Detecting Fake Followers in Twitter: A Machine Learning Approach

Ashraf Khalil, Hassan Hajjdiab, and Nabeel Al-Qirim

***Abstract*—Twitter popularity has fostered the emergence of a new spam marketplace. The services that this market provides include: the sale of fraudulent accounts, affiliate programs that facilitate distributing Twitter spam, as well as a cadre of spammers who execute large scale spam campaigns. In addition, twitter users have started to buy fake followers of their accounts. In this paper we present machine learning algorithms we have developed to detect fake followers in Twitter. Based on an account created for the purpose of our study, we manually verified 13000 purchased fake followers and 5386 genuine followers. Then, we identified a number of characteristics that distinguish fake and genuine followers. We used these characteristics as attributes to machine learning algorithms to classify users as fake or genuine. We have achieved high detection accuracy using some machine learning algorithms and low accuracy using others.**

***Index Terms*—Twitter, security, machine learning, fake follower, social networks.**

I. Introduction

Twitter has become a popular media hub where people can share news, jokes and talk about their moods and discuss news events. In Twitter users can send Tweets instantly to his/her followers. Also, Tweets can be retrieved using Twitter's real time search engine [1]. The ranking of tweets in this search engine depends on many factors, one of which is the user's number of followers. Twitter's popularity has made it an attractive place for spam and spammers of all types. Spammers have various goals: spreading advertising to generate sales, phishing or simply just compromising the system's reputation. Given that spammers are increasingly arriving on twitter, the success of real time search services and mining tools lies in the ability to distinguish valuable tweets from the spam storm [1]. There are various ways to fight spam and spammers such as URL blacklists, passive social networking spam traps, manual classification to generate datasets used to train a classifier that later will be used to detect spam and spammers [2].

So what is Twitter spam? As Twitter describes it in their website [3], Twitter spam is "a variety of prohibited behaviors that violate the Twitter Rules." Those rules include among other things the type of behavior Twitter considers as spamming, such as:

- "Posting harmful links (including links to phishing or malware sites)
- Aggressive following behavior (mass following and mass un-following for attention), particularly by automated means
- Abusing the @reply or @mention function to post unwanted messages to users
- Creating multiple accounts (either manually or using automated tools)
- Having a small number of followers compared to the number of people one is following;
- Posting repeatedly to trending topics to try to grab attention
- Repeatedly posting duplicate updates
- Posting links with unrelated tweets" (Twitter, n.d.).

Twitter actually fights spammers by suspending their accounts. But in general OSN (Online Social Networking) sites do not detect and suspend suspicious user accounts quickly. They are not willing to deploy automated methods to detect and remove spam accounts fearing that this will lead to a serious discontentment among users. Thus, they wait until a sufficient number of users report a specific account as a spam account to suspend it. However, legitimate users are unwilling to invest time to report spammers. Hence spammers are allowed more time to spread spam [4].

Spammers usually try to create a large number of social links (followers) so as to avoid being detected using spam-detection algorithms and also to enable them to quickly spread spam by utilizing one-to-many communication methods provided by OSN [4]. To acquire a large number of followers in Twitter, spammers usually follow a specific strategy. They follow other spammers and target specific legitimate users who frequently follow back. Most of these legitimate users are promoters of products or services who consider it social etiquette to follow back prospective customers. They also target the legitimate users' followers and follow them hoping they will get followed back [4].

In addition, a new sort of spam has emerged in Twitter. Spammers have started to sell fake followers to twitter users. The users who buy those accounts have various reasons. First, having a large number of followers will rank the user's tweets high in Twitter's real-time search engine. Second, there is a tremendous cachet associated with having a large number of Twitter followers. Spammers usually create a large number of followers for the first reason, while celebrities, politicians, start-ups, aspiring rock stars, and reality shows are motivated by the second reason. Fake Twitter followers momentarily made news in July 2012, when Mitt Romney's Twitter follower count jumped by more than 100,000 in one weekend, which is a much faster rate than usual [5].

In such and similar cases, it is useful and even imperative to be able to distinguish fake followers from genuine ones, thus

Manuscript received October 5, 2017; revised December 3, 2017.

