Effective Features Selection for Detecting Fake Accounts on Twitter

Myo Myo Swe
University of Computer Studies, Mandalay (UCSM)
myomyoswe.maeu@gmail.com

Nyein Nyein Myo
University of Computer Studies, Mandalay (UCSM)
vicky.mdy@gmail.com

Abstract

Social networking sites have turned out to be extremely well known as of late. Most internet users utilize them to discover new companions, refreshes their current companions with their most recent feelings and thoughts. Among these sites, Twitter, the quickest developing networking site, additionally pulls in numerous fake users to penetrate genuine users' accounts with a lot of fake news, malware, viruses, and so on. This paper identifies effective features to distinguish fake accounts from legitimate accounts. Firstly, 20 features from users' tweet content and user's profile, which check possibly, are extracted to distinguish fake accounts. Effective features are selected from these 20 features using two feature selection methods. Validations of the effective features are proofed on Decision Tree method. The experimental results of five machine learning classifiers are shown in this paper and Random Forest classifier achieves the best detection accuracy with 95.7%.

1. Introduction

Twitter is an online social networking site and online users can use it to post tweets and interact with others utilizing these. These tweets were limited to 280 characters. Therefore, most of the authors check the content of the tweets whether contains the URLs that link to malicious websites to reach their goal. Only two features; URL rate and interaction rate are used [9].

Malicious users create fake accounts on Twitter to accomplish their pernicious objectives, such as sending spam, spreading fake links, and infiltrating fake news to defame a person or company. All these malignant practices may make noteworthy financial misfortune our general public and even debilitate national security. Twitter gives a few techniques to users to report spam. For examples, Twitter provides a "report as spam" link for a user to report spam account to Twitter. User can write a tweet "@spam @username" format to mention the spam account. All these specially appointed strategies need users require to distinguish spam physically and rely upon their own particular experience [1].

There were some past investigations on the Twitter fake accounts recognition for quite a long time. Their methodologies address this issue as the distinguishing a Twitter account into a fake account or a real account. By studying the fake characteristics, the successful features that are extracted from content-based, network-based, and timing-based of users have been proposed. However, fake users also utilize new strategies to evade detection and existing methods and features are not enough to effectively detect fake account.

The difference characteristics between fake users and genuine users are initiated to examine in this paper. These characteristics are used as features to detect fake users from legal ones. The major contribution of the proposed system is to develop a fake account detection system in an effective manner which can reduce time and memory overhead using effective features. The aim is to select the effective features that can efficiently detect the fake accounts on Twitter. The four steps are fulfilled to achieve this aim.

- Firstly, twenty features are extracted based on user's profile, user's content and user's posting time behavior. User-based features are extracted from profile of the user, content-based features are got from text of tweets and timing-based features are extracted from tweeting time behavior of the user.

- After extracting 20 features, two features selection methods: information gain and gain ratio are applied on these features to choose effective features. Features with highest weights in two features selection methods are selected as effective features.

- We prove that our chosen features are effective based on Decision Tree C4.5 algorithm. The proposed system using these effective features not only can achieve best accuracy but also reduce time and memory overhead.

- The detection approach is evaluated using five machine learning classifiers such as Random Forest, Decision Tree, AdaBoost, Bagging and LogitBoost.

The remaining part of the paper is sorted in the following order: related work describes the previous fake account detection methods in section 2. We discuss the features related to profile, content and tweets posting time of the users to detect fake accounts in section 3. Section 4 explains the important features that can effectively detect the fake accounts. In section 5, the experimental results using C4.5 Decision Tree are shown. Section 6 evaluates the detection approach using five machine learning classifiers. We conclude and describe the future work in section 7.

2. Related Work

Because of the popularity of online social networking sites, fake accounts are transforming into the quickly developing on Twitter. There were some past examinations to handle this issue. A few examinations concentrate on investigating fake accounts characteristics on Twitter.

Alex Hai Wang [1] proposed spam profiles detection based on user based and content based features. The author proposed a directed social graph model using the “follower” and “friend” relationships. The authors compared the results of four machine learning classifier such as Decision Tree, Support Vector Machine, Naïve Bayes and Neural Network; Naïve Bayes achieved the best detection rate amongst all these classification algorithms.

