Tweets and Social Network Data for Twitter Bot Analysis

Jennifer Golbeck, Niloofarsadat Alavi, Hannah K. Bako, Saptarashmi Bandyopadhyay, Calvin Bao, Matthew Barzola, Corie Brown, Kristen M. Byers, Melissa Carraway, Keaunna Cleveland, Brandon Y. Feng, Shayna Foote, Petra Galuscakova, Auguste H. Gezalyan, Rajesh Kumar Gnanasekaran, Saeed Hadadan, SeyyedPooya HekmatiAthar, Jeffrey W Henrikson, Karan D Kaur, Nikita Khatri, Nitzan Koren, Anna Lenhart, Ruiqi Li, Yannan Liu, Jennifer MacDonald, Sean McWillie, Gloria Nzeka, Neviya Prakash, Joanna Purich, Robin Reiss, Aayushi Roy, David Rudd, Anup Sathya, Nicole R. Schneider, Hilary Szu Yin Shiue, Astha Singhal, John Sutton, Gabrielle Trilling, Amulya Velamakanni, Jacob Walter, Tory White, Gareth Thomas Williams, Zachary Wilton, Peratham Wiriyathammabhum, Mischa Wolfinger, Arthur Wu, and Rachael Zehrung

University of Maryland, College Park, MD 20742 USA jgolbeck@umd.edu

Abstract. There are a number of Twitter bot datasets available, and these primarily consist of twitter IDs and labels (bot, human). They do not include social network data and the tweets available are limited. As Twitter shuts down bot accounts, this information is lost, making it difficult for researchers to use these datasets since no features can be retrieved. In this project, we identified eight large datasets with bot/human labels and retrieved screen names, friend lists (people the account is following), and all of the tweets accessible through the Twitter API for accounts in the datasets at the time of collection. For accounts labeled as bots in the original dataset, we also collected Botometer [2] and Bot Sentinel scores where available. Human coders analyzed the bots' content and labeled them as retweet only where appropriate and indicated if they tweeted at all about US politics.
The resulting dataset is publicly available by request (to comply with Twitter terms of service) and will ensure future usefulness of these existing bot datasets.

Keywords: bots · tweets · social media

1 Introduction

Detecting bots on social media is an important challenge, and one that will take on added importance in upcoming elections and amidst the crises and social movements of 2020. Many characteristics can be used in differentiating bots from human accounts - account metadata [8], social network characteristics [4], DNS activity [7], account and linguistic features [5] and more.

There are a number of Twitter bot datasets available, but many are twitter IDs and labels (bot, human). They do not include social network data and the tweets available are limited. As Twitter shuts down bot accounts, this information is lost, making it very difficult for researchers to use these datasets since no features can be retrieved. In this project, we retrieved screen names, friend lists (people the account is following), and all of the tweets accessible through the Twitter API for any accounts the account was following as of March 2020. We expect more of these accounts will be deleted over time. The dataset we provide here will make data about these accounts reliably available for research into the future.

Twitter Terms of Service prevent us from publishing this data directly to the web, but any interested researcher can contact *jgolbeck@umd.edu* to receive a copy of the data.

2 Data Sets

The Bot Repository at https://botometer.iuni.iu.edu/bot-repository/datasets.html provides a number of labeled bot datasets for study. These typically consist of twitter IDs and labels (bot, human). We selected the following datasets, all of which were provided as twitter IDs with labels. Most of these datasets provided a file with data from tweets, but it was limited, averaging 1-2 tweets per account; not enough to use for linguistic analysis or classification. No social network information was included in the datasets. Each dataset name (number of accounts), and description as listed in the Bot Repository is as follows:

- **botometer-feedback-2019**: (529): Botometer feedback accounts manually labeled by K.C. Yang [8]
- **botwiki-2019**: (704) Self-identified bots from https://botwiki.org [8]
- **gilani-2017**: (2,652) Manually annotated human and bot accounts [3]
- **midterm-2018**: (50,538) Manually labeled human and bot accounts from 2018 US midterm elections [8]
- **pronbots-2019**: (21,964) Pronbots shared by Andy Patel (github.com/r0zetta/pronbot2) [8]
- **varol-2017**: (2,572) This dataset contains annotation of 2573 Twitter accounts. Annotation and data crawl is completed in April 2016 [6]
- **verified-2019**: (2,000) Verified human accounts. [8]
- **cresci-stock-2018**: (25,987) A dataset of (i) genuine, (ii) traditional, and (iii) social spambot Twitter accounts, annotated by CrowdFlower contributors [1]

3 Data Collection

For each Twitter ID in the above datasets, we first checked to see if the account was still active by using the Twitter API to retriever the associated screen name. Table 1 shows the number of active accounts for each dataset vs. the number of

accounts listed in the dataset. While some retain most of their users, many are only a fraction of their original size. This emphasizes the need for an archived dataset for future research.

For each currently valid screen name, we then downloaded all accessible tweets for currently active users. The maximum available was 3,200, a constraint imposed by the Twitter API. Our focus was on enabling text and post timing analysis, so our dataset includes tweet ID, date and time of posting, and the tweet text. These are provided in CSV files for each of 12,797 users for whom we were able to retrieve tweets. Tweets were unavailable when accounts were private, had zero tweets, or the API otherwise returned no data.

We also retrieved the friend list for each user. On Twitter, "friends" are the accounts a user is following. Some accounts did not follow any accounts, some were marked private so friends weren't available. We retrieved friends for 16,303 accounts.

