International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 03 - Issue 06 || June 2018 || PP. 141-146

Fake Accounts Classification on Twitter

Ms. Myo Myo Swe[1], Prof. Dr. Nyein Nyein Myo[2]
[1]*Web Data Mining Lab, University of Computer Studies (Mandalay), Myanmar*
[2]*Professor, Faculty of Information Science, University of Computer Studies (Mandalay), Myanmar*

Abstract: Twitter has assembled an interpersonal organization of over a billion people. Some of those individuals don't have the best advantages of their kindred people on the most fundamental level. Unfortunately, Twitterhas a noteworthy issue with spam. There are such a significant number of fake accounts on Twitter.Fake accounts are accounts that are made for many purposes such as seeking out to get information of real users, stealing genuine users' identities, and even destroying real users' reputation. For these reasons, fake accounts are very serious for online users. The aim of this paper is to classify Twitter accounts into fake accounts and legitimate accounts using machine learning classifiers. In this paper, user profile and user generated content are carefully analyzed to extractfake featuresfor classifying fake accounts on Twitter. Five machine learning classifiers such as Decorate, Random Forest, AdaBoost, Decision Tree, and Naïve Bayesian classifiers are applied to classify fake accounts on Twitter. Among these five classifiers, Decorate classifier achieves the best accuracy and the detection rate is 95.1%.

Keywords: Decorate, Random Forest, AdaBoost, Decision Tree, Naïve Bayesian

I. INTRODUCTION

Twitter is an onlineservice and long range informal communication website where individuals impart in short messages called tweets, up to 280 characters long.Its esteem is in its simplicity of sharing and getting to user-generated content, including assessments, comments, likes, news and trending topics.Twitter gives a chance to produce vast movement and income, particularly since it has a huge number of users. These open doors make twitter a prime focus of fake users.It is simple for people to recognize fake users from genuine users, however, the presence of fake users' content consumes genuine users' time and genuine users' consideration, puts genuine users in danger in getting to pernicious and hazardous substance, and debases Twitter's administrations and the general online informal organization.Fake users on Twitter utilize bunch of methods to present undesirable tweets on Twitter. These tweets act either like ads, tricks and help execute phishing assaults or the spread of malware through the inserted URLs.To pick up a more extensive reach to potential casualties, fake users are known to get to know spontaneous messages and disguise noxious components. Most of the fake accounts are bots and previous researchers detected these accounts using tweet similarity feature. To evade this feature, they used rewriting tool such as spinbot that automatically rewrite the original tweets [7].

According to Pear Analytics conducted in 2009, about 4% of all Twitter profiles are fake profiles in Twitter demographics [6]. It is an issue on Twitter, but Twitter cannot effectively solve these problems caused by fake accounts to this day.In previous works, fake accounts are detected based on profile characteristics, user generated content, users' network pattern, users' activity pattern. Former researchers utilized various features and various methods to classify fake accounts on Twitter. In our research paper, the goal is to classify fake accounts from legitimate accounts with best accuracy. Most of the researchers extract the content features based on most recently 20 tweets, some researchers extract these based on most recently 40 tweets, some researchers extract based on most recently 100 tweets and others extract these based on most recently 200 tweets. But, they extract the content-based features based on the number of most recently tweets and they did not set the exact number of tweets to extract content features. Therefore, we extracted the content-based features based on the tweeting post time. In our research work, our major contribution is to classify fake accounts from genuine accounts in an effective manner to reduce not only crawling time of users' tweets but also model building time by analyzing the users' tweets. [8]

The roadmap of the paper isdescribed as follow: section 2 presents the literature study related with the fake accounts detection. In section 3, features that are used to classify are described. Section 4 explains how our fake accounts classification works and reports the results of our proposed work.

II. LITERATURE STUDY

In year 2010, Alex Hai Wang [1] utilized content based features and user based features to detect spam profiles. They proposed a prototype for spam detection to recognize malicious user in Twitter. The "friend" and "follower" relationships are used to create a directed social graph model. To classify spam profile, Naïve

Bayesian classifier was used. According to the spam policy of Twitter, content-based features and user-based features are extracted. They evaluated the detection approach using four machine learning classifiers such as Decision Tree, Support Vector Machine, Naïve Bayesian. Naïve Bayesian classifier gave the best accuracy with 93.5%. Less dataset containing 500 users has been used. In their approach, content-based features are extracted based on most recently 20 tweets.

Many researchers detected the spam accounts on Twitter using a lot of features such as number of followers, number of followings, reputation, follower to following ratio, number of tweets, number of URLs, number of hashtags, and etc. But, Lin et.al [5] detected the spam account based on two features: URL rate and interaction rate. Some spammers' tweets did not contain URLs and hashtag to avoid the detection and URL rate and interaction rate became zero, in this case, this approach was not effective and needed to extract more features. They used most recently 20 tweets to extract the content-based features. J48 classifier was used and 86% precision was achieved.