Based on tweet content and user behavior, Benevenuto et. al. [7] detected spammers using 39 content-based features and 23 user-based features. Dataset containing 355 spammers and 710 non-spammers has been created based on 54 million users on Twitter. Their proposed system achieved 87.4% accuracy with Support Vector Machine classifier.

McCord et.al. [8] utilized traditional machine learning classifiers like Support Vector Machine, Naïve Bayesian, K-Nearest Neighbor and Random Forest to detect spam profiles on Twitter based on user-based and content-based features. In case of unbalanced dataset, spammer detection accuracy is 95.7%.

Amit A. et. al. [2] described two kinds of spammer detection approaches: user centric approach based on user attributes and URL centric approach based on URLs in tweets posted by users. Hybrid features including 15 new features and existing features have been used to identify spammers on Twitter. Decorate classifier achieved the best accuracy amongst three classifiers. This approach has been tested on 31,808 users, but, Twitter is designing millions of users.

To detect spammers on Twitter, Yang et. al. [4] proposed 10 new features consisting of three neighbor-based features, three graph-based features, three automation-based features and one timing-based feature. Random Forest, Decision Tree, Decorate and Bayesian classification methods have been used, and Bayesian classification algorithm achieved the best accuracy with 88.6%. However, new features of this approach are robust for evasion, these features were not only expensive to extract but also difficult to compute.

Meda, Claudia, et al [6] applied three machine learning approaches, namely; Support Vector Machine (SVM), Extreme Learning Machine (ELM) and Random Forest. This approach detected spammers on Twitter based on 13 features. Five machine learning classification algorithms such as Random Forest, Decision Tree, AdaBoost, Bagging and LogitBoost were applied to compare the detection rates. The accuracy of Random Forest classifier was the best in this approach.

Chakraborty et.al. [3] detected abusive users by checking harmful URLs, porn URLs and phishing links. This approach applied two-steps approach. Twenty features were used to detect abusive users and four classifiers were applied. Support Vector Machine classifier achieved the best accuracy with 89%. However, the accuracy of that system was not good. Therefore, in our system, effective features that give the best accuracy are proposed.

The works in literature utilized many features to distinguish fake accounts on Twitter. However, most of the features don't notably affect identifying fake accounts and they did not get the best accuracy. These approaches can cause extra crawling cost and

take a long time for features extraction. This is an issue for these approaches. Therefore, we develop a fake account detection system that can reduce overhead cost and also memory and time overhead by reducing the feature dimensions.

3. Fake Features Identification

The inspiration driving features used to recognize fake accounts on Twitter are described in this section. Fake users and normal users have difference aims of using Twitter to achieve their goals. Therefore, fake user and normal user can be distinguished based on user profiles, user content and their tweeting time behavior. Six user-based features, eleven content-based features and three timing-based features are identified to detect fake accounts in this section.

3.1. User-based features

User-based features are features that are extracted from the user profiles and these features are useful for detecting fake accounts. Six user-based features are described in the following.

3.1.1. Number of Followers

If an account has small number of followers, this account will be suspected to become a fake account.

3.1.2. Number of Followings

An account have large number of followings, it can be recognized as a fake account. Most of the fake accounts have many followings.

3.1.3. Reputation

Reputation is used to measure the relative importance of an account. If the reputation of an account closes to zero, the fake probability of that account is high.

$$\text{Reputation} = \frac{followers}{followers + followings} \quad (1)$$

where,
Reputation = the reputation of the user
followers = the number of followers
followings = the number of followings

3.1.4. Account Age

The less the account is aged, the more it could be considered a fake one.

3.1.5. Follower Rate

This feature reflects the reputation of the users. The higher the follower rate is, the more the account is real.

$$FollowerRate = \frac{NumberOfFollowers}{AccountAge} \quad (2)$$

where,
FollowerRate = the follower rate of the user
NumberOfFollowers = the number of followers
AccountAge = the age of the account

3.1.6. Following Rate

Fake users have aggressive following behavior. If the rate of following is high, the account is more likely to be faked.

$$FollowingRate = \frac{NumberOfFollowings}{AccountAge} \quad (3)$$

where,
FollowingRate = the following rate of the user
NumberOfFollowings = the number of followings
AccountAge = the age of the account

3.2. Content-based features

Content-based features are got from the text of the tweets posted by the users. Content-based features are calculated based on the most recently twenty tweets. Eleven content-based features are described in below.

