Detecting Fake Followers in Twitter: A Machine Learning Approach

Prof. Parinita J Chate, Assistant Professor, Computer Engineering Department, Bharati Vidyapeeth's College of Engineering, Lavale, Pune, India.

Abstract—

Social networking sites such as Twitter and Facebook have been part of many people's lives. Their interaction with social networking has affected their life forever. Accordingly, social networking sites have become among the main channels that are responsible for vast dissemination of different kinds of information during real time events.

Some users create & buy fake followers to increase their popularity 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 10000 purchased fake followers and 5000 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.

Keywords - Twitter, security, Machine learning algorithms, fake follower, social networks, Correlation

I. INTRODUCTION

Microblogging services such as Twitter have become important tools for personal communication as well as spreading news. Twitter users can "follow" accounts that they find interesting, and start receiving status updates that these accounts share in real-time.

In this paper we are highlighting and identifying the fake followers in the platform twitter we approach the methods which can identify fake followers depending upon the some quantity of user used and follow, 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

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.).

Various OSNs have a different measure of user following and social reputation, like followers on Twitter and Instagram, likes on Facebook and ratings on Yelp. However, these reputation metrics can be manipulated in several ways. One of the most prevalent methods to alter social reputation is online black-market services, which help the users to increase their follower/like count. There exist several online services from where an OSN user can purchase bulk followers and likes. The follower black-market services also often have a collusion network model where a user can gain followers for free by following other customers of the service. Users exploit these services to inflate their social media metrics such as – followers, likes and shares (of the user post) in the hope to become more influential and popular on the network.

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. In such and similar cases, it is useful and even imperative to be able to distinguish fake followers from genuine ones, thus we have designed various machine learning algorithms meant to detect fake Twitter followers.

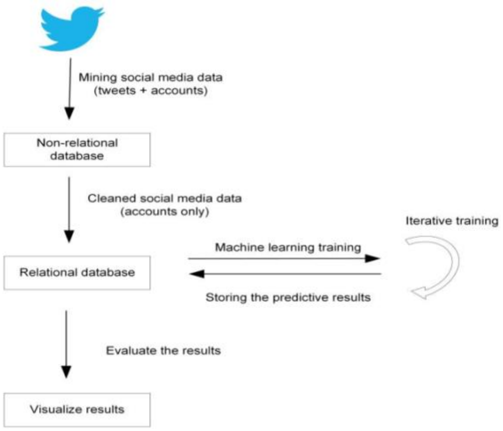


Figure 1: process for indentifying fake followers

II. RELATED WORK

A fake Twitter followers is considered as one form of deception (i.e., deception in both the content and the personal information of the profile as well as deception in having the profile follow others not because of personal interest but because they get paid to do so)." The second characterization for deception is exactly the one we deal with in our paper. We specifically consider fake followers as those Twitter accounts appropriately created and sold to customers, which aim at magnifying their influence and engagement to the eyes of the world, with the illusion of a big number of followers.

So defined fake followers are only an example of anomalous accounts which are spreading over Twitter, Anomalies have been indeed identified in the literature as either spammers (i.e. accounts that advertise unsolicited and often harmful content, containing links to malicious), or bots (i.e., computer programs that control social accounts, as stealthy as to mimic real users), or cyborgs (i.e., accounts that interweave characteristics of both manual and automated behaviour). Finally, there are fake followers,i.e., those accounts massively created to follow a target account and that can be bought from online accounts markets, as Intertwitter.com. We would like to remark that fake followers could be seen as a macro category in the scenario of Twitter anomalous accounts, since subsets of fake followers could include bots or even stolen accounts of real users.

Hereafter, we illustrate methods and approaches that have been proposed in the academic literature to analyze the different phenomena, and we highlight similarities and differences with our approach.

