International Journal of Research in Information Technology (IJRIT)

www.ijrit.com ISSN 2001-5569

Finging Unwanted Messages in Twitter

G. Deepak , V. Shankar

Assoc. Prof. Dept of CSE, Kakatiya Institute of Technology and Science, Warangal.

Prof. & HOD Dept of CSE, Kakatiya Institute of Technology and Science, Warangal.

ABSTRACT: Twitter is social networking site in which people share their views like facebook , whatsapp. It is also called microblogging. Its popularity makes spammers to post unwanted messages. We divide the spammers into categories like bot ,cyborg. For this we divide them we propose a classification system that includes the following four parts: 1) an entropy-based component, 2) a spam detection component, 3) an account properties component, and 4) a decision maker. It uses the combination of features extracted from an unknown user to determine the likelihood of being a human, bot, or cyborg.

1. INTRODUCTION

Now-a-days twitter plays a vital role in social networking site in which people share their views like daily messages and polical information . Twitter was released in the year 2006 and became very popular within short period of time .Its simplicity makes the popularity. The tweet size is limited to 140 characters. Hashtag, namely hour . It ranks the 10th on the top 500 site list according words or phrases prefixed with a # symbol. In November 2009, Twitter by topic. For example, #Justin Bieber and #Women's World emphasized its value as a news and information network by Cup are the two trending hashtags on Twitter in 2011 . changing the question above the tweet input dialog box Symbol @ followed by a username in a tweet enables the from "What are you doing" to "What's happening." To direct delivery of the tweet to that user. Unlike most online some extent, Twitter has transformed from a personal social networking sites (i.e., Facebook and MySpace), microblogging site to an information publish venue. Many Twitter's user relationship is directed and consists of two traditional industries have used Twitter as a new media ends, friend and follower. In the case where the user A adds channel. We have witnessed successful Twitter applications B as a friend, A is a follower of B while B is a friend of A. In business promotion , customer service , political Twitter terms, A follows B (namely, the following relation-campaigning , and emergency communication. ship is unidirectional from A to B). B can also add A as his friend .The growing user population and open nature of Twitter friend (namely, following back or returning the follow), but have made itself an ideal target of exploitation from is not required. When A and B follow each other, the automated programs, known as bots. Like existing bots in relationship becomes bidirectional. From the standpoint of other web applications (i.e., Internet chat

, blogs and information flow, tweets flow from the source (author) to online games), bots have been common on Twitter. subscribers (followers). More specifically, when a user posts Twitter does not inspect strictly on automation. It only requires the recognition of a CAPTCHA image during registration.

2 .RELATED WORK:

Twitter had became very popular in the year 2006, and there are some tweets, like news and blog updates. This supports with the related literature in twittering . Twitter's goal of becoming a news and information net- understand microblogging usage and communities. On the other hand, malicious bots have been greatly studied over 50,000 Twitter users and categorized exploited by spammers to spread spam. The definition of their posts into four main groups: daily chatter (e.g., "going spam in this paper is spreading malicious, phishing, or out for dinner"), conversations, sharing information or unsolicited commercial content in tweets. These bots URLs, and reporting news. Their work also studied 1) the randomly add users as their friends, expecting a few users growth of Twitter, showing a linear growth rate; 2) its to follow back. In this way, spam tweets posted by bots network properties, showing the evidence that the network display on users' homepages. Enticed by the appealing text is scale-free like other social networks and 3) the content, some users may click on links and get redirected to geographical distribution of its users, showing that most spam or malicious sites. Twitter users are from the US, Europe, and Japan. malicious bots and spam tweets, their twittering experience ,the whole Twitter community Twitter users and classified their roles by follower-to-following will be hurt. The objective of this paper is to characterize the following ratios into three groups: 1) broadcasters, which automation feature of Twitter accounts, and to classify them have a large number of followers 2) acquaintances, which into three categories, human, bot, and cyborg, accordingly. have about the same number on either followers or This will help Twitter manage the community better and following; and 3) miscreants and evangelists which follow a large number of other users but have

