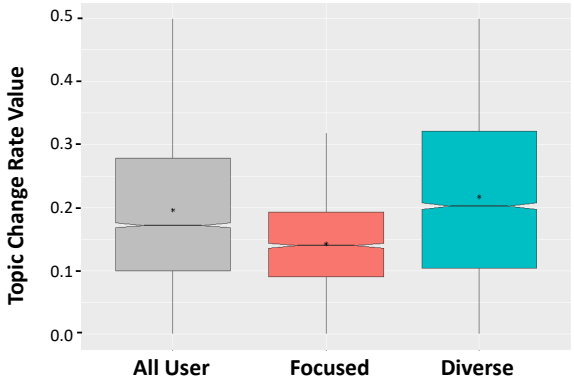

Fig. 8: Topic change rate grouped by clusters where diverse users experience higher change rate than focused users.

This verification experiment successfully reveals the difference between the two kinds of users in terms of focused and diverse interests. The result is roughly consistent with our assumptions and makes an excellent foundation for our subsequent experiment. By analyzing our clustering results, we conclude that our clustering-based features and process can distinguish focused and diverse users effectively.

V. Spam Detection Evaluation

In this section, we evaluate the performance of the proposed two-stage spammer detection approach and the contribution of the proposed two features. For the evaluation metrics, accuracy, precision, recall and F1-score are used to measure the performance. The metrics are defined below.

TABLE VI: Focused and diverse users with different values of topics distributions.

User	Std	Topic Entropy	Topic_0	Topic_1	Topic_2	Topic_3	Topic_4
Focused User 1	0.4446	0.0355	0.0023	0.0007	0.0007	0.9953	0.00077
Focused User 2	0.2985	0.8444	0.1994	0.7167	0.0412	0.0066	0.03590
Diverse User 3	0.1537	**1.3066**	0.1048	0.2951	0.0035	0.2017	0.39466
Diverse User 4	0.0216	**1.6048**	0.2035	0.1996	0.1871	0.1760	0.2336

Accuracy: it is one of the evaluation metrics for classification models, which is the total number of correct predictions divided by the number of users in the testing dataset:

$$Accuracy(\%) = \frac{TP + TN}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} * 100$$

Where TP (True Positives) is the amount of correctly classified spam users, and FN (False Negatives) is the amount of spam users misclassified as legitimate users. FP (False Positives) is the amount of legitimate users incorrectly classified as spam users, and TN (True Negatives) is the number legitimate users correctly classified.

Precision: it is another metric used for evaluating classification model. It is the number of correctly classified spam users divided by the total number of users who are classified as spammers:

$$Precision = \frac{TP}{\mathrm{TP} + \mathrm{FP}}$$

Recall: we measured the recall of the spam users, which is the number of correctly classified spam users, divided by the number of spam users in the testing dataset:

$$Recall = \frac{TP}{\mathrm{TP} + \mathrm{FN}}$$

F-measure: is calculated based on precision and recall as follows:

$$F - measure = 2\ \frac{Precision * \mathrm{Recall}}{Precision + \mathrm{Recall}}$$

Four classification algorithms were used in the experiment, which are Support Vector Machine (SVM), J48, Decision tree and Random Forest. The evaluation results are given in Table 7, from which we can see that Random Forest achieves the best performance. This algorithm has also shown strong results in different spam detection researches [1] [32]. For comparison, we also show the detection performance using the random forest algorithm for each cluster as shown in Table 8.

TABLE VII: Comparisons of different classification algorithms.

