Detection of Spam Tipping Behaviour on Foursquare

Anupama Aggarwal
IIIT - Delhi
New Delhi, India
anupamaa@iiitd.ac.in

Jussara Almeida
UFMG
Brazil
jussara@dcc.ufmg.br

Ponnurangam Kumaraguru
IIIT - Delhi
New Delhi
pk@iiitd.ac.in

ABSTRACT

In Foursquare, one of the currently most popular online location based social networking sites (LBSNs), users may not only check-in at specific venues but also post comments (or *tips*), sharing their opinions and previous experiences at the corresponding physical places. Foursquare tips, which are visible to everyone, provide venue owners with valuable user feedback besides helping other users to make an opinion about the specific venue. However, they have been the target of spamming activity by users who exploit this feature to spread tips with unrelated content.

In this paper, we present what, to our knowledge, is the first effort to identify and analyze different patterns of tip spamming activity in Foursquare, with the goal of developing automatic tools to detect users who post spam tips - *tip spammers*. A manual investigation of a real dataset collected from Foursquare led us to identify four categories of spamming behavior, viz. Advertising/Spam, Self-promotion, Abusive and Malicious. We then applied machine learning techniques, jointly with a selected set of user, social and tip's content features associated with each user, to develop automatic detection tools. Our experimental results indicate that we are able to not only correctly distinguish legitimate users from tip spammers with high accuracy (89.76%) but also correctly identify a large fraction (at least 78.88%) of spammers in each identified category.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and behavioral sciences; H.3.5 [**Information Storage and Retrieval**]: Online Information Services

General Terms

Human factors

Keywords

social networks, location-based social networks, tip spam, user behavior

1. INTRODUCTION

Online Location-based Social Networks (LBSNs) have recently gained a lot of popularity and attracted millions of users in a short span of time. Since most of the mobile phone users have location-sensing enabled in their phones, they can share their location information with their friends easily. [1] LBSNs provide an easy platform for users to *check-in* at the geographical location they are present, spread this information among their network friends and even share their views and opinions about the place they visit in form of comments and photographs.

Foursquare is one of the most popular LBSNs with nearly 30 million users worldwide and over 3 billion check-ins. *Check-ins* may earn users points and help them in getting virtual *badges*. This further incentivises users to actively check-in at venues they visit. Foursquare also provides a platform as a review mechanism about various venues. The users who visit a place can leave a comment (called *tip*) about the place which is publicly available to other users. This feedback mechanism helps others to make an opinion about the specific venue and get some first-hand reviews about the place even before visiting. Since tips on Foursquare are public and easily viewable, they play an important role to share opinion about a venue amongst users and may impact future venue visitors. Tips can be either positive or negative feedback about a venue or even a recommendation. For example, a user may post a tip at a bar to recommend a particular drink or give feedback about the service. Many Foursquare users who want to promote their own brand exploit tips to spread information about themselves. Foursquare usage policy imposes restriction on its users to post tips only which are related to the venue. However, tips are exploited for spamming by unsolicited promotion of brands, spreading unrelated messages and posting malicious content.

There have been few studies to understand user behaviour patterns on Foursquare by studying their check-ins [8], social network graph properties [11] and analyse how users post and respond to tips which show that tips play an important role in determining user behaviour [12]. Some users post tips to provide feedback about a venue, while others use tips to promote a brand. This also indicates that unsolicited advertising and unrelated tips leads to spam.

Unlike a recent effort to detect spam in tip's content [3], the only prior study on that topic, we here aim at detecting users who post spam, considering different patterns of spamming behaviour, observed in real dataset. We refer to such users as *spammers*. We first analyse how user behaviour dif-

[1]Survey: US adults doing mobile check-ins more than doubled. `http://www.tnooz.com/2012/05/11/mobile/survey-us-adults-doing-mobile-check-ins-more-than-doubled/`

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

fers for legitimate and spammers across different features on Foursquare. Then we try to automatically detect users who exploit tips to spread unsolicited messages or post malicious content. In this study, we identify four kinds of irregular user behaviour viz. (i) *Advertising / Marketing* : users who try to promote a specific brand , (ii) *Self-promotion* : users who try to assert their presence at venues without providing any valuable feedback, (iii) *Abusive* : users who post tips only to defame someone in particular and (iv) *Malicious* : users who post malicious content like URLs to malware or phishing websites.