2.1. Spam detection

In recent years, spam detection on Twitter has been the matter of several investigations, approaching the issue from several points of view, Twitter spam has long been a critical but difficult problem to be addressed. So far, researchers have developed a series of machine learning-based methods and blacklisting techniques to detect spamming activities on Twitter. According to our investigation, current methods and techniques have achieved the accuracy of around 80%. However, due to the problems of spam drift and information fabrication, these machine-learning based methods cannot efficiently detect spam activities in real-life scenarios. Moreover, the blacklisting method cannot catch up with the variations of spamming activities as manually inspecting suspicious URLs is extremely time-consuming. In this paper, we proposed a novel technique based on deep learning techniques to address the above challenges. The syntax of each tweet will be learned through WordVector Training Mode. We then constructed a binary classifier based on the preceding representation dataset. In experiments, we collected and implemented a 10-day real Tweet datasets in order to evaluate our proposed method. We first studied the performance of different classifiers, and then compared our method to other existing text-based methods. We found that our method largely outperformed existing methods. We further compared our method to non-text-based detection techniques. According to the experiment results, our proposed method was more accurate

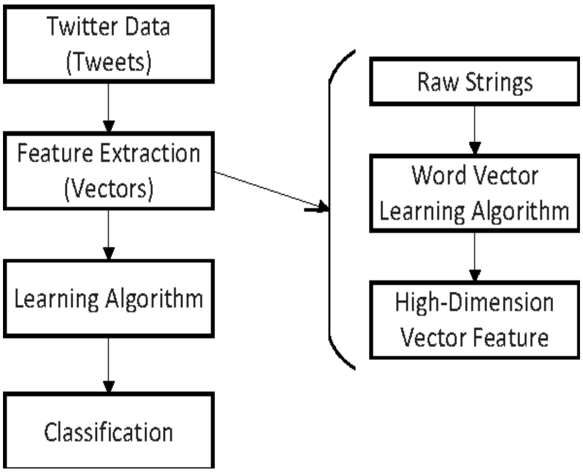

Twitter Data (Tweets)
Feature Extraction (Vectors)
Learning Algorithm
Classification
Raw Strings
Word Vector Learning Algorithm
High-Dimension Vector Feature

Figure 2: New Twitter classification workflow based on deep learning

2.2. Detecting spammers in social networks

The study presented in focuses on spambot detection. The authors exploit five characteristics that can be gathered crawling an account's details, both from its profile and timeline. The characteristics are:

- the number of friends;
- the number of tweets;
- the content of tweets;
- the URL ratio in tweets;
- the relation between the number of friends and followers.

2.3. Fighting evolving Twitter spammers

Twitter spammers often modify their behavior in order to evade existing spam detection techniques. Thus, they suggested considering some new features, making evasion more difficult for spammers. Beyond the features directly available from the account profile lookup, the authors propose some graph-, automation-, and timing-based features.

Table 1: Evaluation of the single feature

feature description		evaluation metrics	
Stringhini		I gain	Pcc
1	number of friends	0.263	0.03
2	number of tweets	0.621	0.289
3	content of tweets	0.444	0.74
4	URL ratio in tweets	0.401	0.353
5	Friends / (followers^2)	0.733	0.169
Yang			
1	age	0.539	0.436
2	bidirectional links ratio	0.905	0.875
3	avg. followers of friends	0.327	0.254
4	avg. tweets of friends	0.203	0.235
5	friends / med. foll. of friends	0.336	0.102
6	api ratio	0.544	0.635
7	api url ratio	0.058	0.113
8	api tweet similarity	0.46	0.748
9	following rate	0.355	0.214

2.4. Do spammers tweet more frequently than legitimate users?

We calculated the average distribution of tweets for each user by subtracting the date of first tweet from the date of last tweet and divided by total number of tweets. The average number of tweets per day was higher among spam accounts than legitimate accounts (legitimate: mean=6.7; spam: mean=8.66).

We also compared retweet and @reply behavior from legitimate accounts versus spammers. Results showed that 19 percent of legitimate tweets and 21 percent of spam tweets contained other hashtags within the set of #robotpickuplines tweets. Use of @replies was slightly higher with 26 percent legitimate uses and 24.8 percent spam uses. A chi–square test was performed and there was no significant difference in retweets or replies between spammers and legitimate users, but there was significant different in number of tweets $\chi^2(1,n=300)=4.464$, $\rho<0.05$, and use of hashtags $\chi^2(1,n=300)=3.847$, $\rho<0.05$.