In the paper, we first conduct a series of measurements few followers. Wu et al. [16] studied the information to characterize the differences among human, bot, and diffusion on Twitter, regarding the production, flow, and cyborg in terms of tweeting behavior, tweet content, and consumption of information. Kwak et al. [17] conducted a account properties. By crawling Twitter, we collect over thorough quantitative study on Twitter by crawling the 500,000 users and more than 40 million tweets posted by entire Twittersphere. Their work analyzed the follower- them. Then, we perform a detailed data analysis, and find a following topology, and found nonpower-law follower set of useful features to classify users into the three classes. distribution and low reciprocity, which all mark a deviation Based on the measurement results, we propose an auto- from known characteristics of human social networks. Kim mated classification system that consists of four major et al. [18] analyzed Twitter lists as a potential source for discovering latent characters and interests of users.

3. Components

Twitter list consists of multiple users and their tweets.

1. the entropy component uses tweeting interval as a Their research indicated that words extracted from each list measure of behavior complexity, and detects the are representative of all the members in the list even if the periodic and regular timing that is an indicator of words are not used by the members. It is useful for automation targeting users with specific interests.

2. The spam detection component uses tweet content to In addition to network-related studies, several previous check whether text patterns contain spam or not3; works focus on sociotechnological aspects of Twitter.

3. the account properties component employs useful information such as its use in the workplace or during account properties, such as tweeting device makeup, major disaster events. URL ration, to detect deviations from normal; and Twitter has attracted spammers to post spam content,

4. The decision maker is based on Random Forest, and due to its popularity and openness.This method uses a static analysis method which was combined with automated reasoning. This method approach assumes that there is no tautology in an SQL query generated dynamically which was verified. This method is efficient in detecting SQL injection attacks but other SQL injection attacks except for tautology cannot be detected.

4. EVALUATION

In this section, we first evaluate the accuracy of our classification system based on the ground-truth set. Then, we apply the system to classify the entire data set of over 500,000 users collected. With the classification results, we further speculate the current composition of Twitter user population. Finally, we discuss the robustness of the proposed classification system against possible evasions.

4.1 Methodology

The components of the classification system collaborate in the following way. The entropy component calculates the entropy (and corrected conditional entropy) of intertweet delays of a Twitter user. The entropy component only processes logs with more than 100 tweets.9 This limit helps reduce noise in detecting automation. A lower entropy indicates periodic or regular timing of tweeting behavior, a sign of automation, whereas a higher entropy implies irregular behavior, a sign of human participation. The spam detection component determines if the tweet content is either spam or not, based on the text patterns it has learned. The content feature value is set to 1 for spam but 0 for nonspam. The account properties component checks all the properties , and generates a real-number-type value for each property. Given a Twitter user, the above three components generate a set of features and input them into the decision maker. For each class, namely human, bot, and cyborg, the decision maker computes a classification score for the user, and classifies it into the class with the highest score. The training of the classification system and cross validation of its accuracy are detailed as follows.

4.2 Classification System Training

The spam detection component of the classification system requires training before being used. It is trained on spam and nonspam data sets. The spam data set consists of spam tweets and spam external URLs, which are detected during the creation of the ground-truth set. Some advanced spam bots intentionally inject nonspam tweets (usually in the format of pure text without URLs, such as adages10) to confuse human users. Thus, we do not include such vague tweets without external URLs. The nonspam data set consists of all human tweets and cyborg tweets without external URLs. Most human tweets do not carry spam. Cyborg tweets with links are hard to determine without checking linked webpages. They can be either spam or nonspam. Thus, we do not include this type of tweets in either data set. Training the component with up-to-date spam text patterns on Twitter helps improve the accuracy. In addition, we create a list of spam words with high frequency on Twitter to help the Bayesian classifier capture
spam content.