We use various Foursquare features to distinguish between legitimate and irregular users [2] and also classify the users with irregular activities in the above four categories. The major contributions of this study are -

- *Characterizing irregular user behaviour.* We found that Foursquare users with irregular tipping activity can be broadly classified into the following categories - (i) Advertising / Marketing, (ii) Self-Promotion, (iii) Abusive or (iv) Malicious. We present characterisation of features of users in each categories.

- *Automatic detection of spammers.* Using machine learning techniques, we were able to distinguish between legitimate and spam users on Foursquare. We could also automatically detect spammers in each category.

The rest of the paper is organised as follows – Section 2 gives a an introduction about the most common Foursquare terms which we will repeatedly use in this paper. This is followed by few related studies. Section 3 describes how we collected data and the features we used for the detection of spam users. We describe our experiment and results in Section 4. We present the conclusion of our study in Section 5 and discuss future work in Section 6.

2. BACKGROUND

In this section we first define some of the important Foursquare terminology and then describe the relevant previous related studies.

2.1 About Foursquare

Foursquare is one of the most popular and widely used LBSN with nearly 30 million users. The prime objective of Foursquare is to enable its users to share their location with their friends. Following are the few terms related to Foursquare which we will repeatedly use -

Venue.

The geographical location where the user is present and wants to share with his friends. Venues on Foursquare have 8 pre-defined categories with subcategories in each of them. The eight primary categories are "Arts and Entertainment" , "Colleges and Universities", "Food", "Great Outdoors", "Nightlife Spots", "Travel Spots", "Shops", "Home, Work and Others". A venue is created by Foursquare users when it is checked in for the first time.

Check-in.

The action of registering at a venue on Foursquare. A user can check-in when he is physically close to that location using a GPS enabled device. When a user checks in at a venue, this information is spread in his network.

Badges.

Based on the number of check-ins and the venue where the user has checked in, he can earn badges which signify a special task completed by the user. For example, user earns a *Newbie* badge when he checks in for the first time ever on Foursquare.

Mayorship.

If a user checks in at a venue more than any other user in past 60 days, then he earns a special badge for that place and is declared the 'mayor' of that specific venue.

Tips.

A user can leave comments about a place in form of 'tips'. Tips can be negative or positive feedback about the venue and publicly visible by anyone and not only by the user's friends. Unlike check-ins, a user may or may not check-in at a venue to be able to provide a tip at that place. When a user checks in at a venue or locates a venue on Foursuare, he can see all the tips which have been posted by other users for that venue. The user may mark a tip as 'like' and even 'save' the tip for future reference.

2.2 Related Work

There have been several studies to detect spam on various online social media like Facebook [4], Myspace [7], Twitter [1], [5] and Youtube [2]. Most of these studies try to identify various attributes of the social network which can help in distinguishing between legitimate and spam users or content. We here focus on tip spamming in Foursquare, aiming at automatically detecting irregular user behaviour.

Tips on Foursquare are user recommendations or opinion about a specific venue. In this regard, there have been studies to detect opinion and review spamming, specially on e-merchandise websites. [6] study the problem of spam reviews on Amazon. They first tried to find duplicate content posted for various products, and then they tried supervised learning mechanism based on various features of Amazon to automatically detect fake and spam reviews posted for various products. Since tips on Foursquare are also opinions of users about that specific venue, we follow a similar methodology to detect users posting spam tips.

Location based social networks (LBSNs) have recently gained popularity among other social networks. LBSNs provide their users to share the location where they are at and also leave recommendations about that place. Other users can benefit from these recommendations before going to that specific venue. There are several recent studies which explore why people use LBSNs and try to analyse the user behaviour. [10] analyse geo-social properties of location sharing social networks and try to study how geographic distance affects social structure by analysing four OSNs which enable location sharing. There have also been studies to understand the user behaviour on Foursquare to analyse spatio-temporal activity patterns of users by studying their check-in behaviour over time [8]. More recently, particularly tips posted on Foursquare have been studied to understand how users interact with each other and post rec-

[2]We use the terms *spammers* and *irregular users* interchangeably throughout the paper

ommendations [12]. This study also provides the evidence of tip-spamming on Foursquare and indicates irregular user activity which is against the Foursquare terms of services.