To classify spammers on Twitter, McCord et.al. [4] utilized four content-based features and seven user-based features. Traditional machine learning classifiers like Random Forest, Support Vector Machine, Naïve Bayesian, and K-Nearest Neighbor classifiers were performed to evaluate the detection approach. In this work, reputation feature was not useful. Content-based features are extracted based on most recently 100 tweets. This approach achieved 95.7% accuracy with Random Forest classifier. They used less dataset. Their approach showed that Random Forest gave the best accuracy by applying unbalanced dataset.

Ten new features: three graph-based features, three neighbor-based features, three automation-based features and one timing-based feature for detecting spammers were proposed by Yang et.al. [9]. Their new features were very robust for evasion but these features were not easily extracted. Random Forest, Decision Tree, Decorate and Bayesian classifier were utilized to detect spammers and their best result was achieved with Bayesian classifier. In this approach, the most recent 40 tweets are used to extract the content-based features.

Chakrabortyet. al. [2]implemented a framework to recognize abusive users who post damaging substance, including hurtful URLs, porn URLs, and phishing joins and redirect away standard client and mischief the protection of informal communities. Two stages in the calculation have been utilized - first is to check the profile of a user sending companion demand to other user with respect to harsh substance and second is to check the likeness of two profiles. After these two stages, it should prescribe whether user ought to acknowledge companion ask for or not. This has been tried on Twitter dataset of 5000 clients which was gathered with REST API. Profile based, content based and timing based features are considered for separating abusive and non-damaging users. They applied four machine learning classifiers like Support Vector Machine, Decision Tree, Random Forest and Naïve Bayes. Among all these classifiers, Support Vector Machine beats all classifiers and model is performing with correctness of 89%. They utilized the most recent 200 tweets for extracting content-based features.

The previous works mentioned above were extracted the content-based features from the most recent 20 tweets, the most recent 40 tweets, the most recent 100 tweets and the most recent 200 tweets. But, in our approach, content-based features are extracted based on the most recent tweets within one month, the most recent tweets within two months and the most recent tweets within four months.

III. FEATURES EXTRACTION

This section describes the features that are used for differentiating Twitter accounts into fake accounts and legitimate accounts. Features for fake accounts classification are categorized into two: (i) user-based features and (ii) content-based features.

1. User-based Features

User-based features are achieved from user's profile and user's relationship networks. Eleven user-based features are utilized in our approach. These eleven features are (1) profile age, (2)favorite_count, (3) follower_count, (4) following_count, (5) geo_enabled_or_not, (6) follower_rate, (7) following_rate, (8) following_follower_ratio, (9) bidirectional_links, (10) protected_or_not, and (11) verified_or_not.

1.1. Profile Age

The more an accountis aged, the more it could be viewed as a decent one. Profile age is computed by subtracting the account creation date from the crawling date.

1.2. Favorite_count

The number of favorite count is high, the more it could be viewed as a suspicious one. Most of the fake accounts have highest number of favorite users.

1.3. Follower_count

If follower_count is high, this is more realistic. Fake accounts have lower number of followers.

1.4. Following_count

The higher the following count is, the more it could be viewed as a fake one. Most of the fake accounts have highest number of following users. Fake users buy followers from fake follower markets.

1.5. Geo_enabled_or_not

Fake users use Twitter for malicious purpose and they do not show their location. Nearly all of the fake users disable their geo location.

1.6. Follower_rate

This feature mirrors the notoriety of the user. The higher the folloer_rate is, the more the user is genuine.

$$Follower\_rate = \frac{Follower\_count}{Profile\_age} \quad (1)$$

1.7. Following_rate

Aggressive following behavior indicate the fake characteristic of user. In the event that the rate of following is high, the user will probably be faked.

$$Following\_rate = \frac{Following\_count}{Profile\_age} \quad (2)$$

1.8. Following_follower_ratio

According to the Socialbakers rule set, the ratio of following to follower of the account is close to 50:1, the account is suspicious.

$$Following\_follower\_ratio = \frac{Following\_count}{Follower\_count} \quad (3)$$

1.9. Bidirectional_links

Genuine users have many bidirectional links. The friends of these users are their relatives, family members, and colleges. Therefore, these users accepted their close friends and they have many bidirectional relationships.

1.10. Protected_or_not

Genuine users do not show profile details and their tweets to friend only. They protect themselves from malicious users' attack.

1.11. Verified_or_not

Most of the genuine accounts are verified by the Twitter.

