A NOVEL FOR ANALYZING AND RECOGNITION OF TWITTER ACCOUNTS

R.NAVEEN KUMAR
Asst.Prof, Dept of CSE,
Mahaveer Institute of Science and Technology

N.ASHWAN KUMAR
Asst.Prof, Dept of CSE,
Mahaveer Institute of Science and Technology

ABSTRACT

Twitter is a new web-appliance in performance twofold roles of on-line social-networking and micro blogging. We have premeditated the difficulty of computerization by bots and cy-borgs on Twitter. As a accepted web-application, Twitter has turn out to be a sole proposal for information distribution with a huge user-base. However, its reputation and very unlock nature have made twitter a very attractive target for misuse by robotic programs, i.e., bots. The problem of bots on Twitter is additional complex by the key role that computerization plays in everyday Twitter usage. Based on the data, we have recognized features that can distinguish humans, bots, and cyborgs on Twitter. valid bots generate a large amount of compassionate tweets deliver news and updating feeds, while malicious bots extend spam or malicious contents. More fascinatingly, in the middle between human and bot, there has emerged cyborg referred to either bot -assisted human or human-assisted bot. To carry users in identifying who they are interacting with, this paper focus on the categorization of human, bot and cy-borg accounts on Twitter.

INTRODUCTION

TWITTER is a accepted online social net-working and micro-blogging tool, which was released in 2006. Remarkable difficulty is its distinguishing feature. Its district interacts via publishing text-based posts, recognized as tweets. The tweet size is limited to 140 characters. Hash tag, namely words or phrases prefixed with a # symbol, can collection tweets by issue. For example, #Justin Bieber and #Women's World Cup are the two trending hash-tags on Twitter in 2011. Pictogram @ followed by a username in a tweet enables the unswerving deliverance of the tweet to that user. contrasting most online social net-working sites (i.e., Facebook and MySpace), Twitter's user alliance is intended at and consists of two ends, associate and admirer. In the case where the user A adds B as a friend, A is a admirer of B while B is a friend of A. In Twitter terms, A follows B (namely, the following relationship is unidirectional from A to B). B can also add A as his friend (namely, following back or recurring the follow), but is not essential. When A and B follow each other, the association becomes bi-directional. From the perspective of in sequence flow, tweets flow from the source (author) to subscribers (followers).

The characterization of spam in this paper is dispersion malicious, phishing, or spontaneous commercial content in tweets. These bots erratically add users as their friends, expecting a few users to follow back. In this way, spam tweets posted by bots exhibit on users' home-pages. Enticed by the pleasing text contented, some users may click on links and get redirected to spam or malicious sites. If human users are enclosed by malevolent bots and spam tweets, their chirping familiarity deteriorates, and ultimately the whole Twitter population will be hurt. In the paper, we first conduct a series of dimensions to distinguish the difference between human, bot, and cyborg in terms of tweeting behavior, tweet content, and account properties. By crawling Twitter, we accumulate over 500,000 users and more than 40 million tweets posted by them. Then, we achieve a comprehensive data analysis, and locate asset of helpful features to categorize users into the three classes.

A FEW CHIRPS ABOUT TWITTER

Online social networks (OSNs) have emerged recently as the most popular

application since the Web began in the early 1990s.Coincident with the growth of Web 2.0 applications (such as mashups, user generated content) and users being treated as_rst class objects, frequent social networks along with thousands of helper applications have arisen. Well known ones include Facebook, MySpace, Friendster, Bebo, hi5, and Xanga, each with over forty million registered users. Many applications have been created to use the distribution platform provided by OSNs. For example, popular games like Scrabulous, allow many thousands of users on Facebook to play the game with their social - network friends. A few slighter networks with supercial similarities to the larger OSNs have started recently. Some of these began as simple helper applications that work well with the larger OSNs, but then become popular in their own right. A key distinguishing factor of these smaller networks is that they afford a new means of communication. In the case of Twitter it is Short Message Service (SMS), a store and promote best effort delivery system for text messages. In the case of qik, it is streaming video from cell phones, another small OSN, allows people to share their .activity stream", while Dodge ball lets users update their status along with _ne-grained geographical information, allowing the system to locate friends nearby. GyPSii, a Dutch OSN is aimed at the mobile market exclusively, combining geo-location of users with image uploading and works on various cell phones including Apple's iPhone. Close to Twitter, a mobile OSN that encourages constant updates is Bliin . For example, Twitter messages can be received by users as a text message on their cell phone, through a Facebook application that users have added to their Facebook account to see the messages when they log in, via email, as an RSS feed, or as an Instant Message (with a choice of Jabber, Google Talk etc.). Figure 1 shows the diverse input and output vectors to send and receive Twitter status update messages (.tweets"). Twitter isan example of a micro-content OSN, as opposed to say, YouTube, where individual videos uploaded are much larger. Individual tweets are limited to 140 characters in Twitter. Twitter began in October 2006 and is written using Ruby on Rails . Our study _ends that users from a dozen countries are heavily represented in the user population but significantly less than the U.S. Recently, Twitter has made interesting inroads into novel domains, such as help during a large-scale _re emergency , updates during riots in Kenya , and live traffic updates to track commuting delays .