Ashraf Khalil and Hassan Hajjdiab are with Abu Dhabi University, Abu Dhabi, UAE (e-mail: ashraf.khalil@adu.ac.ae, hassan.hajjdiab@adu.ac.ae).

Nabeel Al-Qirim is with United Arab Emirates University, Al Ain, UAE (e-mail: nalqirim@uaeu.ac.ae).

doi: 10.18178/ijmlc.2017.7.6.646

we have designed various machine learning algorithms meant to detect fake Twitter followers. We describe our results in this paper, the remainder of which is organized as follows. The first section presents background information about Twitter. The second section presents a literature review about Twitter spam. In the third section, we present the method we followed to detect fake followers. Finally, section four gives conclusion and suggestions for future work.

II. Background

Twitter is an information sharing network where users follow other users' newsfeed to get information about various topics. Each entry, or 'tweet', is a status message of 140 characters or less posted by a user for his or her audience of followers [1]. This message might convey their feelings, news events, plans, jokes, etc. Some users follow their friends and family to stay in touch, and others follow news outlets and celebrities' accounts. Unlike Facebook, Twitter links are bi-directional; a user can have followers and followees. In other words, if you follow someone in Twitter, you can see all his tweets if his account is public, but this does not mean he can see your tweets. If the user follows you back then he will be one of your followers and then can see your tweets [4].

Tweet content includes language conventions particular to Twitter and other peculiarities [6]:

- The string "RT" is an acronym for a "re-tweet", which is put in front of a tweet to indicate that the user is repeating or reposting someone else's tweet. For instance, "RT @Omer I'm voting for Obama".
- The hash-tag "#" is used to mark, organize and filter tweets according to topics or categories. People use the hash-tag symbol before relevant keywords in their tweets to categorize those tweets and make them more easily identifiable in Twitter Search. For example, "I love #Obama".
- The string format "@username1" indicates that a message is a reply to a user whose user name is "username1" or mentions him in the tweet. For example, "@Ahmed how are you bro?"
- Emoticons (e.g., the smiley ":-)" denoting a humorous comment) and colloquial expressions (e.g., "looove", where the repeated letter serves as emphasis) are frequently used in tweets.
- External Web links (e.g., "http://amze.ly/8K4n0t") are commonly found in tweets to provide a reference to some external sources.
- Users press the favourite button (presented in Figure 1.) to express their like of other people's tweets.

Fig. 1. Twitter's favorite button.

III. Literature Review

The prevalence of fake accounts and/or 'bots' is continuously evolving, and feature based machine-learning detection systems employing highly predictive behaviors provide unique opportunities to develop an understanding of how to discriminate between bots and humans, i.e. between real vs. fake accounts on social media. [8]

Ferrara *et al.* (2016), in their Taxonomy of Social Bots Detection Systems, have commended the use of machine-learning methods for bot detection based on the identification of highly revealing features that differentiate them from humans (real users). By focusing on differences in behavioural patterns between bots and humans, these features can be easily encoded and adopted by way of machine learning techniques to identify and classify accounts into the category of bot or human based on their observed behaviors.

Creating and using a baseline dataset of verified human and fake follower accounts, Cresci *et al.* (2015), in their seminal work, have exploited the baseline dataset to train a set of machine-learning classifiers built according to the reviewed rules and features set by media and academia (based on existing literature) respectively. Their results show that while the rules proposed by media are weak in detecting fake followers in a satisfactory fashion, features proposed in the past by academia for spam detection are a lot more accurate and efficient in providing the required results. Building on the most promising features, they have revised these classifiers, developing a Class A classifier, which is lightweight and general, yet it is able to correctly classify more than 95% of the accounts of the original training set, tested for global sensitivity.

