TwiBot-22: Towards Graph-Based Twitter Bot Detection

Leading authors: Shangbin Feng*, Zhaoxuan Tan*, Herun Wan*, Ningnan Wang*, Zilong Chen*, Binchi Zhang*
Q Zheng, W Zhang, Z Lei, S Yang, X Feng, Q Zhang, H Wang, Y Liu, Y Bai, H Wang, Z Cai, Y Wang, L Zheng, Z Ma, J Li, M Luo

Twitter Bot Detection

Twitter bot detection aims to automatically detect automated accounts that are operated to achieve malicious purposes, such as spreading misinformation, promoting hate speech, and manipulating public opinion.

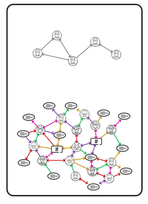
Existing models could be categorized as:

- **Feature-based**: where feature engineering is conducted with user metadata and tweets while combined with simple classifiers
- **Text-based**: where word embeddings and language models are adopted to identify bots based on user tweets and descriptions
- **Graph-based**: where network and graph mining models are adopted to analyze the structure of Twitter to identify bots

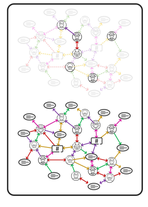
Graph-based Twitter bot detection approaches are the most advanced, achieves state-of-the-art performance, and helps to tackle the many challenges in bot detection such as bot evolution and generalization.

However, graph-based models are poorly supported by existing datasets!

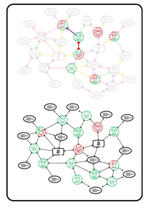
- Only **2** out of the **18** datasets provide the Twitter network structure.
- Existing datasets suffer from **limited dataset scale**, **incomplete graph structure**, and **low annotation quality**.



Limited Dataset Scale
Existing datasets contain at most 10k users, while online conversations and discussions about heated topics often involve hundreds of thousands of users.



Incomplete Graph Structure
Only users and follow relationships are provided, while Twitter is a heterogeneous information network with many types of entities and relations.



Low Annotation Quality
Crowdsourcing is often adopted for data annotation, while whether crowd workers could be trusted to identify advanced and evasive bots remain debated.

The TwiBot-22 Benchmark

To address these issues, we propose TwiBot-22, which
- establish the largest benchmark to date with ~1M users
- provides diversified entities and relations on the Twitter network
- has considerably improved annotation quality through weak supervision

Table 1: Statistics of the 9 datasets. TwiBot-20 contains unlabelled users so that # User ≠ # Human + # Bot. C-15, G-17, C-17, M-18, C-S-18, C-R-19, B-F-19 are short for cresci-2015, gilani-2017, cresci-2017, midterm-18, cresci-stock-2018, cresci-rtbust-2019, botometer-feedback-2019. C-17 contains only "post" edges between users and tweets, which is not a graph-based dataset.

Dataset	C-15	G-17	C-17	M-18	C-S-18	C-R-19	B-F-19	TwiBot-20	TwiBot-22
# Human	1,950	1,394	3,474	8,092	6,174	340	380	5,237	860,057
# Bot	3,351	1,090	10,894	42,446	7,102	353	138	6,589	139,943
# User	5,301	2,484	14,368	50,538	13,276	693	518	229,580	1,000,000
# Tweet	2,827,757	0	6,637,615	0	0	0	0	33,488,192	88,217,457
# Human Tweet	2,631,730	0	2,839,361	0	0	0	0	33,488,192	81,250,102
# Bot Tweet	196,027	0	3,798,254	0	0	0	0	33,488,192	6,967,355
# Edge	7,086,134	0	6,637,615	0	0	0	0	33,716,171	170,185,937

Benchmarking Twitter Bot Detection

Armed with TwiBot-22, we provide a comprehensive benchmarking of Twitter bot detection to enable a rethinking of research progress.
- **35** bot detection **models**, covering both the classic and the advanced
- **9** representative and publicly available bot detection **datasets**

Table 2: Average bot detection accuracy and standard deviation of five runs of 35 baseline methods on 9 datasets. **Bold** and underline indicate the highest and second highest performance. The F, T, and G in the "Type" column indicates whether a baseline is feature-based, text-based, or graph-based. Cresci *et al.* and Botometer are deterministic methods or APIs without standard deviation. "/" indicates that the dataset does not contain enough user information to support the baseline. "-" indicates that the baseline is not scalable to the largest TwiBot-22 dataset.