Note that for this data collection process, the Twitter API was not entirely reliable. It is possible, and we noticed in spot checks, that it failed to retrieve data that was public and should have been accessible. All the data collected is accurate, but may be incomplete. Even with that limitation, we believe this archive is an important resource for bot researchers to use in conjunction with the labeled bot datasets that are already publicly accessible.

For accounts labeled as bots in the dataset, we collected additional metadata. This is primarily comprised of features from Botometer [2], the online service rank's an account's bot-ness using over a thousand network, user, social, temporal, content, and sentiment features.

Of 6,857 bot-labeled accounts that were still active at the time of our first data crawl, we were able to collect external metadata for 5,949. The external metadata features include the following:

- Botometer scores: from the Botometer at Indiana University https://botometer.osome.iu.edu/, we retrieved the majority tweet language and scores the following attributes, all given on a 0-5 scale:
  - Astroturf
  - Fake follower
  - Financial
  - Self-declared
  - Spammer
  - Other
  - All Features score
  - Language-independent score (an average over features that do not include language analysis)
- Bot Sentinel Scores - Using the Bot Sentinel at https://botsentinel.com/, we obtained scores where available. This is a 0-100 score, and we were able to obtain it for 1,830 of the 5,949 accounts (30.1%)
- US Politics content - researchers surveyed the tweets and coded this as Y if there were any recent tweets related to US politics and N if there were none

– Retweet only - Using the collected tweets, bots were marked Y if they only retweeted others with no original posts and N if there were any tweets authored by the account. This supports future research on bots designed only to amplify through retweet.

Table 1. Datasets with accounts provided and accounts active at the time of data collection

Dataset	**Accounts Listed**	**Accounts Still Active**
botometer-feedback-2019	529	438
botwiki-2019	704	564
gilani-2017	2,652	2,483
midterm-2018	50,538	7,741
pronbots-2019	21,964	1,871
varol-2017	2,572	2,194
verified-2019	2,000	1,890
cresci-stock-2018	25,987	13,221
Total	106,946	30,402

4 Data Description

Our dataset has 21,194,137 tweets for 12,797 accounts, averaging 1,656.2 tweets per account. There are 7,982,300 social network connections over 16,303 accounts, averaging 489.6 friends per account.

The most popular accounts followed in this data were among the most popular accounts on Twitter. Table 2 shows the 20 most followed accounts and the number of followers within this dataset.

Among the bots, Table 3 shows the average and standard deviation of scores obtained from the Botometer and Bot Sentinel. While for most Botometer categories, the average score is relatively low, the composite scores of All Features and Language-Independent are much higher, as would be expected for bots. The Bot Sentinel scores, which are presented as a percentage, are quite low.

Bots tweeted in 21 different languages, but the vast majority (91.9%, N=5,465) tweeted in English, followed by Arabic (5.3%, N=318).

5 Conclusion

We have collected data from Twitter on all the accounts accessible from 8 different labeled bot datasets. Data includes screen names mapped to the Twitter IDs; the following list for each account; the ID, timestamp, and text of all tweets we could obtain through the API; and external metadata describing the bots'

Table 2. Top 20 most followed accounts in our dataset

Account	**Followers in our dataset**
BarackObama	4,942
TheEllenShow	3,358
cnnbrk	3,281
nytimes	3,162
jimmyfallon	2,922
StephenAtHome	2,494
CNN	2,288
instagram	2,274
ConanOBrien	2,248
AP	2,222
ActuallyNPH	2,187
POTUS44	2,143
BillGates	2,131
TheOnion	2,058
YouTube	2,021
jtimberlake	2,020
Oprah	2,007
taylorswift13	1,970
HillaryClinton	1,947
Twitter	1,912

characteristics and behavior. We believe this dataset provides a valuable archive for researchers. Already, 70% of the accounts included in the original labeled datasets have been removed from Twitter. As bots continue to be removed, the value of the datasets decreases unless an archive of data from these accounts is available. Our archive can be used for training and analysis as bot detection research progresses.

References

1. Stefano Cresci, Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi, and Maurizio Tesconi. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*. 963–972.
2. Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*. 273–274.
3. Jana Diesner, Elena Ferrari, and Guandong Xu. 2017. *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*.
4. Jennifer Golbeck. 2019. Benford's Law can detect malicious social bots. *First Monday* 24, 8 (2019).

Table 3. Botometer scores (0-5 scale) and Bot Sentinel Scores (0-100 scale)

Attribute	**Average**	**SD**
Astroturf	0.28	0.37
Fake follower	1.71	0.83
Financial	2.86	0.95
Self declared	0.16	0.43
Spam	0.77	0.59
Other	2.08	1.15
All Features	3.32	0.90
Language-Independent	2.78	0.94
Bot Sentinel (0-100 scale)	8.73	19.72

5. Onur Varol, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2018. Feature engineering for social bot detection. In *Feature engineering for machine learning and data analytics*. CRC Press, 311–334.
6. Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Eleventh international AAAI conference on web and social media*.
7. Ricardo Villamarín-Salomón and José Carlos Brustoloni. 2009. Bayesian bot detection based on DNS traffic similarity. In *Proceedings of the 2009 ACM symposium on Applied Computing*. 2035–2041.
8. Kai-Cheng Yang, Onur Varol, Pik-Mai Hui, and Filippo Menczer. 2020. Scalable and generalizable social bot detection through data selection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 1096–1103.