2.5. Do spammers have more friends than followers?

We then examined structural properties of the network. We hypothesized that follower–to–friend ratio would be higher for legitimate accounts than for spammers because spambots may auto–follow Twitter users *en masse*. We calculated the ratio to be not significantly different (1.38 for legitimate, 1.12 for spammers). However, the total number of followers and friends for spammers was three times that of legitimate users (see Figures 3 and 4).

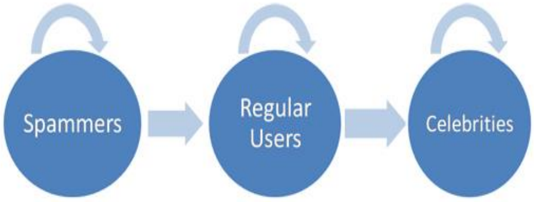


Figure 3: Average number of spam follower and friends.

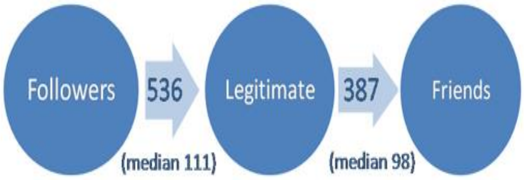


Figure 4: Number of legitimate follower and friends.

The following points shows some points of span detection

- Cyborgs and bots detection
- Fake followers and Account Markets Analysis
- Differences and similarities with our approach
- Grey literature and Online Blogs

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.

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.

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 2: FABRICIO SPAMMER DETECTION ALGORITHM RESULTS

True classification		Predicated Classification	
		Spammer	Non-spammer
	Spammer	69.1%	28.9%
	Non-spammer	3.4%	90.4%

Besides studying the spammer detection problem, they studied the problem of detecting spam tweets

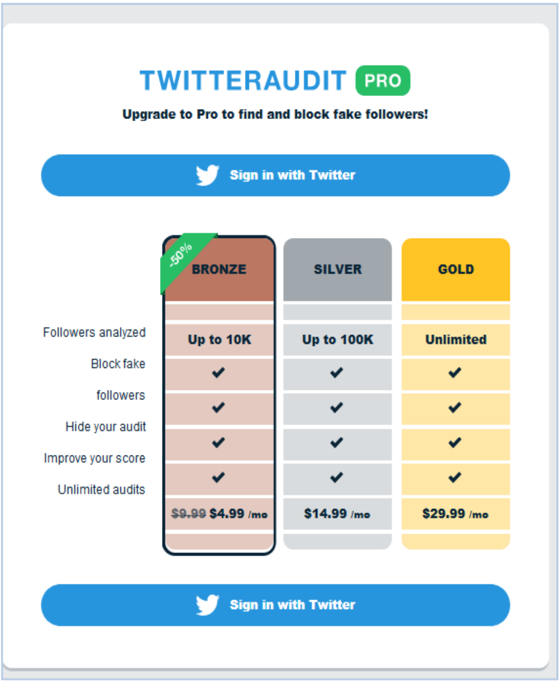
IV. TWITTER AUDIT

Based on analysis and detecting some of the tools are used for auditing twitter fake followers; one of the tools we used is TwitterAudit Pro.

TwitterAudit Pro is a tool to help you find and block fake followers on Twitter. TwitterAudit is the most popular tool, widely used and trusted by over 1 million Twitter users. The Pro service takes a deeper look at your account and lets you block followers, as well as make your account private. You also get unlimited free basic audits for any other user.

Does it really work?

TwitterAudit finds "low-quality" followers in the sense that they may be either bots, troll-types, or simply inactive. Keep in mind that no algorithm is perfect and as a result real followers may sometimes be mis-identified as fake. You will be able to mark a follower as "not fake", which will still improve your score, once that follower is manually reviewed.



TWITTERAUDIT PRO

Upgrade to Pro to find and block fake followers!

Sign in with Twitter

	-50% BRONZE	SILVER	GOLD
Followers analyzed	Up to 10K	Up to 100K	Unlimited
Block fake followers	✓	✓	✓
Hide your audit	✓	✓	✓
Improve your score	✓	✓	✓
Unlimited audits	✓	✓	✓
	~~$9.99~~ $4.99 /mo	$14.99 /mo	$29.99 /mo

Sign in with Twitter

Figure 5: How many of your followers are real.