4.3 Cross Validation of Accuracy

We use Weka, a machine learning tool , to implement the Random Forest-based classifier. We apply cross validation with 10-folds to train and test the classifier over the ground-truth set. The data set is randomly partitioned into 10 complementary subsets with equal size. In each round, one out of 10 subsets is retained as the test set to validate the classifier, while the remaining nine subsets are used as the training set to train the classifier. At the beginning of a round, the classifier is reset and retrained. Thus, each round is an independent classification procedure, and does not affect subsequent ones. The individual results from 10 rounds are averaged to generate the final estimation. The advantage of cross validation is that, all samples in the data set are used for both training and validation, while each sample is validated exactly once. The confusion matrix demonstrates the classification results. The “Actual” denote the actual classes of the users, and the “Classified” columns denote the classes of the users as decided by the classification system. For example, the cell in the junction of the “Human” row and column means that 1,972 humans are classified (correctly) as humans, whereas the cell of “Human” row and “Cyborg” column indicates that 27 humans are classified (incorrectly) as cyborgs. There is no misclassification between human and bot. We examine the logs of those users being classified by mistake, and analyze each category as follows: For the human category, 1.4 percent of human users are classified as cyborg by mistake. One reason is that the overall scores of some human users are lowered by spam content penalty. The tweet size is up to 140 characters. Some patterns and phrases are used by both human and bot, such as “I post my online marketing experience at my blog at http:// bit.ly/xT6klM. Please ReTweet it.” Another reason is that the tweeting interval distribution of some human users is slightly lower than the entropy means, and they are penalized for that. For the bot category, 2.3 percent of bots is wrongly categorized as cyborg. The main reason is that, most of them escape the spam penalty from the spam detection component. Some spam tweets have very obscure text content, like “you should check it out since it’s really awesome. http://bit.ly/xT6klM”. Without checking the spam link, the component cannot determine if the tweet is spam merely based on the text. For the cyborg category, 3.3 percent of cyborgs are misclassified as human, and 5.1 percent of them are misclassified as bot. In analyzing those samples misclassified as human, we find out a common fact that, some owners of cyborg accounts interact with followers from time to time, and use manual devices to reply or retweet to followers. Besides, the manual behavior of owners increases the entropy of tweeting

interarrivals. The two factors tend to influence the classifier to make decisions in favor of human. The difficulty here is that, a cyborg can be either a human-assisted bot or a bot-assisted human. A strict policy could categorize cyborg as bot, while a loose one may categorize it as human.

Confusion Matrix

9. The intertweet span could be wild on Twitter. An account may be inactive for months, but suddenly tweets at an intensive frequency for a short-term, and then enters hibernation again. It generates noise to the entropy component. Thus, the entropy component does not process logs with less than 100 tweets. Besides, in practice, it is nearly impossible to determine automation based on a very limited number of tweets.

10. A typical content pattern is listed as follows: Tweet 1, A friend in need is a friend in deed. Tweet 2, Danger is next neighbor to security. Tweet 3, Work home and make $3k per month. Check out how, http:// tinyurl.com/bF234T. There is negligible misclassification between human and bot. The classifier clearly separates these two classes.

Overall, our classification system can accurately differentiate human from bot. However, it is much more challenging for a classification system to distinguish cyborg from human or bot. After averaging the true positive rates of the three classes with equal sample size, the overall system accuracy can be viewed as 96.0 percent. Among the set of features used in classification, some play a more important role than others. Now, we evaluate the discrimination weight of each feature. In every test, we only use one feature to independently cross validate the ground-truth set. Table 4 presents the results sorted on accuracy. The entropy feature has the highest accuracy at 82.8 percent. It effectively captures the timing difference between regularity of automated behavior and complexity of manual behavior. Limited by tweet size, bot usually relies on URLs to redirect users to external websites. This fact makes the URL ratio feature have a relatively high accuracy at 74.9 percent. Recognizing the tweeting device makeup (manual or automated) and detecting spam content also help the classification. By comparing the collective performance in Table 3 and individual performance in Table 4, we observe that, no single feature works perfectly well, and the combination of multiple features improves the classification accuracy.

4.4 **Twitter Composition**

We further use the classification system to automatically classify our whole data set of over 500,000 users. We can speculate the current composition of Twitter user population based on the classification results. The system classifies 53.2 percent of the users as human, 36.2 percent as cyborg, and 10.5 percent as bot. Thus, we speculate the population proportion of human, cyborg and bot category roughly as on Twitter.

5.5 Resistance to Evasion

Now, we discuss the resistance of the classification system to possible evasion attempts made by bots. Bots may deceive certain features, such as the followers to friends ratio as mentioned before. However, our system has two critical features that are very hard for bots to evade. The first feature is tweeting device makeup, which corresponds to the manual/auto device percentage in Table 4. Manual device refers to web and mobile devices, while auto device refers to API and other auto-piloted programs (see Section 3.3, Q5). Tweeting via web requires a user to log in and manually post via the Twitter website in a browser. Posting via HTTP form is considered by Twitter as API.