3.2.1. Number of Tweets

Fake users post more tweets to be more active and more willing to interact with others. They posted tweets in a specific time interval using particular automated tweeting tools and software such as Twitter API and AutoTwitter. Therefore, number of tweets is an attribute for detection.

3.2.2. Number of URLs

Twitter allows users to post tweets with 280 characters so that most of the users use shortened URLs to post their tweets. Twitter cannot check the shortened URLs and fake users also utilize these shortened URLs that are leading to the malicious pages by clicking these links. Therefore, number of URLs containing in tweets is the significant feature of fake accounts detection.

3.2.3. URLs Ratio

If the rate is high, the probability of the fake is also high. To lure the legitimate users to malicious pages, fake users posts malicious URLs in their tweets.

$$URLRatio = \frac{NumberOfTweetsContainingURLs}{TotalNumberOfTweets} \quad (4)$$

where,
URLRatio = the ratio of the URLs in the tweets
NumberOfTweetsContainingURLs = the number of tweets containing URLs
TotalNumberOfTweets = the total number of tweets that the user posted

3.2.4. Number of Mentions (@username)

Most of the spammers post numerous @ usernames as spontaneous mentions in their tweets. In this event, users incorporate an excessive number of answers/says in their tweets, Twitter will think about that user as suspicious.

3.2.5. Mentions Ratio

If the mentions ratio is high, the fake probability is high.

$$MentionRatio = \frac{NumberOfTweetsContainingMention}{TotalNumberOfTweets} \quad (5)$$

where,
MentionRatio = the ratio of the mentions
NumberOfTweetsContainingMention = the number of tweets containing mention

3.2.6. Number of Hashtags (#)

Fake users use popular hashtags (#) to attract the legitimate users. Therefore, number of hashtags is a feature to detect fake account on Twitter.

3.2.7. Hashtags Ratio

Fake users utilize hashtags to grab the attention of the legitimate users so that hashtags ratio can be used for detecting fake accounts. The high the hashtags ratio is, the more the account suspious.

$$HashtagRatio = \frac{NumberOfTweetsContainingHashtags}{TotalNumberOfTweets} \quad (6)$$

where,
HashtagRatio = the ratio of the hashtag
NumberOfTweetsContainingHashtags = the number of tweets containing the hashtags

3.2.8. Number of Retweets (@RT)

Users can use retweets with the symbol @RT to share other users' tweets. This is another content-based feature for detecting fake accounts.

3.2.9. Number of Spam Words

Some tweets of fake users convey unequivocal spam information. Spam words dataset are created to catch spam content based on our perception and existing list of spam trigger words [5].

3.2.10. Spam Words Ratio

On the off chance that the spam words proportion is high, the likelihood of fake is additionally high.

$$SpamWordsRatio = \frac{NumberOfSpamWords}{TotalNumberOfWords} \quad (7)$$

where,
SpamWordsRatio = the ratio of spam words
NumberOfSpamWords = the number of spam words in the users' tweets
TotalNumberOfWords = the total number of words containing in tweets posted by the user

3.2.11. Total Number of Words

This feature also indicates the fake trademark. The tweets of fake users contain more words to describe the product details of advertisement or more porn URLs links.

3.3. Timing-based features

Some fake users are bots and they have automation behavior for tweeting post. However, normal users have abnormal behavior. Therefore, timing-based features are used as strong separators

between automation behavior and normal behavior. Three timing-based features are explained in below.

3.3.1. Mean Time between Tweets (μ)

Fake users are dominatingly observed to make posts at a quicker rate when contrasted with legitimate users. This is an essential perception and we trust this component would enable us to catch fake users.

$$\mu = \frac{\sum(TimestampOfTweet(i) - TimeStampOfTweet(j))}{TotalNumberOfTweet - 1} \quad (8)$$

where,
μ = mean time between tweets
TimestampOfTweet(i) = the timestamp of the i^{th} tweet posted by the user
TimestampOfTweet(j) = the timestamp of the j^{th} tweet posted by the user