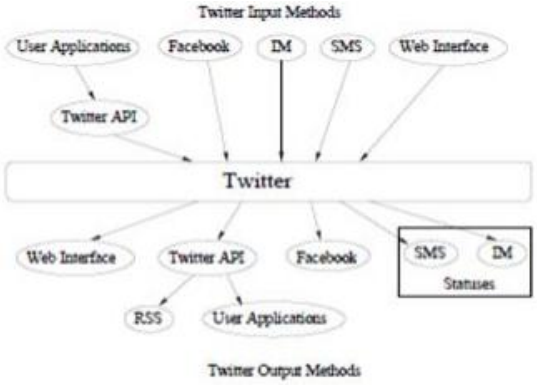

Figure 1: Twitter input and output methods

Figure 2: Example of the Twitter public timeline

Research Background : Data Collection Using the Twitter search API6 we collected publicly available tweets during the four events of study. As a security feature Twitter users can choose to make their profile either public or private. All tweets sent by a public profile are publicly available for anyone to view, even those without an account. These public tweets are also aggregated into a tweet stream called the public timeline (see Figure 2 for an example), which lets anyone view what people are tweeting about at a given time. If a user marks their profile as private, their tweets can only be viewed by other users that they have given permission to follow them, so these tweets are not ones we could sample. Figure 2: Example of the Twitter public timeline Data collection timeframes (see Table 1) for each event were determined by the nature of the event. Both the DNC and the RNC started on a Monday and ended on a Thursday. However, there were many pre-convention activities and so data capture began the Thursday before continuing until the last day of the convention, rendering eight successive days of data collection for each event. For the two hurricanes, data collection began the day each hurricane was officially named and continued until the hurricane was declared over. Table 1 also describes how many tweets were captured in each data set, including the number of unique Twitter users sending these tweets . Tweets were selected using high-level, case-insensitive search terms (see Table 1). Ideally we would have included searches based on location but, unfortunately, the location field on a user profile is an editable field that is only precise or updated if the user chooses to do so. We found that inclusion of a location search returned too many irrelevant tweets and so we did not use this information in the data collection.

Event	Data Collection Timeframe	Search Terms	# Tweets	Avg. # Tweets per Day	# Users
Conventions					
DNC	21 Aug 2008 – 28 Aug 2008	denver, dnc	21,139	2,642	9,417
RNC	8 Aug 2008 – 4 Sep 2008	rnc, st paul, saint paul	17,588	2,199	8,613
Hurricanes					
Gustav	25 Aug 2008 – 4 Sep 2008	gustav, hurricane	38,373	3,488	14,478
Ike	1 Sep 2008 – 14 Sep 2008	ike, hurricane	59,963	4,283	20,689

Table 1

Daily Twitter Activity

Twitter movement varied over the days of each event, with the graphs of this activity (see Figure 3) corresponding with the significant happenings of the events they reflect. For example, both the DNC and RNC show the number of tweets, according to our sampling method, was highest on the designated days of each convention—August 25-28, 2008 and September 1-4, 2008 correspondingly. Hurricane Gustav experienced the highest number of tweets according to our sampling method on September 1, 2008, the day it hit landfall in the US. For Hurricane like two spikes in activity appear, one when it made landfall in Cuba on September 8, and another when it model and fall in the US on September 13,