A slightly different approach has been used by Zhang and Lu (2016), who proposed a novel method of fake account detection in Weibo, the Chinese counterpart of Twitter. This detection framework is based on the premise of why such accounts exist in the first place, i.e. 'to follow their targets en masse'. Hence it has been observed that there is high overlap between the follower lists of the customers of such accounts. They have investigated the top Weibo accounts whose follower lists duplicate or nearly duplicate each other. Using a sample-based approach in their experiment, they found 395 near-duplicates, leading to 11.90 million fake accounts (4.56 % of total users) sending 741.10 million links (9.50 % of the entire edges). They have further characterized four typical structures of spammers, clustering them into 34 groups, and analysing the properties of each group.

One of the early efforts to detect and fight spam and spammers on Twitter was by Fabricio *et al.* [1]. They manually annotated a sample that consists of 8207 users, 335 of which were spammers and 7852 were non-spammers. They selected 710 of the legitimate users to include in their collection which is twice the number of the labelled spammers. Thus, the total size of their labelled collection was 1065 users. They identified two sets of attributes that distinguish spammers from non-spammers. The first set they called "content attributes" and the second set "user behaviour attributes". The content attributes include properties of the text of the tweets posted by the users. For instance, they captured the maximum, minimum, average and median of the following metrics: number of hash tags per number of words on each tweet, number of URLs per words, number of words of each tweet, number of characters of each tweet, number of URLs on each tweet, number of hash tags on each tweet, number of numeric characters (i.e. 1,2,3) that appear in the text, number of users mentioned in each tweet, number of times the tweet has been retweeted (counted by the presence of "RT @username" in the text).

From the content attribute sets they collected, they inspected 3 attributes that they believed would largely distinguish spammers from non-spammers. Those attributes are 1) Fraction of tweets containing URLs 2) Fraction of tweets containing spam word 3) Average number of tweets that are hashtags. They drew the cumulative distribution function (CDF) for those three attributes, and they found that those attributes indeed distinguish spammers from non-spammers. Since spammers tend to have more tweets containing URLs, more tweets with spam words and higher numbers of hashtags per tweet [1].

They also identified 23 user behavior attributes, some of which are:- number of followers, ratio of followers per followees, number of tweets according to age of user account, number of times the user was mentioned, number of times the user was replied to, existence of spam words in the user screen name, number of tweets posted per day and per week, etc. Also, they inspected 3 user behavior attributes they believed would distinguish spammers from non-spammers. They drew the CDF for these 3 attributes: number of followers per number of followees, the age of the user account, and the number of tweets received. They found that spammers have a higher ratio of followers to followees and they explain this by the fact that spammers try to follow a large number of users in hope that they will be followed back. They also found that spammers' accounts are mostly new since they frequently get blocked and immediately create new accounts. Finally, they reported that non-spammers receive a much larger number of Tweets from their followees compared to spammers (Benevenuto, Magno, Rodrigues, & Almeida, 2010).

Fabricio *et al.* (Benevenuto, Magno, Rodrigues, & Almeida, 2010) used the 39 content attributes and the 23 user behavior attributes to distinguish spammers from non-spammers using a supervised machine learning algorithm which is SVM. They performed 5-fold cross-validations for testing. Their results are presented in the confusion matrix table below.

TABLE I: FABRICIO *ET AL.* [3] SPAMMER DETECTION ALGORITHM RESULTS

		Predicated Classification	
		Spammer	Non-spammer
True classification	Spammer	70.1%	29.9%
	Non-spammer	3.6%	96.4%

Besides studying the spammer detection problem, they studied the problem of detecting spam tweets [1]. Detecting spam tweets can be very useful for real time search, while detecting spammers is helpful in suspending spammers' accounts. To detect spam Tweets, they considered the following attributes: number of words from a list of spam words, number of hashtags per words, number of URLs per words, number of words, number of numeric characters on the text, number of characters, number of URLs, number of hashtags, number of mentions, number of times the tweet has been replied to, and whether the tweet was posted as a reply. They used SVM classifier with these attributes. They used a labeled collection of tweets that are labeled as spam or non-spam. The classifier was able to identify 78.5% of spam and 92.5% of non-spam.