V. METHOD

5.1. Data Collection and Reduction

We collected data from 507 active Twitter users who col-lectively provided us with a corpus of 522,368 tweets span-ning the 15 months between August 2010 and October 2011. In addition to the tweets, we also have snapshots of friends and followers taken at periodic intervals (a total of five periods, each approximately three months in duration). We were interested in discovering the relationship between the factors discussed above within each three-month period and the subsequent changes in follower counts at the end of that period. To build our dataset, Twitter accounts were ob-tained by recording unique account IDs that appeared on the public timeline during a two-week period in August 2010, and then screened for certain attributes. The subset selected for inclusion in this study consisted of those ac-counts that met the following four criteria when sampled approximately every three months.

Tweet in English, as determined by inspecting the users' profiles for the designated language via Tweepy2, a Twitter API library for Python, as well as Python's Nat-ural Language Tool Kit3 (NLTK) for language detection on the users' 20 most recent tweets. This filter is neces-sary for our linguistic predictors (described later), alt-hough it may restrict the generalizability of our results.

VI. DATASET AND LABELLED COLLECTION

6.1. Baseline datasets - We present the datasets of Twitter accounts we used to conduct our empirical study and that will be used throughout the paper. We detail how we collected each of them and how we verified if they were genuine humans or fake followers. Despite the final size of the baseline dataset, to perform our research, we altogether crawled 3 millions of Twitter accounts and about 1 millions of tweets. To foster in-vestigation on the novel issue of fake Twitter followers, our baseline dataset has been made publicly available for the research purpose.

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.

6.2. Collecting Labeled Data

ML problems start with data - preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called *labeled data*. In supervised ML, the algorithm teaches itself to learn from the labeled examples that we provide.

Each example/observation in your data must contain two elements:

- The target – The answer that you want to predict. You provide data that is labeled with the target (correct answer) to the ML algorithm to learn from. Then, you will use the trained ML model to predict this answer on data for which you do not know the target answer.
- Variables/features – These are attributes of the example that can be used to identify patterns to predict the target answer.

VII. ANALYSIS OF THE IDENTIFIED CUSTOMERS

In this section, we analyze the characteristics of the 684 market customers that we identified. These customers are detected by the dynamic classifier, so not surprisingly, their dynamic characteristics are very similar to the ones of the customers in the training dataset Ac. We found that their static characteristics show strong customer-like signals too: Figure 6 shows the distribution of the number of followers of the identified customers, compared to legitimate users. The identified customers typically have more than 1,000 followers, which is more than the number of followers for 80% of the regular Twitter user population. In addition, Figure 7 shows that the follower-to-friend ratio of identified customers is typically higher than one.

As we said, by manually looking at the identified customers we found that those accounts mostly belong to small businesses or wannabe celebrities, who try to boost their popularity. We then wanted to assess whether the purchase of followers actually helps in this process. To this end, we analyzed the influence score of these customers, according to Klout . The CDF of the influence of the identified customers is shown in Figure 8, indicating that about half the customers have a Klout score lower than 45. However, the median Klout score of Twitter accounts is precisely 45 (on a scale 1-100). This shows that, although purchasing followers can boost an account's social network, it does not really help in making the account popular. Since the followers did not willingly follow the profile, it is unlikely that they will engage the user, and share her content.

As a last element, we wanted to understand whether Twitter is detecting and blocking these customer accounts. After one week from detecting them, only two accounts had been suspended by Twitter. This shows that it is hard for Twitter to detect which accounts purchased followers, and that the type of techniques proposed in this paper could actively help Twitter in fighting this phenomenon.

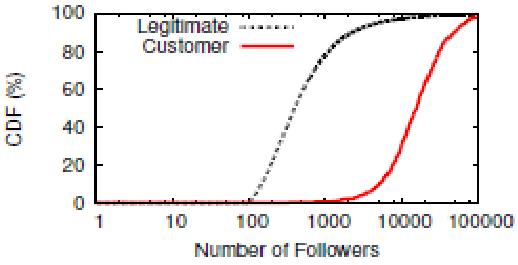


Figure 6: Followers of identified customers and legitimate users.

