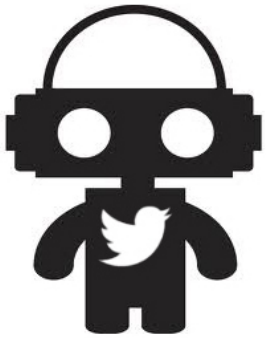
TWITTER BOT PROFILING

OVERVIEW

Submit Request



This dataset comprises a set of Twitter accounts in Singapore that are used for social bot profiling research conducted by the Living Analytics Research Centre (LARC) at Singapore Management University (SMU). Here a bot is defined as a Twitter account that generates contents and/or interacts with other users automatically (at least according to human judgment). In this research, Twitter bots have been categorized into three major types:

- Broadcast bot. This bot aims at disseminating information to general audience by providing, e.g., benign links to news, blogs or sites. Such bot is often managed by an organization or a group of people (e.g., bloggers).
- Consumption bot. The main purpose of this bot is to aggregate contents from various sources and/or provide update services (e.g., horoscope reading, weather update) for personal consumption or use.
- Spam bot. This type of bots posts malicious contents (e.g., to trick people by hijacking certain account or redirecting them to malicious sites), or promotes harmless but invalid/irrelevant contents aggressively.

This categorization is general enough to cater for new, emerging types of bot (e.g., chatbots can be viewed as a special type of broadcast bots). **Figure 1** illustrates the three bot types together with human accounts, where the arrow direction represents the flow of information.

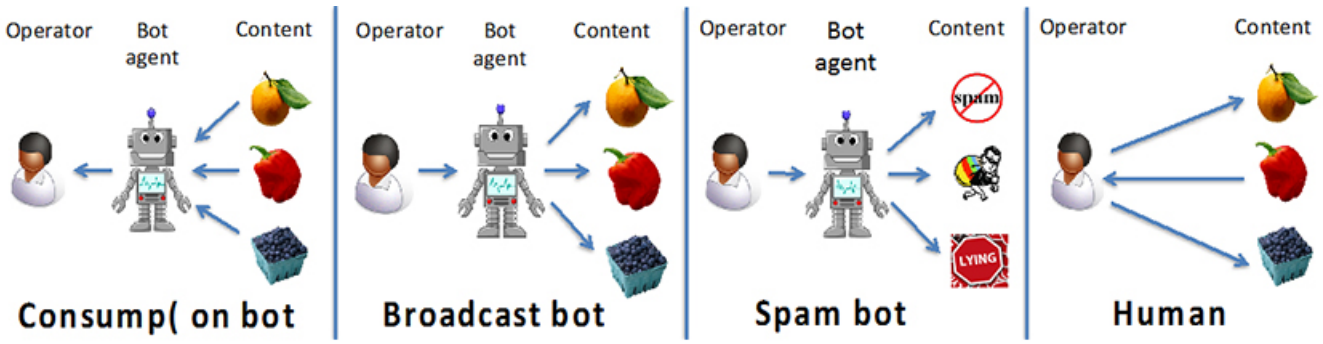


Figure 1. Categorization of Twitter bots

DESCRIPTION

The dataset was collected from 1 January to 30 April 2014 via the Twitter REST and streaming APIs. Starting from popular seed users (i.e., users having many followers), their follow, retweet, and user mention links were crawled. The data collection proceeds by adding those followers/followees, retweet sources, and mentioned users who state Singapore in their profile location. Using this procedure, a total of 159,724 accounts have been collected.

To identify bots, the first step is to check active accounts who tweeted at least 15 times within the month of April 2014. These accounts were then manually checked and labelled, of which 589 bots were found. As many more human users are expected in the Twitter population, the remaining accounts were randomly sampled and manually checked. With this, 1,024 human accounts were identified. In total, this results in 1,613 labelled accounts.

Table 1 summarizes the basic (count) statistics of the dataset. The labelling was done by four volunteers, who were carefully instructed on the definitions of the three bot types. The volunteers agree on more than 90 % of the labels, and any labelling differences in the remaining accounts are resolved by consensus. Also, if an account exhibits both human and bot characteristics, the label was determined based on the majority posting patterns.

Table 1. Statistics of the bot dataset

Labelled data				UnLabelled data
Consumption bot	Broadcast bot	Spam bot	Human account	158,111
313	171	105	1,024	

CITATION

Kindly cite the following paper if you use the dataset:

R. J. Oentaryo, A. Murdopo, P. K. Prasetyo, and E.-P. Lim, "On profiling bots in social media," in Proceedings of the International Conference on Social Informatics (SocInfo'16), Bellevue, WA, 2016, pp. 92-109.

Link: http://link.springer.com/chapter/10.1007%2F978-3-319-47880-7_6

Disclaimer: The bot labels in this dataset were obtained from the observation period of 1 January - 30 April 2014. Given the rapidly changing nature of bot behavior, however, these labels may no longer be relevant today. As such, when analyzing the labels, you are advised to use the corresponding tweet posts/contents from the observation period mentioned above.

Last updated on **04 Jan 2017** .

SMU
SINGAPORE MANAGEMENT UNIVERSITY

WHERE TO FIND US

Living Analytics Research Centre

School of Information Systems
Singapore Management University
80 Stamford Road
Singapore 178902

Tel: 65 6808 5227 | larc@smu.edu.sg