Thomas *et al.* [3] collected 1.8 billion tweets sent by 32.9 million Twitter accounts. Then, they identified the accounts suspended by Twitter for abusive behavior. They found that 1.1 million accounts were suspended. To verify that the 1.1 million suspended accounts were spam accounts they drew a random sample consisting of 100 accounts. They then analyzed the Tweets posted by those accounts to find common spam keywords, frequent duplicate tweets, tweet content that appears across multiple accounts, the landing page of each tweet's URL and the overall posting behavior of each account. They found that 93 accounts were suspended for posting spam and unsolicited product advertisements; 3 accounts were suspended for exclusively retweeting content from major news accounts and the remaining 4 were suspended for aggressive marketing and duplicate posts. Thus they discerned that the majority of the suspended accounts are created by spammers. They found that only 8% of URLs appearing in fraudulent accounts appeared in blacklists. The problem is that those blacklists are used as techniques to identify social network spam which mean social networks should not rely on blacklists to detect spam.

IV. METHOD

A. Dataset and Labelled Collection

In order to use machine learning to identify fake twitter accounts, we needed a labeled collection of users, preclassified as fake or genuine. To acquire fake users, we created a new account and we used Fiverr, an online classified website for cheap marketing services which has several ads offering 1,000 Twitter followers for $5. We actually got 13000 Twitter followers with $5. To get genuine users we chose a university Twitter account which had 5386 twitter followers. We manually verified that those followers are real students or the other accounts managed by the university.

B. Identifying Attributes

Unlike genuine users, fake followers' accounts are created to generate revenue by following other users. Thus, we believe that they exhibit a unique behavior patter in Twitter. Although they are considered a type of spam, fake followers' accounts exhibit different behavior from twitters spammers. Twitter spammers usually post many tweets in order to spread their spam messages knowing that excessive posting of spam messages will put them at the risk of getting exposed and suspended by twitter but their goal is to send their spam message to as many users as they can [3]. Whereas, fake followers' accounts want to avoid risk of getting exposed as much as they can.

Thus, they follow a very conservative approach in twitter. They actually post less than usual users. In order to verify this assumption, we considered six attributes, which are number of followers, number of followees, number of favored Tweets, number of lists a user is a member of, number of Tweets the user has posted and number of followees per followers (our intuition was that this fraction is way too small for fake followers in comparison with real users).

To verify that those attributes are indeed useful to distinguish between fake followers' accounts and genuine users' accounts, we present the cumulative distribution function (CDF) for the six attributes. In all the figures the x-axis represents the CDF while the y-axis represents the attributes for both fake and real followers. The y-axis is represented in the logarithmic scale. In Fig. 2, we can see that CDF for fake followers is different than for real users; real users have a higher number of followers. In Fig. 3, we show that number of status updates (No. of user's Tweets) for real users are far more than fake user accounts. Also, in Fig. 4 we can see that fake followers are followed by constant number of users while real users can have as few as zero followers or high number of followers.

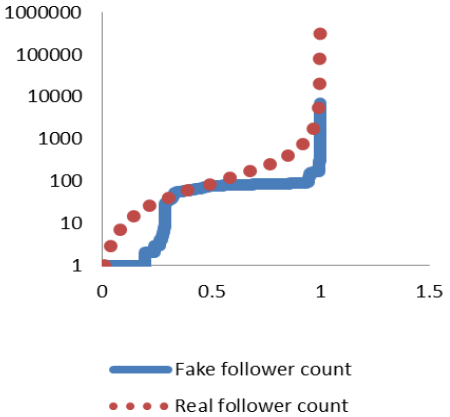

Fig. 2. CDF for number of Tweets the fake followers and real users have posted.

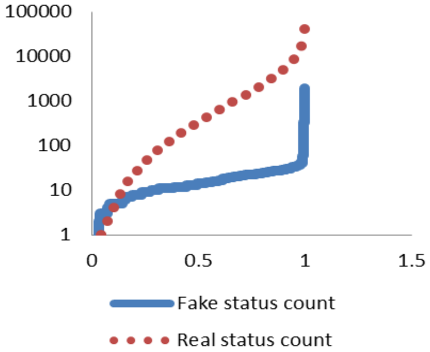

Fig. 3. CDF for number of followers for fake followers and real users' account.