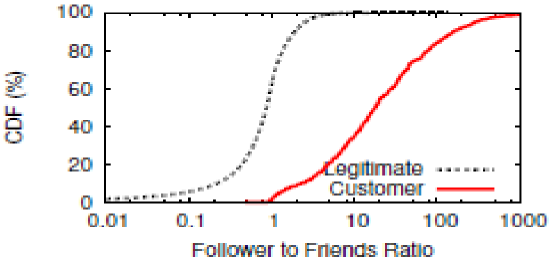


Figure 7: Follower-friend ratio of identified customers and legitimate users.

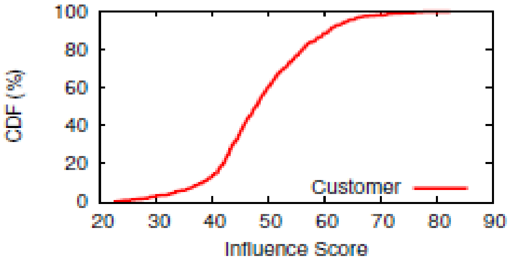


Figure 8: Influence (Klout) score of identified customers.

VIII. 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. 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. 9, we can see that CDF for fake followers is different than for real users; real users have a higher number of followers. In Fig. 10, 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. 11 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.

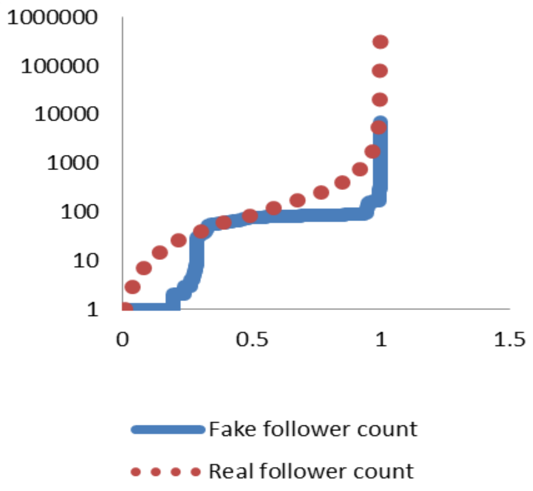


Figure 9: CDF for number of Tweets the fake followers and real users have posted.

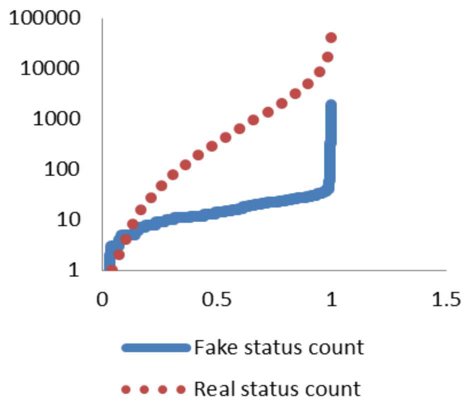


Figure 10: CDF for number of followers for fake followers and real users' account.

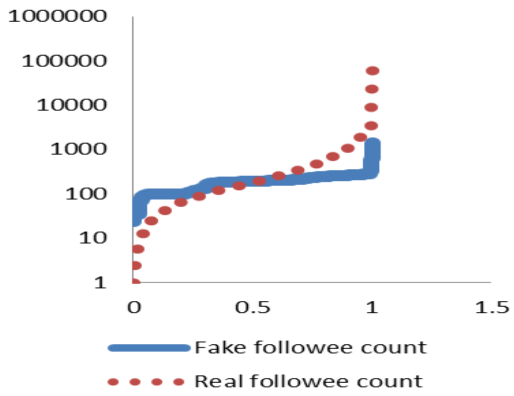


Figure 11: CDF for number of users who follow fake follower compared to who follow real follower.

In Fig. 12, 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. 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 3.