2008. Figure 3: Graphs of the number of daily tweets our research sampled using specific keywords. Similarly, the number of tweets collected for each event corresponds with the size and impact of each event (see Table 1). Tweets collected for the DNC, the larger of the two conventions studied, outnumbered those collected for the RNC by more than 20%. Comparison of the two hurricanes, shows that Hurricane Ike which had the larger impact, financially speaking— estimated $27 billion in damages (Masters,2008b)— had much higher tweet activity than Hurricane Gustav— estimated $4-14 billion in damages (Masters,2008a). Because we cannot be sure our search selection yielded completely comparable samples, we can only speculate that there is a correlation here.

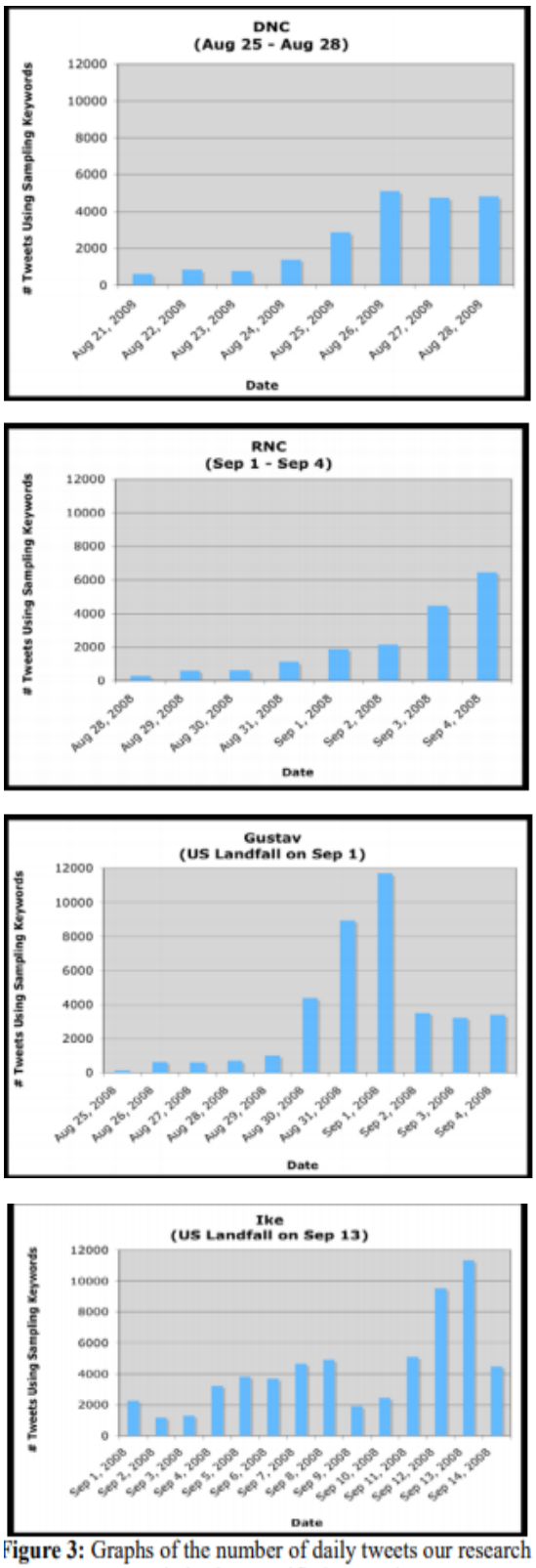

Figure 3: Graphs of the number of daily tweets our research sampled using specific keywords.

Number of Tweets per User

To understand how many tweets each user in our data contributed to the Twitter conversation around each event, we determined the tweet count for each

user. Users within each information set were then sorted according to their tweet count, after which we calculated the percentage of users who contributed one tweet for each event. We then performed the same percentage calculation for those who contribute two tweets up to seven tweets. Somewhat unpredictably, we found the percentage of users who sent a certain number of tweets to be unfailing across events. This suggest similar patterns of macro Twitter behavior: that the number of Twitter senders decreases as the number of messages sent increases. This supports—but does not prove—the idea that people serve as "information hubs" (Palen and Liu, 2007) to collect and deploy information, but that many others "participate" in the event in a more peripheral fashion.