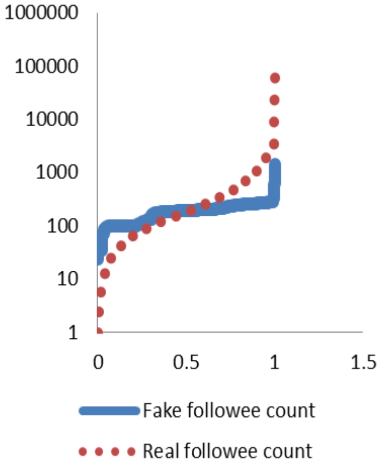

Fig. 4. CDF for number of users who follow fake follower compared to who follow real follower.

In Fig. 5, it is clear that number of tweets that fake followers favored is zero, though it is not clear in the figure because we presented the y-axis in the logarithmic scale.

To support the results we got from drawing the CDF, we decide to use the attributes selection method available on Weka [7]. This attribute selection method will rank the attributes based on their importance in classifying the dataset as fake followers and genuine followers. We used the well-known feature selection method, namely, info gain. The results are presented in Table II.

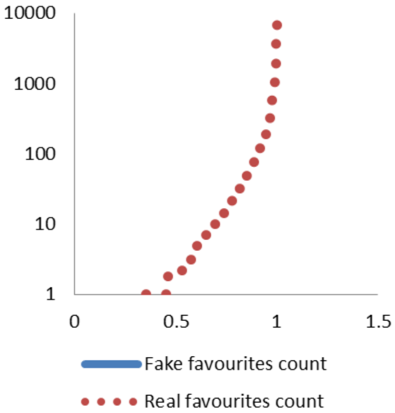

Fig. 5. CDF for number of Tweets the fake and real users favored.

TABLE II: RANKING OF THE ATTRIBUTES EXTRACTED

Rank	Info Gain	Attributes
1	0.604	status Count
2	0.529	followees Count
3	0.489	followers Count
4	0.442	favourites Count
5	0.347	followees/followers
6	0.262	Listed Count

We used 10-fold cross-validations with many machine learning algorithms. The accuracy of each algorithm is presented in Table III. We kept all the algorithms in their default settings in Weka.

TABLE III: DIFFERENT MACHINE LEARNING ALGORITHMS AND THEIR CORRESPONDING ACCURACY

The Algorithm	The accuracy
SVM	60.48%
Simple Logistic	90.02 %
Instance-based classifier using 1 nearest neighbour	98.74

V. CONCLUSION

In this paper, we investigated fake followers' accounts in Twitter. We collected a large sample which consists of 13000 fake followers and we also collected 5386 genuine followers and manually versified them. We identified number of characteristics that distinguish fake and genuine followers such as number of tweets and number of followers. Then, we used these characteristics as attributes to machine learning algorithms to classify users as fake or genuine. We achieved high detection accuracy using machine learning algorithms.

Machine learning algorithms are essential to the detection of fake accounts on Twitter and other similar social media. Knowing the key features and behaviorial differences between humans with real accounts as opposed to bots operating via fake accounts is key to the detection and elimination of fake followers. This study attempts to identify the most efficient approach for detecting fake accounts on Twitter. Our findings identify a system that can eliminate the nuisance caused by fake accounts in Twitter as well as other social networks this can be extended to.

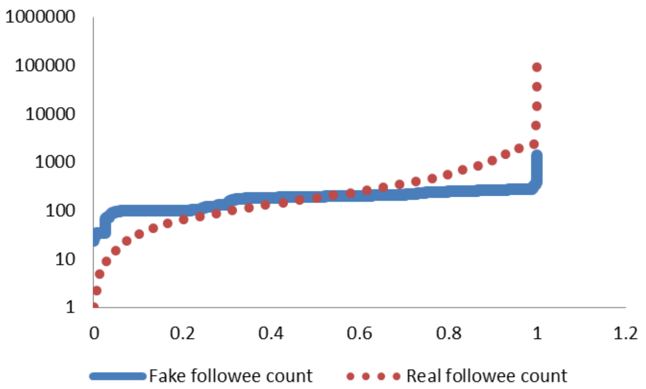

Fig. 6. CDF for number of users who follow fake follower compared to who follow real follower.