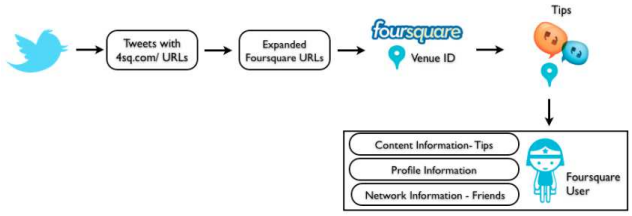
The closest previous effort to our present work is that by Costa et al. [3]. In that work, the authors analysed tip spamming on LBSNs and proposed an automatic tip spam detection method based on supervised learning techniques. Our work differs from [3] because we are here interested in understanding and identifying irregular tipping behaviour as well as developing automatic tools to detect *users* involved in such activities (opposed to detect the spam content, as in [3]). Moreover, unlike [3], where the authors did not distinguish among different types of spamming activity, we here identify four kinds of irregular (i.e., spam related) tipping behaviors, namely - Advertizing, Self-promotion, Abusive and Malicious. In this study, we analyse different features which are characteristic to these categories and compare them with legitimate user activity. We then build an automated mechanism to detect such users.

3. METHODOLOGY

3.1 Data Collection

For our experiments, we had to collect tips posted at various venues. Venues on Foursquare have a unique ID which is a random 16 - 24 digit alphanumeric string. Since it is computationally impossible to spawn over all such IDs, we followed the following data collection methodology. We first used Twitter streaming API to look for tweets which had '4sq.com' as a pattern. Such tweets are those which indicate the check-in activity of a Foursquare user which is also shared on Twitter. We used this url to find the venue ID and then extracted all the tips posted at that particular venue. In the end we had 2,400,594 tips posted in total by 613,298 Foursquare users. We use this data for all the experiments in our study.

For all of the 613,298 Foursquare users, we collected their additional Foursquare profile information for the analysis which included various attributes like the number of photos they have posted, their number of friends, tips, badges, mayorships and other profile data publicly available.



3.2 User Categories/Behaviour on Foursquare

Based on the content of tips Foursquare users post, we identified irregular tipping behaviour. According to Foursquare usage policy, *"Spam is any content including links to websites selling software, realtor contact info, a listing for your business, or other promotion)"*. Foursquare also says that if same or similar content is posted across various distinct venues then that is considered a violation of its terms of services. Also, the tips posted should be related to the venue and should be helpful recommendations.

In accordance with these terms of services, we found irregular tip activity pattern among Foursquare users. However, the users who violate the terms of services by posting unrelated tips can be further divided into four categories. Manual inspection of our dataset and the content of the tips led us to the following categories of irregular behaviour:

Advertising / Marketing.

Users who try to promote a specific brand or other venue. Such users often post about their brand/venue in form of tips at multiple venues. The content of the tip is often same or similar and talks only about a specific brand which the user wants to promote.

Self-promotion.

We found that some users post repeated and unrelated tips not with an intent of advertising but asserting their presence at a particular location. For example we found a user who posted a tip - "I am the mayor of *Place-XYZ*!!" at multiple venues. We found a large fraction of such users in our dataset.

Abusive.

Since Foursquare lets the users earn badges and mayorships, we observed that many users check-in to gain these points. We found a large number of users who posted particularly abusive words about other Foursquare users in form of tips at various venues. Some users also wrote derogatory words about a specific individual and not the service provided at a venue. We marked such users as 'Abusive'.

Malicious.

We found a small fraction of Foursquare users posting tips with URLs to malicious or phishing websites.

In all our further experiments, we will consider 'legitimate' users and 'spammers' divided into the aforementioned four categories.