Method	Precision	Recall	F1-Score	Accuracy
SVM	0.882	0.877	0.862	87.75%
J48	0.954	0.953	0.954	95.37%
Decision Tree	0.949	0.95	0.949	94.92%
Random Forest	**0.962**	**0.963**	**0.963**	**96.25%**

The results in Table 7 and Table 8 are obtained by using our two-stage approach. To evaluate the performance of the two-stage approach, we conducted another experiment which does not include the clustering stage. In this experiment, we classify users without clustering them in order to determine the detection rate without our proposed method of clustering users based on the topical features. We trained the data using the Random Forest algorithm as one group (without clustering stage) using the existing features and our proposed features, and we got an accuracy of 94.65% as shown in Table 9, which is worse than the accuracy 96.25% produced by using clustering. The results indicate that our proposed method performs well. It shows that spammers' behavior cannot be characterized with unified features, and the technique of grouping users based on

TABLE VIII: Detection result for each cluster using random forest algorithm.

Focused users			**Diverse user**		
Correctly Classified Instances	2469	**96.8235 %**	Correctly Classified Instances	6222	**95.6789 %**
Incorrectly Classified Instances	81	3.1765 %	Incorrectly Classified Instances	281	4.3211 %
Precision	**Recall**	**F-Measure**	**Precision**	**Recall**	**F-Measure**
0.967	**0.968**	**0.968**	**0.957**	**0.957**	**0.957**

focused and diverse interest has confirmed that characterizing users in different groups with different features is crucial to detecting spammers with higher accuracy. This finding of the current study is consistent with that of [14] and [9], who found that the strength of classification relies mainly on the selection of the most appropriate features for spammer detection.

TABLE IX: Detection results with clustering stage and without clustering stage.

	Precision	Recall	F-Measure	Accuracy
With clustering Stage	**0.962**	**0.963**	**0.963**	**96.25%**
Without Clustering Stage	0.947	0.947	0.947	94.65%

A remarkable observation is that, on diverse group, spammers show similar behavior to that of legitimate users. This behavior of having different topics of interest is very common for a large number of legitimate users and spammer try to mimic this behavior by post different messages and topics. Our proposed method of measuring similarity between the user's description and tweets has uncovered this evasive behavior with less effort than existing approaches do [2, 18]. To test the effectiveness of this proposed feature, we train the data without this feature for diverse users and the detection result decreased by 1% as Table 10 shows.

TABLE X: The effectiveness of the Interest consistency feature for classification results.

	Precision	Recall	F-Measure	Accuracy
With our proposed feature	**0.957**	**0.957**	**0.957**	**95.67%**
Without the proposed feature	0.947	0.947	0.947	94.66%

Existing features shown in Table 4 were used to test the classification system. These features have been widely adapted in many previous methods. To further evaluate the effectiveness of the proposed feature, the methods proposed in [13, 24] are chosen as the baselines. We selected these baselines because they used some of the existing features and also used the same dataset as us. The comparison results are provided in Table 11.

VI. Discussion

The strong relationship between spammers' behavior and features has been described in the literature. However, most of these previous works use unified features to represent spammers, without considering that spammers behave differently and that this results in distinct patterns and features that need consideration. The present study aimed to integrate existing features and new features into a framework from the perspective of user level of interests in order to increase the level of detection and provide more reliable features that cannot be easily evaded by spammers.

We believe that our work provides good suggestions for micro-blogging systems to consider focused-interest and diverse-interest users. However, effective features need to be defined to determine focused-interest and diverse-interest users with reliable measurements. We used the LDA topic model to model user information interest using users' tweet content and then we applied cosine similarity between user's description and tweets to cluster users into two separate groups, i.e., focused and diverse users. However, more information about the user's interest can be used in addition to the user's description. Using heuristics, for example, the underlying page posted by the user or considering changes of profile description, would help to establish a greater understanding of users' interests.