Furthermore, currently it is impractical or expensive to run a bot on a mobile device to frequently tweet. As long as Twitter can correctly identify different tweeting platforms, device makeup is an effective metric for bot detection. The second feature is URL ratio. Considering the limited tweet length that is up to 140 characters, most bots have to include a URL to redirect users to external sites. Thus, a high URL ratio is another effective metric for bot detection. If we exclude the features of URL ratio and tweeting device makeup, and retrain the classifier, the overall classification accuracy drops to 88.9 percent. Bot may try to bypass some features when it knows our detection strategy. For timing entropy, bot could mimic human behaviors but at the cost of much reduced tweeting frequency. For spam content, bot could intermix spam with ham tweets to dilute spam density. We will continue to explore new features emerging with the Twitter development for more effective bot detection in the future.

5. CONCLUSION

In this paper, we have studied the problem of automation by bots and cyborgs on Twitter. As a popular web application, Twitter has become a unique platform for information sharing with a large user base. However, its popularity and very open nature have made Twitter a very tempting target for exploitation by automated programs, i.e., bots. The problem of bots on Twitter is further complicated by the key role that automation plays in everyday Twitter usage.

To better understand the role of automation on Twitter,we have measured and characterized the behaviors of humans, bots, and cyborgs on Twitter. By crawling Twitter, we have collected one month of data with over 500,000 Twitter users with more than 40 million tweets. Based on the data, we have identified features that can differentiate humans, bots, and cyborgs on Twitter. Using entropy measures, we have determined that humans have complex timing behavior, i.e., high entropy, whereas bots and cyborgs are often given away by their regular or periodic timing, i.e., low entropy. In examining the text of tweets, we have observed that a high proportion of bot tweets contain spam content. Lastly, we have discovered that certain account properties, like external URL ratio and tweeting device makeup, are very helpful on detecting automation. Based on our measurements and characterization, we have designed an automated classification system that consists of four main parts: the entropy component, the spam detection component, the account properties component, and the decision maker. The entropy component checks for periodic or regular tweet timing patterns; the spam detection component checks for spam content; and the account properties component checks for abnormal values of Twitter-account-related properties. The decision maker summarizes the identified features and decides whether the user is a human, bot, or cyborg. The Feature Weights effectiveness of the classification system is evaluated through the test data set. Moreover, we have applied the system to classify the entire data set of over 500,000 users collected, and speculated the current composition of Twitter user population based on the classification results.

REFERENCES

[1] “Top Trending Twitter Topics for 2011 from What the Trend,” http://blog.hootsuite.com/top-twitter-trends 2011/, Dec. 2011.

[2] "Twitter Blog: Your World, More Connected," http://blog.twitter.com/2011/08/your-world more-connected.html, Aug. 2011.

[3] Alexa, "The Top 500 Sites on the Web by Alexa," http://www.alexa.com/topsites, Dec. 2011.

[4] "Amazon Comes to Twitter," http://www.readwriteweb.com/archives/amazon_comes_to_twitter.php, Dec. 2009.

[5] "Best Buy Goes All Twitter Crazy with @Twelpforce," http://twitter.com/in_social_media/status/2756927865, Dec. 2009.

[6] "Barack Obama Uses Twitter in 2008 Presidential Campaign,"http://twitter.com/BarackOba/, Dec. 2009.

[7] J. Sutton, L. Palen, and I. Shlovski, "Back-Channels on the Front Lines: Emerging Use of Social Media in the 2007 Southern California Wildfires," Proc. Int'l ISCRAM Conf., May 2008.

[8] A.L. Hughes and L. Palen, "Twitter Adoption and Use in Mass Convergence and Emergency Events," Proc. Sixth Int'l ISCRAM Conf., May 2009.

[9] S. Gianvecchio, M. Xie, Z. Wu, and H. Wang, "Measurement andClassification of Humans and Bots in Internet Chat," Proc. 17thUSENIX Security Symp., 2008.