3.3 Annotation

After we manually identified the various user tipping behaviour on Foursquare, we manually annotated a sample of our dataset to classify each user in legitimate, Advertising, Self-promotion. Abusive or Malicious categories. We found that a single Foursquare user may exhibit multiple tipping behaviour. For example, a Foursquare user may post legitimate tips, but may also sometimes post malicious content at few venues. However, tip distribution across users is very uneven. A few users post a large number of tips [12]. Thus manually analyzing all tips posted by a user is very costly. Also, in case of spamming behaviour, the tips are heavily repeated. Therefore we showed the annotator a sample of the tips of the Foursquare user. Incase the tips were repeated, we showed the text only once and marked how many times the tip was repeated. We then asked the annotators to first mark a sample user in either one or two of the given 5 categories, and then for each selected category, give a score of 1 - 3 viz 1 for strongly agree, 2 for agree and 3 for maybe. Then we calculated augmented kappa score [9] for each annotated sample point to find the inter-annotator agreement for the primary category of the Foursquare user. We considered the weights of the score of categories as 0.67 for 'strongly agree', 0.22 for 'agree', 0.11 for 'maybe'.

The augmented kappa score [9] was calculated using the formula -

$$K' = \frac{p(A) - p(E)}{1 - p(E)}$$

where, p(A) is the observed probability and a(E) is the expected probability. To compute p(A) we first compute the annotation matrices for each annotator - $M_{annotator}$. For each annotator, M_i has N rows, i.e. the number of users he has annotated, and M columns, i.e., the total number of categories. $M_i[x, y]$ is the score corresponding to the category y annotator i has marked user x. For example, if annotator i marked the user x as Malicious with 'strongly agree' and 'Abusive' as 'maybe', then $M_i[x, \textit{Malicious}] = 0.67$ and $M_i[x, \textit{Abusive}] = 0.11$ and $M_i[x, \textit{Advertising}] = M_i[x, \textit{Self-Promotion}] = M_i[x, \textit{Legitimate}] = 0$. Similarly, we computed the annotator matrix for both the users and then computed the *Agreement matrix Ag*. Given two annotator matrices M_A and M_B, Agreement matrix Ag is computed such that

$$Ag[x, y] = M_A[x, y] * M_B[x, y]$$

Finally we compute p(A) = $\frac{\alpha}{N}$, where α is the sum of all cells of Ag.

To compute p(E), we use the relative frequencies of each annotator's labelling preference. We compute the relative frequency vectors for each category for an annotator defined by

$$Freq_i[y] = \Sigma_{x=1}^{N} \frac{M_i[x, y]}{N}$$

Then, using the frequency matrices for both the annotators A and B, we compute p(E) defined by

$$p(E) = \Sigma_{y=1}^{M} Freq_A[y] * Freq_B[y]$$

Using the above definitions of p(E) and p(A), we compute the Augmented Kappa Score and take into consideration the annotated users for which the value came out to be more than 0.7 and take the corresponding label as the primary category of the Foursquare user. This indicates a high inter-annotator agreement and therefore we considered only these annotated users for our training set.

As a result of manual annotation and selecting annotated uses with high inter annotator agreement, we had 2000 legitimate users and 1900 spammers.

3.4 Features used for Classification

Foursquare users who post tips with the intent of advertising, marketing a product, defaming someone or spreading malicious content have a characteristically different behaviour from regular Foursquare users. Such behaviour can be captured by studying the various attributes of the social profile of such users. In this section we analyse different Foursquare attributes which can be a used to distinguish between regular and irregular Foursquare users. To capture the entire profile information of a user, we look at the following - user attributes, social attributes and content attributes.

User attributes are the properties of the Foursquare profile of the user. Though the check-in history of a user is not publicly available, we capture the number of checkins of the user. We also look at the number of badges and mayorships earned by the user and the total number of tips he has posted. Usually users post tips when they check-in at a particular venue. However, Foursquare enables users to post a tip even if they have not visited that specific venue. This is often exploited by spammers to post unrelated tips at several venues. Therefore, there is a strong correlation between the number of check-ins and number of tips posted by the user. This is used as one of the most significant attributes for detection. Other attributes like presence of profile picture and the badges earned are also important features. Legitimate users usually put up their display photographs and also try to earn more points to gain badges. This is less commonly observed with spammers.

Next, we analyse social attributes of the user which describe how he is connected to other users on Foursquare. Foursquare has a uni-directional friendship network. To capture this, we looked at the the number of friends the user has on Foursquare. Unlike other social networks, where the amount of content penetration depends on the number of friends, Foursquare users do not necessarily need more friends to do so. This is because Foursquare users communicate with other users by posting tips, which are publicly viewable. We found that spammers have significantly lesser number of friends than legitimate users.