The Twitter dataset has restrictions and imposes certain constraints on data collection. The size of the dataset has been slashed. We could not have accessed all Twitter accounts when we crawled the user descriptions. Also, users with insufficient numbers of tweets were excluded as it is difficult to understand user interests from a limited number of tweets. These restrictions may affect the quality of our approach, but our proposed approach performed well in detecting spam users. The experiment conducted in this research project considered the fact that spammers cannot be represented by constant features, and the level of focused interest enabled additional inferences about the effective features. The recommended method and the experiment conducted in this study to detect spammers presented the following strengths:

- This work considered user interests as playing a key role, since the engagement of users in any activity is driven by their interests; further, this feature is difficult for a spammer to manipulate, given that the behavior of spammers lacks a focused interest.
- The framework can handle smart spammers, given that spammers tend to set up fake accounts that appear to be legitimate or to compromise legitimate accounts to hide behind such account. However, the proposed method can identify this tricky behavior implicitly and explicitly through tweet content and profile description.

VII. Conclusion and Future Work

Due to the ability of spammers to use different strategies to evade detection, we conducted an extensive study of user

TABLE XI: The effectiveness of the Interest consistency feature for classification results.

Models	Precision	Recall	F1-Score	Accuracy
Reference [24]	Not provided	Not provided	0.888	88.98%
Reference [13]	0.895	0.951	0.922	Not provided
Our model	**0.962**	**0.963**	**0.963**	**96.25%**

interest evolution patterns. We propose a method to quantify changes in user interest and to depict user topic evolution patterns to understand the degree of focus interest. Based on the level of focus interest among users, we put forward a framework that combines the clustering algorithm with supervised machine learning to detect spammers in online social networks. Our experiment, based on a real-world dataset, reveals the differences between spammers and legitimate users in terms of focus level of interest and shows that user interest evolution patterns are indeed sufficient to represent and detect spammers with different features.

There are many potential directions for future work on this research project. It would be interesting to explore user interest in a dynamic way through different activities to characterize user interest evolution patterns comprehensively. In addition, our detection approach is offline, so it would also be interesting to extend it as an online real-time detection system that is deployed on online social network. Moreover, our proposed method includes a supervised learning stage that is needed to obtain labelled data to train the classification model. However, as it is hard to get enough labelled data because of certain factors, it would be good development to add the ability to perform detection without training data. The main idea of our proposed method is that it does not view social network spammers through constant behavior or represent them with constant features. If we model the user's behavior from a user interest perspective, then the differences between legitimate users and spammers become more evident. This is the most important feature of our work in the design of spammer detection systems in online social networks.

REFERENCES

[1] Benevenuto, F., et al. Detecting spammers on twitter. in *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*. 2010.

[2] Egele, M.S., Gianluca Kruegel, Christopher Vigna, Giovanni, COMPA: Detecting Compromised Accounts on Social Networks. *NDSS*. 2013, San Diego, CA United States: NDSS.

[3] Egele, M., et al., Towards detecting compromised accounts on social networks. *IEEE Transactions on Dependable and Secure Computing*, 2017. 14(4): p. 447-460.

[4] Boshmaf, Y., et al., Íntegro: Leveraging victim prediction for robust fake account detection in large scale OSNs. *Computers & Security*, 2016. 61: p. 142-168.

[5] Martinez-Romo, J. and L. Araujo, Detecting malicious tweets in trending topics using a statistical analysis of language. *Expert Systems with Applications*, 2013. 40(8): p. 2992-3000.

[6] Sedhai, S. and A. Sun. Hspam14: A collection of 14 million tweets for hashtag-oriented spam research. in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2015. Santiago, Chile: ACM.

[7] Hua, W.Z., Yanqing. Threshold and Associative Based Classification for Social Spam Profile Detection on Twitter. in *Semantics, Knowledge and Grids (SKG), 2013 Ninth International Conference on*. 2013. Beijing, China: IEEE.

[8] Fu, Q., et al., Combating the evolving spammers in online social networks. *Computers & Security*, 2018. 72: p. 60-73.

[9] Lee, K., B.D. Eoff, and J. Caverlee. Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter. in *International Conference on Weblogs and Social Media ICWSM*. 2011. AAAI.

[10] Nilizadeh, S., et al. POISED: Spotting Twitter Spam Off the Beaten Paths. in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017. Dallas, Texas, USA: ACM.