REFERENCES

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, 2010, “Detecting spammers on twitter,” in *Proc. Anti-Abuse and Spam Conference on Collaboration, Electronic messaging*, Washington: Redmond.

[2] K. Thomas, C. Grier, D. Song, and V. Paxson, “Suspended accounts in retrospect: an analysis of twitter spam,” in *Proc. the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, 2011.

[3] Twitter. *How to Report Spam on Twitter*. [Online]. Available: https://support.twitter.com/articles/64986-how-to-report-spam-on-twitter

[4] G. Saptarshi, K. Gautam, and G. Niloy, “Spammers' networks within online social networks: A case-study on Twitter,” in *Proc.World Wide Web Conference 2011*, Hyderabad, 2011.

[5] A. Considine. (August 22, 2012). Buying their way to twitter fame. [Online]. Available: http://tinyurl.com/n763xe7

[6] C. Malu, D. Umeshwar, H. Meichun, G. Riddhiman, D. Mohamed, L. Yue, and S. Mark, 2011, “LCI: A social channel analysis platform for live customer intelligence,” in *Proc. SIGMOD '11 Conf. NY*, USA.

[7] Weka 3: Data Mining Software in Java. (2012). [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka/

[8] B. Y. E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, “The rise of social bots,” *Commun. ACM*, vol. 59 No. 7, pp. 96–104, 2016.

[9] S. Cresci, R. Di, M. Petrocchi, A. Spognardi, and M. Tesconi, “Fame for sale : Efficient detection of fake Twitter followers,” *Decis. Support Syst.*, vol. 80, 2015, pp. 56–71, October 2015.

[10] Y. Zhang and J. Lu, “Discover millions of fake followers in Weibo,” *Soc. Netw. Anal. Min.*, 2016.

Ashraf Khalil earned his Ph.D. in computer science from Indiana University, Indiana, USA.

He is currently working as an associate professor at Abu Dhabi University, United Arab Emirates and the Director of Research. His research interests are ubiquitous computing, social and mobile computing, persuasive computing, and human computer interaction. The results of his research studies are published in the top conferences in the field such as CHI, CSCW, and INTERACT.

His work has also been featured in public media including the UAE's The National, the USA's National Public Radio, and Slashdot. Since he moved to the UAE in 2006, he is an active member of the Gulf community of academicians. He is interested in applying diverse technological innovations in addressing pertinent problems. He believes technology should not be confined to the work environment but should be woven into every aspect of our daily lives.

Hassan Hajjdiab received his B.Sc. in computer and communication engineering from the American University of Beirut, Lebanon and M.Sc. and Ph.D. degrees in Computer Science from the University of Ottawa, Canada.

He is an associate professor at Abu Dhabi University, United Arab Emirates and the Chair of the Computer Science and Information Technology department. His research interests are mobile computing, computer vision, image processing and remote sensing. Dr. Hajjdiab is a member of the IEEE computer Society and has authored or co-authored many technical publications in refereed journals and conference.

Nabeel Al-Qirim has a PhD in information systems from Deakin University, Australia. He is an associate professor in the College of Information Technology in UAE University.

His interests are mainly in information systems and education/pedagogy research and are mostly related to studying the adoption and diffusion of different “complex” technologies in different contexts using both qualitative (case studies, focus group, action research) and quantitative (surveys) methods.

He authored one research book and three edited books in the area of Small to Medium-Sized Enterprises (SMEs) and E-Commerce. He published more than 100 research papers in refereed and highly-impact international outlets (i.e., PACIS, AMCIS, ECIS, IFIP, BLED, EM, Computer & Education, ECRA, Medical Informatics, Telemedicine and eHealth, etc.). He is also a member in leading academic and professional associations (IEEE senior member, ACM, AIS) and in the editorial advisory board of several journals (JIKM, IJCEC, JECO, BPMJ, IJNOV, IDI, IJCA).