The content attributes play a major role in detection of spammers. The content generated on Foursquare is in form of tips which are posted by users at various venues. We look at several features of tips to analyse whether the user is involved in spamming activity or not. We also find whether the tips posted by the user have any spam words (e.g. 'lottery', 'free') present or not. The users who spread malicious content and aim for viral advertising, use a large number of spam words. Also, specifically the users who aim at advertising, often spread the same or similar tips across several venues. We capture this by calculating Jaccard coefficient. For example, if T_1 and T_2 are tips posted by the user, then J(T_1 , T_2) is defined as the total number of words common in T_1 and T_2 divided by the total number of words in T_1 and T_2. We use this to find how much similar content the user is posting via tips at various venues. If the similarity coefficient is high, then there is a high possibility that the user is spamming. We also analyse the content of the tip and find if the user has posted any URLs. We use Google Safebrowsing API to find whether the posted URL is malicious or not. Spammers with the intent to spread malicious content often use URLs in the tips which redirects to external content like malware or phishing webpages. Therefore, we also calculate the average number of URLs the user posts in his tips.

Next, to find the most informative features amongst these to discriminate the legitimate users from spammers, we apply χ^2 (Chi-square) feature selection method. Table 1 gives a list of all the features we assessed and their χ^2 rankings. Based on this ranking, we take the best 15 features for our experimentation and to automatically distinguish Foursquare spammers from legitimate users. We observe that these most discriminative features are distributed across all the feature sets. Content attributes play a major role in determining a user as legitimate or spammer since the properties of the tips posted by the user indicates the quality of content posted by him.

The above three set of attributes are used in our further experiments to automatically detect spammers. The next section will illustrate how these features vary for legitimate users and spammers. We will also analyse how these features differ across various categories of spammers which we defined earlier. For example, users who promote advertising often

Table 1: Features used to detect Foursquare spammers and their χ^2 rankings

User Attributes	**4**	Number of check-ins
	5	Number of badges
	11	Number of mayorships
	3	Ratio of check-ins and tips
	12	Ratio of check-ins and badges
	16	Presence of profile picture
	1	Number of tips
	15	Number of photos posted
Social Attributes	**6**	Number of friends
	17	Number of lists created
	19	Number of lists saved
Content Attribute	**9**	Average number of characters in tips
	8	Average number of words in tips
	7	Number of URLs posted
	2	Similarity score of tips
	13	Average number of spam words in tips
	20	Total number of likes for the tips
	10	Ratio of number of likes and number of tips
	18	Average number of numeric characters
	14	Average number of phone-numbers posted in tips

have a URL in their tips to redirect to external content, however, those who post abusive content do not have URLs. We will analyse this behaviour in the next section.

4. EXPERIMENTAL SETUP AND RESULTS

In this section, we provide our detailed experimental analysis. We first analyse how various deterministic features differ for legitimate users and spammers. We draw distribution graphs of few of the discriminative features and observe the difference in user behaviour for both the classes. Next, based on these discriminative features, we build automated mechanisms to differentiate between spammers and legitimate users. For this purpose, we compare the results of three machine learning techniques and evaluate their performance to distinguish between spammers and legitimate users. Then, to determine the category of each spammer, we apply a hierarchical clustering algorithm to divide spammers into further four categories, viz. Advertising, Self-promotion, Abusive and Malicious.

4.1 Comparison of features across spam and legitimate category

In this section, we compare various Foursquare features which are characteristically different for legitimate and spam users. We look at various types of attributes described in the section above and compare user behaviours across legitimate and spam class. We find that user behaviour for spam users is significantly different from legitimate users.

Figure 1 shows that the frequency distribution of badges for legitimate users is a power law. Few legitimate users earn large number of badges. However, the spammers have lesser badges on an average as compared to the legitimate users.

The number of checkins are shown in Figure 2. The figure shows that about 20% of the spammers have zero check-ins, which is a very different behaviour from that of legitimate users. The spammers are not interested to check-in at places, but rather more interested to post tips. Tips can be posted at venues without requiring the user to check-in at that venue. Spammers exploit this feature and post unrelated tips at multiple venues.