URL Tweets :Twitter allows users to include URLs in their tweets. This is useful for multiple reasons. Sometimes the 140character limit for Twitter messages can be too constricting when a user wants to convey large amounts of information. Other times, tweets serve as pointers to resources that followers might find interesting or important. Readers of the tweet can then follow the URL to a website with a click on the link. Again, we wanted to compare how many tweets in our data sets contain URLs with the number of tweets containing URLs found in a random sample of all tweets appearing in Twitter during our collection time frame. Using the same sample of random tweets we collected in the last section we were able to make this comparison (see Table 3).We found the percentage of tweets containing URLs to be notably lower in the general sample than that of our convention and hurricane data samples (see Table 3). This observed behavior supports the idea that users are serving as information brokers, and distributing web-based information resources to others during times of non-routine events. Also distinguished is the difference in percentage of URL tweets between the two conventions and the two hurricanes. Roughly 40% of the convention tweets contained URLs, while around 50% of the hurricane tweets restricted URLs. What could explain this difference is that emergency events have higher information demands than mass convergence but non-emergency events.

Event/Data Set	Avg. # URL Tweets per Day	Avg. # of Sampled Tweets per Day	Percentage of URL Tweets
Conventions			
DNC	1,143	2,642	43.25%
RNC	805	2,199	36.59%
Hurricanes			
Gustav	1,827	3,488	52.38%
Ike	2,136	4,283	49.87%
Sample of the General Population Tweets During Same Time Period			
General	180	732	24.57%

Table 3. Percentage of Tweets in each data set that contain an URL.

EXISTING SYSTEM

Twitter has been extensively used since 2006. To enhanced comprehend micro-blogging convention and community, premeditated over 70,000 Twitter users and categorized their posts into four main groups: daily prattle, conversations, distribution in organize or URLs, and reporting news. Their work also studied 1) the growth of Twitter, showing a linear growth rate; 2) its network properties, showing the evidence that the networks scale-free like other social networks; and 3) the geological distribution of its users, showing that most Twitter users are from the US, Europe, and Japan. A group of over 100,000Twitter users and classified their roles by follower-to-following ratios into three groups: 1) broadcasters, which have a large number of followers; 2) connections, which have about the same number on either group or subsequent; and 3) miscreants and evangelists, which track a large number of other users but have few group. The in sequence propagation on Twitter, regarding the production, flow, and spending of in sequence. The quantitative study on Twitter by crawling the entire Twitter sphere. Their work analyzed the follower-following topology, and found manpower-law follower division and low

reciprocity, which all mark a divergence from known characteristics of human social networks. Twitter list as a probable foundation for discover concealed typeset and benefit of users. A twitter list consists of various users and their tweet. Their research indicated that words extracted from each list are diplomat of all the members in the list even if the words are not used by the members. It is useful for targeting users with specific interests. The behaviors of spammers on Twitter by analyzing the tweets originate from suspended users in retrospect. They found that the current market for Twitter spam uses a sundry set of spamming techniques, including a variety of strategies for creating Twitter accounts, generating spam URLs, and distributing spam.

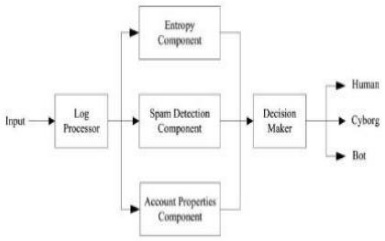

Fig No 01 Existing Approach

PROPOSED SYSTEM

The proposed system using twitters data study, and finds a set of useful features to classify users into the three classes. Based on the quantity results, we propose an mechanical organization system that consists of four major components:

1. The entropy component uses tweeting time as a measure of behavior complication, and detects the periodic and regular timing that is an indicator of mechanization. 2. The spam discovery component uses tweet content to check whether text patterns contain spam or non-spam. 3. The account property module employs useful account property, such as tweeting device structure, URL ration, to detect deviations from normal. 4. The decision maker is based on Naïve Bayes, and it uses the grouping of the features generated by the above three components to classify an unknown user as human, bot, or cyborg. To together data, we arbitrarily choose different samples and classify them by manually checking their user logs and homepages. The ground-truth set includes 2,000 users per class of human, bot, and cyborg, and thus in total three are 6,000 confidential samples. The system classify Twitter users into three categories: human, bot, and cyborg. The system consists of quite a few mechanisms: the entropy module, the spam detection constituent, the account properties component, and the decision maker. The elevated design of our Twitter user categorization system is shown in fig.2. 1. Entropy Component: The entropy component detects episodic or regular timing of the messages posted by a Twitter user. On one hand, if the entropy or correct conditional entropy is low for theater tweet delays, it indicates periodic or regular behavior, a sign of mechanization. More expressly, some of the messages are posted via automation, i.e., the user may be probable bot or cyborg. On the other hand, a high-entropy indicates irregularity, a sign of human participation.

2. **Spam Detection Component**: The spam recognition component examines the content of tweets to detect spam. We have observed that most spam tweets are generated by bots and only very few of them are manually posted by humans. Thus, the presence of spam patterns usually indicates automation. Since tweets are text, determining if their content is spam can be reduced to a text classification problem.

3. Account Properties Component: Twitter account-related properties are very supportive for the user classification. The first property is the URL ratio. Thus, high ratio (e.g., close to 1) suggests a bot and a low ratio implies a human. The second property is tweeting device makeup. The third property is the followers to friend's ratio. The fourth property is link protection, i.e., to decide

whether external links in tweets are malicious/phishing URLs or not. We run a batch script to check a URL in five blacklists: Google Safe Browsing, Phishing Tank, URIBL, SURBL, and Spamhaus. If the URL appears in any of the blacklists, the characteristic of link safety is set as false. The fifth property is whether a Twitter account is verified. The sixth property is the account registration date. The last two properties are the hash tag ratio and mention ratio.Hashtag ratio of an account is defined as the number of hash tags included in the tweets over the number of tweets posted by the account. Mention ratio is defined similarly.

4. Decision Maker: We select Random Forests the machine learning algorithm, and implement the decision maker based on it. Each decision tree is built top- down in a recursive manner. Forever node in the construction path, m features is randomly selected to reach a decision at the node. The node is then associated with the feature that is the most informative. Entropy is used to compute the information gain contributed by each of the features (namely, how informative a feature is). In other words, the recursive algorithm applies a greedy search by selecting the candidate characteristic that maximizes the heuristic splitting measure.

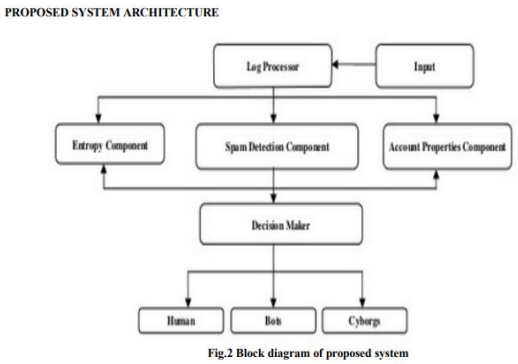

Fig.2 Block diagram of proposed system

Conclusion and Future Scope

In this paper, we have considered the difficulty of automation by bots and cyborgs on Twitter. The complexity of bots on Twitter is further complex by the key role that automation plays in everyday Twitter usage. We have collected one month of information with good number of Twitter users with more than 40 million tweets. Based on the data, we have recognized features that can differentiate humans, bots, and cyborgs on Twitter. Lastly, we have uncovered that certain account properties, like external URL ratio and tweeting device makeup, are very helpful on detecting automation. In the future, there is a prospect to block the automated tweets by using any engineering technique and also there is a scope to extend this work, which restricts discarding of huge data. Based on the data, we have recognized features that can distinguish humans, bots, and cyborgs on Twitter. Using entropy measures, we have resolute that humans have composite timing behavior, i.e., high entropy, whereas bots and cyborgs are often given away by their regular or interrupted timing, i.e., low entropy. In examining the text of tweets, we have observed that a high quantity of bot tweets enclose spam content.

REFERENCES:

[1].Top Trending Twitter Topics for 2011 from What the Trend,"http://blog.hootsuite.com/top-twitter-trends2011/, Dec. 2011.

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

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

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

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

[6]. 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

[7].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.

[8] 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.

[9] 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.