Method	Type	C-15	G-17	C-17	M-18	C-S-18	C-R-19	B-F-19	TwiBot-20	TwiBot-22
SGBot	F	77.1 (±0.2)	**78.6** (±0.8)	92.1 (±0.3)	99.2 (±0.0)	81.3 (±0.1)	80.9 (±1.5)	75.5 (±1.9)	81.6 (±0.5)	75.1 (±0.1)
Kudugunta *et al.*	F	75.3 (±0.1)	70.0 (±1.1)	88.3 (±0.2)	91.0 (±0.5)	77.5 (±0.1)	62.9 (±0.8)	74.0 (±4.7)	59.6 (±0.7)	65.9 (±0.0)
Hayawi *et al.*	F	84.3 (±0.0)	52.7 (±0.0)	90.8 (±0.0)	84.6 (±0.0)	50.0 (±0.0)	51.2 (±0.0)	77.0 (±0.0)	73.1 (±0.0)	76.5 (±0.0)
BotHunter	F	96.5 (±1.2)	76.4 (±1.0)	88.1 (±0.2)	**99.3** (±0.0)	81.2 (±0.2)	81.5 (±1.7)	74.7 (±1.0)	75.2 (±0.4)	72.8 (±0.0)
NameBot	F	77.0 (±0.0)	60.8 (±0.0)	76.8 (±0.0)	85.1 (±0.0)	55.8 (±0.0)	63.2 (±0.0)	71.7 (±0.0)	59.1 (±0.1)	70.6 (±0.0)
Abreu *et al.*	F	75.7 (±0.1)	74.3 (±0.1)	92.7 (±0.1)	96.5 (±0.1)	75.4 (±0.1)	80.9 (±0.1)	**77.4** (±0.1)	73.4 (±0.1)	70.7 (±0.1)
Cresci *et al.*	T	37.0	/	33.5	/	/	/	/	47.8	-
Wei *et al.*	T	96.1 (±1.4)	/	89.3 (±0.7)	/	/	/	/	71.3 (±1.6)	70.2 (±1.2)
BGSRD	T	87.8 (±0.6)	48.5 (±8.4)	75.9 (±0.0)	82.9 (±1.5)	50.7 (±1.3)	50.0 (±4.9)	59.6 (±3.1)	66.4 (±1.0)	71.9 (±1.8)
RoBERTa	T	97.0 (±0.1)	/	97.2 (±0.0)	/	/	/	/	75.5 (±0.1)	72.1 (±0.1)
T5	T	92.3 (±0.1)	/	96.4 (±0.0)	/	/	/	/	73.5 (±0.1)	72.1 (±0.1)
Efthimion *et al.*	FT	92.5 (±0.0)	55.5 (±0.0)	88.0 (±0.0)	93.4 (±0.0)	70.8 (±0.0)	67.6 (±0.0)	69.8 (±0.0)	62.8 (±0.0)	74.1 (±0.0)
Kantepe *et al.*	FT	97.5 (±1.3)	/	98.2 (±1.5)	/	/	/	/	80.3 (±4.3)	76.4 (±2.4)
Miller *et al.*	FT	75.5 (±0.0)	51.0 (±0.0)	77.1 (±0.2)	83.7 (±0.0)	52.5 (±0.0)	54.4 (±0.0)	**77.4** (±0.0)	64.5 (±0.4)	30.4 (±0.1)
Varol *et al.*	FT	93.2 (±0.5)	/	/	/	/	/	/	78.7 (±0.6)	73.9 (±0.0)
Kouvela *et al.*	FT	97.8 (±0.5)	74.7 (±0.9)	98.4 (±0.1)	97.0 (±0.1)	79.3 (±0.3)	79.7 (±1.2)	71.3 (±0.9)	84.0 (±0.4)	76.4 (±0.0)
Santos *et al.*	FT	70.8 (±0.0)	51.4 (±0.0)	73.8 (±0.0)	86.6 (±0.0)	62.5 (±0.0)	73.5 (±0.0)	71.7 (±0.0)	58.7 (±0.0)	-
Lee *et al.*	FT	98.2 (±0.1)	74.8 (±1.2)	**98.8** (±0.1)	96.4 (±0.1)	81.5 (±0.4)	**83.5** (±1.9)	75.5 (±1.3)	77.4 (±0.5)	76.3 (±0.1)
LOBO	FT	**98.4** (±0.3)	/	96.6 (±0.3)	/	/	/	/	77.4 (±0.2)	75.7 (±0.1)
Moghaddam *et al.*	FG	73.6 (±0.2)	/	/	/	/	/	/	74.0 (±0.8)	73.8 (±0.0)
Alhosseini *et al.*	FG	89.6 (±0.6)	/	/	/	/	/	/	59.9 (±0.6)	47.7 (±8.7)
Knauth *et al.*	FTG	85.9 (±0.0)	49.6 (±0.0)	90.2 (±0.0)	83.9 (±0.0)	**88.7** (±0.0)	50.0 (±0.0)	76.0 (±0.0)	81.9 (±0.0)	71.3 (±0.0)
FriendBot	FTG	96.9 (±1.1)	/	78.0 (±1.0)	/	/	/	/	75.9 (±0.5)	-
SATAR	FTG	93.4 (±0.5)	/	/	/	/	/	/	84.0 (±0.8)	-
Botometer	FTG	57.9	71.6	94.2	89.5	72.6	69.2	50.0	53.1	49.9
Rodriguez-Ruiz *et al.*	FTG	82.4 (±0.0)	/	76.4 (±0.0)	/	/	/	/	66.0 (±0.1)	49.4 (±0.0)
GraphHist	FTG	77.4 (±0.2)	/	/	/	/	/	/	51.3 (±0.3)	-
EvolveBot	FTG	92.2 (±1.7)	/	/	/	/	/	/	65.8 (±0.6)	71.1 (±0.1)
Dehghan *et al.*	FTG	62.1 (±0.0)	/	/	/	/	/	/	86.7 (±0.1)	-
GCN	FTG	96.4 (±0.0)	/	/	/	/	/	/	77.5 (±0.0)	78.4 (±0.0)
GAT	FTG	96.9 (±0.0)	/	/	/	/	/	/	83.3 (±0.0)	79.5 (±0.0)
HGT	FTG	96.0 (±0.3)	/	/	/	/	/	/	**86.9** (±0.2)	74.9 (±0.1)
SimpleHGN	FTG	96.7 (±0.5)	/	/	/	/	/	/	86.7 (±0.2)	76.7 (±0.3)
BotRGCN	FTG	96.5 (±0.7)	/	/	/	/	/	/	85.8 (±0.7)	**79.7** (±0.1)
RGT	FTG	97.2 (±0.3)	/	/	/	/	/	/	86.6 (±0.4)	76.5 (±0.4)