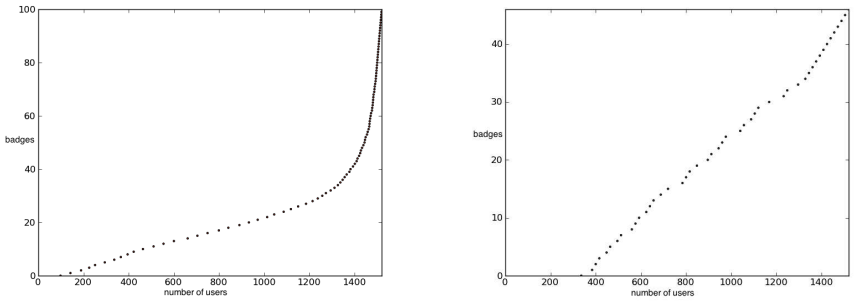
Figure 3 shows the distribution of number of friends for spammers and legitimate users. In case of legitimate users, the distribution is power-law. Similar to the check-in behaviour, spammers on Foursquare have very less friends. About 10% of the spammers do not have any friends. Foursquare is a LBSN, hence, legitimate users intend to have friends to share their location. However, the tip spammers have very less friends as tips posted are publicly available and they need not have friends to gain visibility.

Distribution of the number of tips shown in Figure 4 is the most discriminative feature. The spammers post a large number of tips, many of them are often repetitive. Legitimate users however exhibit a power-law in their tip distribution graph. The average number of tips posted by spammers is much higher than those posted by legitimate users.

4.2 Classification of users in legitimate and spam category

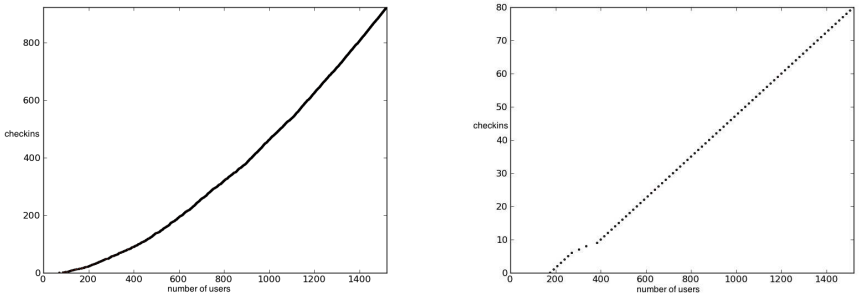
For our experiments, we compared various supervised classification algorithms to evaluate their performance on detection of legitimate and spam users. As a baseline we used KNN binary classification. The algorithm compares k nearest neighbours of all the data points in the dataset and then the class is assigned which is voted most among these k neighbours. For example, if k is 1, the datapoint will have the same label as the immediate neighbour. If k is 3, then the algorithm will chose the label which occurs most often in the surrounding 3 neighbouring data points. With KNN binary classification, we achieved an average accuracy of 84.89%. However, we did not obtain high precision and recall values for spammers and legitimate users. The precision for spammer class is 83.2%, which signifies that a large fraction of spammers were not correctly classified.

We then used decision trees and random forest. Decision tree algorithm is based on a predictive model which creates a



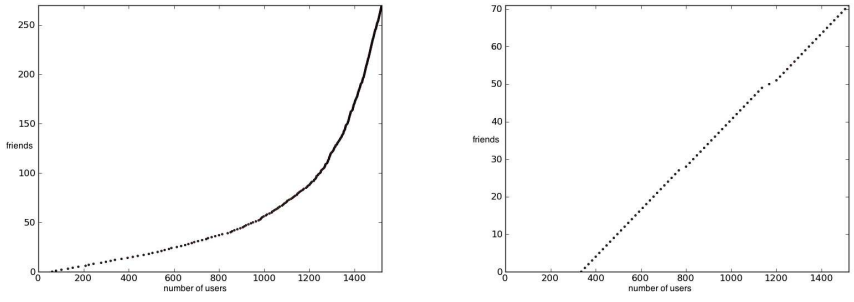
(a) Legitimate Users (b) Foursquare Spammers