[10] B. Stone-Gross, M. Cova, L. Cavallaro, B. Gilbert, M. Szydlowski, R. Kemmerer, C. Kruegel, and G. Vigna, "Your Botnet Is My Botnet: Analysis of a Botnet Takeover," Proc. 16th ACM Conf. Computer and Comm. Security, 2009.

[11] S. Gianvecchio, Z. Wu, M. Xie, and H. Wang, "Battle of Botcraft: Fighting Bots in Online Games with Human Observational Proofs," Proc. 16th ACM Conf. Computer and Comm. Security, 2009.

[12] A. Java, X. Song, T. Finin, and B. Tseng, "Why We Twitter: Understanding Microblogging Usage and Communities," Proc. Ninth WebKDD and First SNA-KDD Workshop Web Mining and Social Network Analysis, 2007.

[13] B. Krishnamurthy, P. Gill, and M. Arlitt, "A Few Chirps about Twitter," Proc. First Workshop Online Social Networks, 2008.

[14] S. Yardi, D. Romero, G. Schoenebeck, and D. Boyd, "Detecting Spam in a Twitter Network," First Monday, vol. 15, no. 1, Jan. 2010.

[15] A. Mislove, M. Marcon, K.P. Gummadi, P. Druschel, and B.Bhattacharjee, "Measurement and Analysis of Online Social Networks," Proc. Seventh ACM SIGCOMM Conf. Internet Measurement,2007.

[16] S. Wu, J.M. Hofman, W.A. Mason, and D.J. Watts, "Who Says What to Whom on Twitter," Proc. 20th Int'l Conf. World Wide Web, pp. 705-714, 2011.

[17] H. Kwak, C. Lee, H. Park, and S. Moon, "What Is Twitter, a Social Network or a News Media?" Proc. 19th Int'l Conf. World Wide Web, pp. 591-600, 2010.

[18] I.-C.M. Dongwoo Kim, Y. Jo, and A. Oh, "Analysis of Twitter Lists as a Potential Source for Discovering Latent Characteristics of Users," Proc. CHI Workshop Microblogging: What and How Can We Learn From It?, 2010.

[19] D. Zhao and M.B. Rosson, "How and Why People Twitter: The Role that Micro-Blogging Plays in Informal Communication at Work," Proc. ACM Int'l Conf. Supporting Group Work, 2009.

[20] K. Starbird, L. Palen, A. Hughes, and S. Vieweg, "Chatter on the Red: What Hazards Threat Reveals about the Social Life of Microblogged Information," Proc. ACM Conf. Computer Supported Cooperative Work, Feb. 2010.
[21] B.J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter
Power: Tweets as Electronic Word of Mouth," Am. Soc. For Information Science and Technology, vol. 60, no. 11, pp. 2169-2188,2009.
[22] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@spam: The Underground on 140 Characters or Less," Proc. 17th ACM Conf. Computer and Comm. Security, pp. 27-37, 2010.
[23] K. Thomas, C. Grier, D. Song, and V. Paxson, "Suspended Accounts in Retrospect: An Analysis of Twitter Spam," Proc. ACM SIGCOMM Conf. Internet Measurement Conf., pp. 243-258, 2011.
[24] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System," Proc. Seventh ACM SIGCOMM Conf. Internet Measurement, 2007.
[25] M. Cha, A. Mislove, and K.P. Gummadi, "A Measurement-Driven Analysis of Information Propagation in the Flickr Social Network," Proc. 18th Int'l Conf. World Wide Web, 2009.
[26] M. Xie, Z. Wu, and H. Wang, "Honeyim: Fast Detection and\ Suppression of Instant Messaging Malware in Enterprise-Like Networks,," Proc. 23rd Ann.Com

ABOUT AUTHORS

1. **G.DEEPAK is** currently pursuing her M.Tech Computer Science & Engineering Department in Kakatiya Institute of Technology and Science, Warangal.He received her B.Tech in Computer Science and Engineering from Kamala institutes of technological sciences, Singapur. His area of interests includes Data mining .
2. **V.SHANKAR** is currently working as a Assoc. Prof. Department of Computer Science & Engineering, Kakatiya Institute of Technology and Science, Warangal. His research interests include software engineering.