2. Content-based Features

Content-based features are get from user's tweets. We use 13 content-based features in our approach. These thirteen features are (1) tweet_count, (2) retweet_count, (3) hashtag_count, (4) mention_count, (5) source, (6) spam_word_count, (7) spam_word_ratio, (8) hashtag_ratio, (9) mention_ratio, (10) mean_time_within_tweets, (11) maximum_idle_time_within_tweets (12) standard_deviation_within_tweets and (13) tweet_similarity_score.

2.1. Tweet_count

Counterfeit users present more tweets to be more dynamic and all the more ready to cooperate with others. They posted tweets in a particular time interim utilizing specific robotized tweeting instruments and programming, for example, Twitter API and AutoTwitter. In this manner, number of tweets is a quality for recognition.

2.2. Retweet_count

Twitter users can utilize retweets with @RT sign to share other users' tweets. This is another feature for recognizing counterfeit users.

2.3. Hashtag_count

To persuade the legitimate users, counterfeit users utilize well known hashtags (#). In this manner, number of hashtags is a feature to recognize fake account.

2.4. Mention_count

The higher the mention_count is, the more fake probability is.

2.5. Source

Users that post tweets from various APIs are suspicious.

2.6. Spam_word_count

The tweets of fake users contain spam words like "diet", "make money", "work from home" or "free". These spam words are caught by using spam trigger words [].

2.7. Spam_word_ratio

Higher spam_word_ratio indicates the more suspicious fake probability of an account.

$$Spam\_word\_ratio = \frac{Spam\_word\_count}{Tweet\_count} \quad (4)$$

2.8. Hashtag_ratio

Well known hashtags (#) are utilized to persuade the honest users. In this manner, number of hashtags is an element to identify malicioususers on Twitter.

$$Hashtag\_ratio = \frac{Hashtag\_count}{Tweet\_count} \quad (5)$$

2.9. Mention_ratio

Counterfeit users use hashtags to get the consideration of the honest to goodness users with the goal that hashtags proportion can be utilized for identifying counterfeit users. The high the hashtags proportion is, the more the record suspicious.

$$Mention\_ratio = \frac{Mention\_count}{Tweet\_count} \quad (6)$$

2.10. Mean_time_within_tweets(μ)

Malicious users are dominatingly seen to make posts at a speedier rate when appeared differently in relation to honest to goodness clients. This is a basic observation and we believe this feature would empower us to get phony users.

$$\mu = \frac{\sum(TimestampOfTweet(i) - TimeStampOfTweet(j))}{TotalNumberOfTweet - 1} \quad (7)$$

where,

μ = mean_time_within_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

2.11. Maximum_idle_time_within_tweets (Max)

Malicious users are believed to be discrete in their posting conduct. They post tweets in blasts. This feature would empower us to get the level of progress of this direct among fake and genuine users.

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

where,

Max = Extreme idle duration time between tweets

2.12. Standard_deviation_within_tweets(□)

This feature can recognize counterfeit users and real users.

$$\sigma = \sqrt{\frac{\sum(X - \mu)}{TotalNumberOfTweets - 1}} \quad (9)$$

where,

σ = standard deviation between tweets

X = timestamp of the tweet

μ = mean time between tweets

2.13. Tweet_similarity_score

Most fake users we watched sent fundamentally the same as tweets. Tweet similarity score are calculated using cosine sine similarity method.

IV. EVALUATION

We utilize the dataset from (Yang et al, 2012) which are gathered from April 2010 to July 2010 to evaluate the proposed system [3]. This dataset has 11000 users and their 1354616 tweets. The dataset contains 1000 fakeusers and 10000 legitimate users. In this manner, the proportion of fake to realusers is 1:10. Therefore, this is unbalanced dataset.The users who are not tweeting in English dialect are removed; the proposed technique is assessed on 1000 users which are 500 fake users and 500 real users.Experiments are carried out on a 2.10 GHz Intel Core i3 processor and 2GB RAM running on 32 bits Window 7 OS.

The performance executions are performed based on 10-fold cross validation. Precision, Recall, F-Measure and classification accuracy are calculated to compare the classification accuracy of most recent tweets

within 1 month, 2 months and 4 months. Decorate, Random Forest, Decision Tree, AdaBoost and Naïve Bayes classifiers are used for classification. Their results are compared. Table 1 shows the classification accuracy based on most recent tweets within one month. According to the results shown on Table 1, Decorate achieves the best classification result followed by Random Forest, Decision Tree, AdaBoost and Naïve Bayesian.