Figure 1: Number of badges on Foursquare



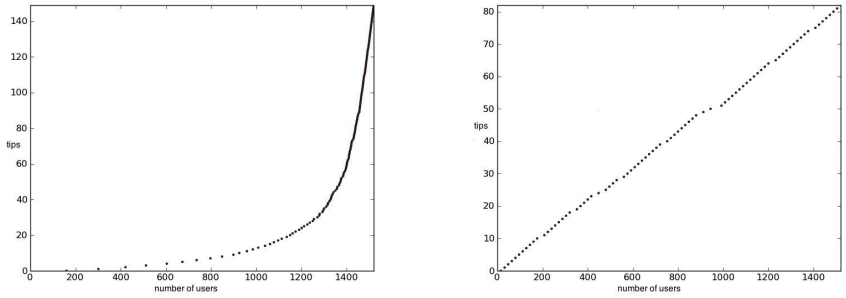
(a) Legitimate Users (b) Foursquare Spammers

Figure 2: Number of checkins on Foursquare



(a) Legitimate Users (b) Foursquare Spammers

Figure 3: Number of friends on Foursquare



(a) Legitimate Users (b) Foursquare Spammers

Figure 4: Number of tips on Foursquare

classification tree. It creates a model that predicts the category of the target data point by learning simple decision rules inferred from the data features. We use *'DecisionTreeClassifier'* module provided by *'scikit'* library. We received an accuracy of 89.53%. Using decision tree algorithm, we improved the precision and recall metrics and obtained a high precision of 89.2% for the 'safe' class.

We achieved the highest accuracy with random forest of about 89.76% for dividing the users into legitimate or spam category. For each data point to be classified, Random Forest algorithm randomly chooses a subset of features which are used for classification. It selects the most important features of the data point hence improves the predictive accuracy and controls over-fitting. We use *'RandomForestClassifier'* module provided by *'scikit'* library. To accurately choose the value of the parameters, we applied the standard grid search parameter optimization algorithm to determine the optimum number of features and the number of trees. As a result, we obtained the value of 'number of features' as 15 and 'number of trees' as 175. We hence use these values for our experiments. We slightly improved the accuracy as well as the precision-recall metrics when we used Random Forest algorithm. We received a high precision (90.2%) and recall (90.3%) for the 'safe' class. The precision-recall metrics remained low for the 'spam' class. One of the reasons for this could be that a lot of spammers on Foursquare exhibit mixed behaviour by sometimes posting legitimate content and sometimes posting spam tips. Such erratic behaviour is hard to determine for our automated system.

All of the classification results to categorise a user into spam or legitimate category are described in Table 2. For all the experiments, we used 5-fold cross validation. In this process the entire dataset was divided into 5 distinct sets, 4 were used for training and 1 was used for testing. This process was repeated 5 times such that each set is used in training as well as testing. This 5-cross fold validation was repeated 10 times and hence we obtained 50 different results and then computed the average of all the 50 runs. The results do not differ more than 1% from the average with a 95% confidence.

4.3 User categorisation for spam class

In this section we evaluate how we can classify each user into different spam categories which we described, i.e., Advertising, Self-promotion, Abusive and Malicious. We used hierarchical clustering on the dataset to divide them into different categories based on the 15 most informative features.

We used the Expectation-Maximization (EM) clustering algorithm. We used the EM implementation in Weka, which determines the number of clusters automatically for the given dataset using the mentioned features. It is based on 10-fold cross validation where the data is divided into 10 parts; 9 are used as training set and 1 is used as testing set. This process is repeated 10 times, such that each part has been used for both training and testing. For each run, it builds clusters on the training set and computes the log-likelihood of each instance in the testing-set. Then, the log-likelihood values for each instance is summed and averaged over all the 10-folds. The final number of clusters is determined by finding the maximum of these average log-likelihood values. We performed EM clustering over the entire dataset of annotated Foursquare users. We found that in the first split, the users were divided into two distinct clusters, viz spam and legitimate with an average accuracy of 82.56% and 83.82%. The second split resulted into 3 clusters of users for Advertising, Self Promotion and Abusive. Since our dataset had very small sample of malicious users and the Foursquare features of malicious users are not very distinct from users of other spam categories, we could not correctly detect any of these malicious users automatically. We were able to detect users belonging to Advertising, Self-promotion and Abusive categories with an accuracy of 88.23%, 87.23% and 78.88%.