[11] Shen, H., et al., Discovering Social Spammers from Multiple Views. *Neurocomputing*, 2016. 255: p. 49-57.

[12] Dang, Q., et al., Detecting cooperative and organized spammer groups in micro-blogging community. *Data Mining and Knowledge Discovery*, 2017. 31: p. 573–605.

[13] Liu, L., et al., Detecting" Smart" Spammers on Social Network: A Topic Model Approach. *arXiv preprint arXiv*:1604.08504, 2016.

[14] Alfifi, M. and J. Caverlee. Badly Evolved? Exploring Long-Surviving Suspicious Users on Twitter. in *International Conference on Social Informatics*. 2017. Cham: Springer

[15] Almaatouq, A., et al., If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts. *International Journal of Information Security*, 2016. 15(5): p. 475-491.

[16] Kaur, R., S. Singh, and H. Kumar, Rise of spam and compromised accounts in online social networks: A state-of-the-art review of different combating approaches. *Journal of Network and Computer Applications*, 2018. 112: p. 53-88.

[17] Sedhai, S. and A. Sun, Semi-Supervised Spam Detection in Twitter Stream. *IEEE Transactions on Computational Social Systems*, 2018. 5(1): p. 169-175.

[18] Ruan, X., et al., Profiling Online Social Behaviors for Compromised Account Detection. *Information Forensics and Security, IEEE Transactions on*, 2016. 11(1): p. 176-187.

[19] Thomas, K.G., Chris Song, Dawn Paxson, Vern. Suspended accounts in retrospect: an analysis of twitter spam. in the *2011 ACM SIGCOMM conference on Internet measurement conference*. 2011. New York, USA: ACM.

[20] Zhu, Y., et al. Discovering Spammers in Social Networks. in *Twenty-Sixth AAAI Conference on Artificial Intelligence AAAI*. 2012.

[21] Song, L., R.Y. Lau, and C. Yin. Discriminative Topic Mining for Social Spam Detection. in *Pacific Asia Conference on Information Systems PACIS*. 2014. AIS Electronic Library.

[22] Blei, D.M., A.Y. Ng, and M.I. Jordan, Latent dirichlet allocation. *Journal of machine Learning research*, 2003. 3(Jan): p. 993-1022.

[23] Gao, H., et al. Towards Online Spam Filtering in Social Networks. in *NDSS*. 2012. NDSS.

[24] Lee, K.C., James Webb, Steve. Uncovering social spammers: social honeypots+ machine learning. in the *33rd international ACM SIGIR conference on Research and development in information retrieval*. 2010. New York, USA: ACM.

[25] Yang, C., R. Harkreader, and G. Gu, Empirical evaluation and new design for fighting evolving Twitter spammers. *Information Forensics and Security, IEEE Transactions on*, 2013. 8(8): p. 1280-1293.

[26] Chu, Z., I. Widjaja, and H. Wang. Detecting social spam campaigns on twitter. in *International Conference on Applied Cryptography and Network Security*. 2012. Springer.

[27] Shen, Y., et al. Automatic fake followers detection in Chinese micro-blogging system. in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. 2014. Cham Switzerland: Springer.

[28] Alfifi, M. and J. Caverlee. *Badly Evolved? Exploring Long-Surviving Suspicious Users on Twitter*. 2017. Cham: Springer International Publishing.

[29] Hall, M.A., Correlation-based feature selection for machine learning, in Computer Science. 1999, Waikato: Hamilton, NewZealand. p. 171.

[30] Witten, I.H., et al., *Data Mining: Practical machine learning tools and techniques*. Fourth Edition ed. 2016, United States: Morgan Kaufmann.

[31] Arthur, D. and S. Vassilvitskii. k-means++: The advantages of careful seeding. in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. 2007. Society for Industrial and Applied Mathematics.

[32] Castillo, C., et al. Know your neighbors: Web spam detection using the web topology. in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. 2007. ACM.