Table 1. Results Based on Tweets within One Month

	Accuracy	Precision	Recall	F-Measure
Naïve Bayesian	82.3%	0.759	0.946	0.842
AdaBoost	90.2%	0.922	0.878	0.9
Decision Tree	93.5%	0.939	0.93	0.935
Random Forest	94.1%	0.942	0.94	0.941
Decorate	94.3%	0.942	0.944	0.943

Table 2. Results Based on Tweets withinTwo Months

	Accuracy	Precision	Recall	F-Measure
Naïve Bayesian	83.8%	0.933	0.728	0.818
AdaBoost	89.6%	0.878	0.92	0.898
Decision Tree	94.8%	0.934	0.964	0.949
Random Forest	93.1%	0.91	0.956	0.933
Decorate	94.8%	0.943	0.952	0.949

Table 2 describes the classification results of tweets within two months. In this case, Decorate also achieve the best accuracy with 94.3% and Naïve Bayesian gives the bad results with 82.3%.

Table 3. Results Based on Tweets withinFour Months

	Accuracy	Precision	Recall	F-Measure
Naïve Bayesian	81.7%	0.918	0.696	0.792
AdaBoost	90.9%	0.888	0.936	0.911
Decision Tree	93.3%	0.934	0.932	0.933
Random Forest	95.1%	0.931	0.974	0.952
Decorate	94.9%	0.946	0.952	0.949

Table 3 shows the results based on the tweets posted within three months. Decorate also achieves the best classification result. Decorate gives the best result in three cases and results of using tweets within three months are the best. According to the results, we can be seen clearly that classification accuracy of using tweets within three months is the better than the two cases.

Table 4. Performance Comparison of tweets within one month, two months and four months

	Tweets within one Month	Tweets within two Months	Tweets within four Months
Features Extraction Time	35 seconds	59 seconds	87 seconds
Model Building Time (Decorate)	4.15 seconds	5.41 seconds	6.96 seconds
Classification accuracy	94.3%	94.8%	94.9%

The results of fake accounts classification using tweets within one month, tweets within two months and tweets within four months are compared and these results are shown in Table 4. These experiments are conducted using Decorate classifier. The classification accuracy of tweets within four months achieves the best accuracy with 94.9% that is a little more than the accuracy of tweets within one month. However, the time taken to build model of tweets within one month is 4.15 seconds and the time taken to build model of tweets within four months is 6.96 seconds that is nearly double of tweets within one month. Extracting features using tweets within two months take more time than that of the tweets within one month. The classification accuracy of using tweets within one month is acceptable. The feature extraction time and model building time of approach using tweets within one month is less than the approach using tweets within two months and four months. Therefore, the approach using tweets within one month is the best approach because it achieves good accuracy and can reduce time overhead significantly.

V. CONCLUSION

In this paper, Twitter accounts are classified into fake accounts and genuine accounts using five machine classifiers such as Naïve Bayesian, Decision Tree, Random Forest, AdaBoost and Decorate. Among these classifiers, Decorate achieves the best classification results. Contents based features are extracted based on tweets within one month, tweets within two months and tweets within four months. The approach using contents based features that are extracted from the tweets within four months achieves a little more accuracy than the approach using contents based features that are extracted from the tweets within one month but features extraction time of tweets within four months is significantly high. Therefore, our approach using tweets within one month achieves good result and can decrease time overheads. Our approach can extract features from tweets that are written in English language. Therefore, this is a limitation of our approach. Only 1KS-10KN dataset is used for evaluation, in future, we will evaluate on another dataset such as Social Honeypot dataset and we will try to extract features that are adaptable for other social networking sites such as Facebook, SinaWeibo, etc.

REFERENCES

[1] Alex Hai Wang, Don't follow me: Spam detection in twitter, Security and Cryptography (SECRYPT), Proc. of the 2010 International Conference on. IEEE, 2010

[2] AyonChakraborty, JyotirmoySundi, and SomSatapathy, SPAM: a framework for social profile abuse monitoring, CSE508 report, Stony Brook University, Stony Brook, NY, 2012

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

[4] M. McCord, and M.Chuah, Spam detection on twitter using traditional classifiers, International conference on Autonomic and trusted computing. Springer, Berlin, Heidelberg, 2011

[5] 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

[6] Raghuram, M.A., Akshay, K. and Chandrasekaran, Efficient user profiling in twitter social network using traditional classifiers, Advances In Intelligent Systems Technologies and Applications, Springer, Cham, 399-411

[7] Vishwarupe, V., Bedekar, M., Pande, M. and Hiwale, Intelligent Twitter Spam Detection: A Hybrid Approach, In Smart Trends in Systems, Security and Sustainability, 189-197, Springer, Singapore

[8] Wang, A.H., Detecting spam bots in online social networking sites: a machine learning approach, In IFIP Annual Conference on Data and Applications Security and Privacy, Springer, Berlin, Heidelberg, 2010, June , 335-342

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