The above experiments show that using Foursquare features, we were able to detect spam and legitimate users. We further show that these features can be used to distinguish different behaviour and intent of spam users. We were able to automatically differentiate between the user behaviour of spam users by analysing the difference in various features.

5. CONCLUSION

In this study, we approached the problem of analysing user behaviour on Foursquare based on their tipping activity and detected Foursquare spammers. We collected data from Foursquare about tips posted at various venues and the users who post these tips and analysed how users exploit tips to spread unwanted spam. We observed that the Foursquare spammers can be further divided into four categories viz. Advertising, Self-promotion, Abusive and Malicious. We first analysed the discriminative features to distinguish spammers from legitimate users. We found that spammers behave much differently from legitimate users.

Table 2: Different classification algorithms used for Foursquare spammers detection

Classification Algorithm	Precision (Spam)	Precision (Safe)	Recall (Spam)	Recall (Safe)	Accuracy
KNN	83.2%	86.6%	86.3%	83.5%	84.89%
Decision Tree	88.1%	89.2%	88.3%	85.8%	89.53%
Random Forest	89.3%	90.2%	88.33%	90.3%	89.76%

Few of the most discriminative features are the number of tips posted by the users, number of checkins made and the number of badges earned. We found 15 significant features which we then used to automatically detect spammers using machine learning techniques. We obtained an accuracy of 89.76% with Random Forest classifier to distinguish spammers from legitimate users.

Next, we classified the spammers into four broad categories based on manual observation of our dataset. We used hierarchical clustering algorithm to automatically divide the spammers into these four categories using the same discriminative features. We were able to to detect users belonging to Advertising, Self-promotion and Abusive categories with an accuracy of 88.23%, 87.23% and 78.88%. The malicious users in our dataset were very small. Hence, we could not identify such users automatically. However, analysis of URLs posted by the users along with the tips and a lookup in the spam and malware blacklists like Google Safebrowsing and PhishTank indicates a small presence of such users who post malicious URLs in their tips.

Therefore, our experiments show that we could automatically distinguish spammers from legitimate users with a high accuracy of 89.76%. Our method was also able to categorize spammers into different categories based on their tipping behaviour with a minimum accuracy of 78.88%.

6. FUTURE WORK

We envision the following directions in which this work can be further extended. We can further refine our methodology by use of other classification algorithms. In our dataset, we also observed that a high fraction of Foursquare users had connected their accounts with Twitter and Facebook. Many of the Foursquare spammers also posted the same content as in their tips to Twitter. This behaviour shows that the spammers want to spread the spam using multiple social networks. In future, we intend to study this behaviour and analyse how spammers can leverage multiple social media to spread the same spam campaign. As in case of spammers belonging to Advertising category, for some users, we observed different Foursquare users posting links to the same product. Correlation of content and the URLs posted by different users can help us in identifying several spam campaigns.

7. REFERENCES

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting spammers on twitter. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, volume 6, 2010.

[2] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 620–627. ACM, 2009.

[3] H. Costa, F. Benevenuto, and L. H. de Campos Merschmann. Detecting tip spam in location-based social networks. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing (SAC)*, 2013.

[4] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th annual conference on Internet measurement*, pages 35–47. ACM, 2010.

[5] S. Ghosh, B. Viswanath, F. Kooti, N. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. Gummadi. Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st international conference on World Wide Web*, pages 61–70. ACM, 2012.

[6] N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, pages 219–230, 2008.

[7] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots+ machine learning. In *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 435–442. ACM, 2010.

[8] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. *ICWSM*, 2011.

[9] A. Rosenberg and E. Binkowski. Augmenting the kappa statistic to determine interannotator reliability for multiply labeled data points. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 77–80. Association for Computational Linguistics, 2004.

[10] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora. Distance matters: Geo-social metrics for online social networks. In *Proceedings of the 3rd conference on Online social networks*, pages 8–8. USENIX Association, 2010.

[11] S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. *Proceedings of ICWSM*, 11, 2011.

[12] M. Vasconcelos, S. Ricci, J. Almeida, F. Benevenuto, and V. Almeida. Tips, dones and todos: uncovering user profiles in foursquare. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 653–662. ACM, 2012.