3.3.2. Extreme Idle Duration Time between Tweets (Idle)

Fake users are seen to be discrete in their posting conduct. They post tweets in bursts. This feature would enable us to catch the level of change of this conduct amongst fake and real users.

$$Idle = \frac{Max(TimeStampOfTweet(i) - TimeStampOfTweet(j))}{TotalNumberOfTweet - 1} \quad (9)$$

where,
Idle = Extreme idle duration time between tweets

3.3.3. Standard Deviation between Tweets (σ)

This feature can distinguish fake users and legitimate users.

$$\sigma = \sqrt{\frac{\sum(X - \mu)}{TotalNumberOfTweets - 1}} \quad (10)$$

where,
σ = standard deviation between tweets
X = timestamp of the tweet
μ = mean time between tweets

4. Effective Features Selection

In section 3, 20 features (6 user-based features, 11 content-based features and 3 timing based features) are identified. However, most of the features are not essential for detecting and some features are not easy to extract and some can cause computation heavy. Therefore, in order to choose effective features, two features selection methods, namely, information gain and gain ratio; have been used. In information gain feature selection method, reputation, number of tweets, number of spam words, follower rate and mean time between tweets features get the high weight. Based on the weights of gain ratio feature selection method, reputation is the highest rate, followed by mean time between tweets, number of spam words, follower rate and number of tweets. These features are not only common but also high rank in both feature selection methods. Therefore, reputation, mean time between tweets, number of spam words, follower rate and number of tweets are selected the best features for detection of fake account. Table 1 shows top 5 effective features and their respective weights based on information gain and gain ratio.

Table 1. Top 5 features and their respective weights

Rank	Information Gain		Gain Ratio	
	Feature	Weight	Feature	Weight
1	Reputation	0.5597	Reputation	0.302
2	Number of tweets	0.5268	Mean time between tweets	0.2728
3	Number of spam words	0.4733	Number of spam words	0.2245
4	Follower rate	0.4696	Follower rate	0.2209
5	Mean time between posts	0.3857	Number of tweets	0.217

5. Experiments

To test the system, the dataset from (Yang et al, 2012) which are collected from April 2010 to July 2010 is utilized [10]. In this dataset, the information contains 11000 users and 1354616 tweets of them. The dataset have 1000 fake users and 10000 honest users. Therefore, the ratio of fake to honest users is 1:10. However, the users who are not tweeting in English language are excluded; the proposed method is evaluated on 1000 users which are 500 fake users and 500 normal users.

Experiments are performed on a 2.10 GHz Intel Core i3 processor and 2GB RAM running Window 7. For features selection, information gain and gain ratio operations were done in WEKA toolkit. After features reduction, J48 in WEKA toolkit is used as a C4.5 decision tree for detection. The experimental results for fake accounts detection before applying features selection and after applying features selection are shown in Table 2.

Table 2. Experimental results before and after features selection

	20 features (before features selection)	5 features (after features selection)
Number of Leaves	30	15
Size of the tree	59	29
Time taken to extract features	9 seconds	2 seconds
Time taken to build model	0.06 seconds	0.01 second
Correctly classified instances	926	935

The experimental results in table 2 show that the proposed method gives better and robust data representation as it reduces the number of attributes size. Time overhead is reduced nearly 5 times in features extraction and 6 times in model building. Before feature selection, number of nodes is 59, but, after feature selection, number of leaves is 29. Therefore, our proposed method can effectively reduce the memory overhead and time overhead based on C4.5 decision tree method. The predicted accuracy of proposed method is 93.5%. Therefore, the proposed system not only can reduce building time and memory overhead but also can achieve the best accuracy.

6. Evaluation

The performance executions detailed here are based on 10-fold cross validation. The contrast the detection with 20 features and 5 top effective features are described in this experiment. Five performance metrics: Detection Rate (DR), False Positive Rate (FPR), Precision, F-Measure and Matthew's Correlation Coefficient **(**MCC) are conducted to compare the results of 20 features and top 5 features.

Table 3 shows the precision measurements comes about for 10-fold cross validation by applying five classifiers based on 20 features set. The detection rate of Random Forest is 95.4% and this is the best accuracy amongst the other classifiers. False positive rate of Random Forest is 0.042 and it correctly predicts 954 users of total 1000 users.