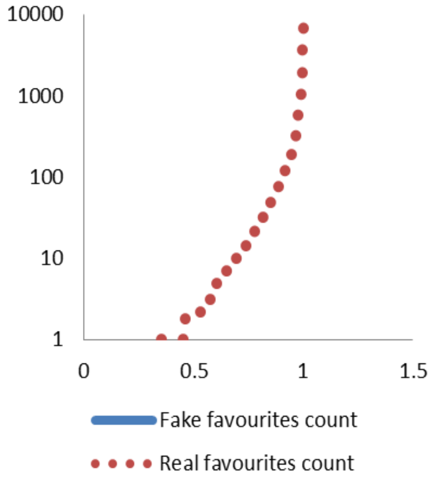


Figure 12: CDF for number of Tweets the fake and real users favored.

TABLE 3: 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 4. We kept all the algorithms in their default settings in Weka.

TABLE 4: 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

IX. CONCLUSION

In this paper, we focused on efficient techniques for fake Twitter follower’s detection. To reach the goal, we investigated fake followers‟ accounts in Twitter. We collected a large sample which consists of 10000 fake followers and we also collected 5000 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.

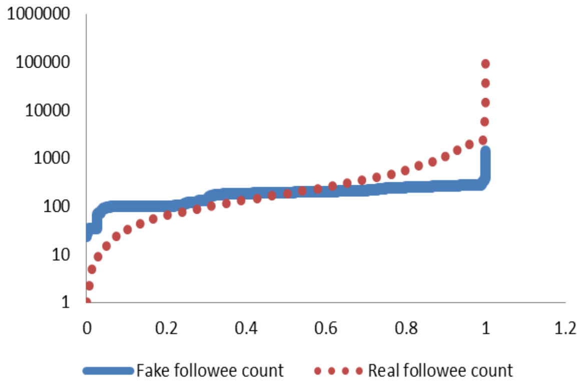


Figure 13: CDF for number of users who follow fake follower compared to who follow real follower.

Acknowledgements

The authors would like to thank Bharati Vidyapeeth’s College of Engineering, Lavale, Pune for supporting and guiding me. Also I am thankful to Principal Sir, Staff who encourage me all the time and willingness to provide assistance whenever it was needed.

References

1. C. Smith, By The Numbers: 254 Amazing Twitter Statistics.,

http://goo.gl/o1lNi8, Last checked 03/09/15 (Oct. 2014).

2. Statistic Brain, Twitter statistics, http://goo.gl/XEXB1, Last checked 03/09/15 (July 2014).
3. CAI, Z., AND JERMAINE, C. The latent community model for detecting sybils in social networks. In Symposium on Network and Distributed System Security (NDSS) (2012).
4. 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.
5. BILGE, L., STRUFE, T., BALZAROTTI, D., AND KIRDA, E. All Your Contacts Are Belong to Us: Automated Identity Theft Attacks on Social Networks. In World Wide Web Conference (WWW) (2009).
6. Twitter. How to Report Spam on Twitter. [Online]. Available: https://support.twitter.com/articles/64986-how-to-report-spam-on-twitter
7. 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.
8. J. H. Parmelee and S. L. Bichard, Politics and the twitter revolution: How tweets influence the relationship between political leaders and the public, Lexington books, 2011.
9. L. Bilge, T. Strufe,D. Balzarotti, and E.Kirda, “All your contacts are belong to us: automated identity theft attacks on social networks,” in Proceedings of the 18th international conference on World wide web, pp. 551–560,Madrid, Spain, 2009.
10. S. Gurajala, J. S. White, B. Hudson, B. R. Voter, and J. N. Matthews, ``Pro_le characteristics of fake Twitter accounts," Big Data Soc., vol. 3, no. 2, p. 2053951716674236, 2016, doi: 10.1177/2053951716674236.
11. C. Xiao, D. M. Freeman, and T. Hwa, ``Detecting clusters of fake accounts in online social networks," in Proc. 8th ACMWorkshop Artif. Intell. Secur., 2015, pp. 91_101.
12. Ablon, L., Libicki, M. C., and Golay, A. A. (2014). Markets for Cybercrime Tools and Stolen Data: Hackers’ Bazaar. RAND Corporation, Santa Monica, CA.