Table 3. Detection accuracy based on 20 features

Classifiers	DR	FPR	Precision	F-Measure	MCC
Random Forest	0.954	0.042	0.958	0.954	0.908
Decision Tree	0.926	0.080	0.939	0.900	0.805
AdaBoost	0.902	0.074	0.922	0.900	0.805
Bagging	0.938	0.060	0.940	0.938	0.876
LogitBoost	0.925	0.074	0.926	0.925	0.850

Performance matrices of applying five machine learning classifiers based on top 5 effective features are showed in table 4. The accuracy of Random Forest classifier is 95.7% and it gives the best accuracy, followed by Bagging, LogitBoost, Decision Tree and AdaBoost. False positive rate of Random Forest based on top 5 effective features is also 0.042 and this rate is equal to the false positive rate based on 20 features. The accuracy of all of the classifiers based on top 5 features is much higher than that of 20 features. 478 users are correctly predicted by Random Forest classifier based on top 5 features. By comparing the results of 20 features and top 5 features, the results of top 5 features are better than that of 20 features. Therefore, top 5 features are more efficient to detect fake users, and also require less computation time. Therefore, the proposed method achieves good accuracy in all classifications.

Table 4. Detection accuracy based on 5 effective features

Classifiers	DR	FPR	Precision	F-Measure	MCC
Random Forest	0.957	0.042	0.958	0.957	0.914
Decision Tree	0.935	0.060	0.939	0.926	0.852
AdaBoost	0.904	0.074	0.923	0.902	0.809
Bagging	0.940	0.060	0.938	0.940	0.880
LogitBoost	0.932	0.068	0.932	0.932	0.864

7. Conclusion and Future Work

We proposed an approach for recognizing fake accounts on Twitter; the proposed approach depends on deciding the importance features for the detection process. We collect and analyze features in previous studies, and then utilize these 20 features on detection based on five machine learning classifiers, namely, Random Forest, Decision Tree, AdaBoost, Bagging and LogitBoost. According to the results, Random Forest classifier achieves the best accuracy in both cases. Therefore, Random Forest classifier is the best classification algorithm for detection of fake accounts on Twitter. The proposed approach has achieved just five successful features for fake accounts discovery from 20 features. According to experimental results, the proposed method can effectively detect the fake account and can reduce in time overhead and memory overhead. In future, new features that are adaptable for other social networking sites will be extracted.

References

[1] Alex Hai Wang, "Don't follow me: Spam detection in twitter." *Security and Cryptography (SECRYPT),*

Proceedings of the 2010 International Conference on. IEEE, 2010.

[2] Amit A. Amleshwaram, Narasimha Reddy, Sandeep Yadav, Guofei Gu, and Chao Yang, "Cats: Characterizing automation of twitter spammers." *Communication Systems and Networks (COMSNETS), 2013 Fifth International Conference on*. IEEE, 2013.

[3] Ayon Chakraborty, Jyotirmoy Sundi, and Som Satapathy, "SPAM: a framework for social profile abuse monitoring." *CSE508 report, Stony Brook University, Stony Brook, NY* (2012).

[4] Chao Yang, Robert Harkreader, and Guofei Gu, "Die free or live hard? empirical evaluation and new design for fighting evolving twitter spammers." *Recent Advances in Intrusion Detection*. Springer Berlin/Heidelberg, 2011.

[5] Chu, Zi, Indra Widjaja, and Haining Wang. "Detecting social spam campaigns on twitter." *International Conference on Applied Cryptography and Network Security*. Springer, Berlin, Heidelberg, 2012.

[6] Claudia Meda, Federica Bisio, Paolo Gastaldo, and Robolfo Zunino Diten, "Machine learning techniques applied to Twitter spammers detection." *International Carnahan Conference on Security Technology*. 2014.

[7] Fabrucui Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgilio Almeida, "Detecting spammers on twitter." *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*. Vol. 6. No. 2010. 2010.

[8] M. McCord, and M.Chuah, "Spam detection on twitter using traditional classifiers." *international conference on Autonomic and trusted computing*. Springer, Berlin, Heidelberg, 2011.

[9] Po-Ching Lin, and Po-Min Huang. "A study of effective features for detecting long-surviving Twitter spam accounts." *Advanced Communication Technology (ICACT), 2013 15th International Conference on*. IEEE, 2013.