Results and Discussion

- Graph-based approaches consistently outperform other models. In fact, all the top-5s on TwiBot-20 and TwiBot-22 are graph-based.

Implication: future research on the network structure of Twitter

- TwiBot-22 establishes the largest benchmark while exposing the scalability issue of certain approaches.

Implication: emphasize scalability in bot detection research

- Model performance on TwiBot-22 is on average 2.7% lower on TwiBot-22 than TwiBot-20, which features older generation of bots.

Implication: combating the everlasting bot evolution

Generalization Test

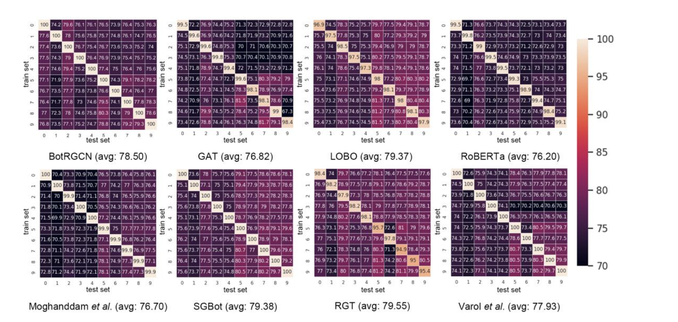


Figure 3: Training models on fold i and testing on fold j. We present model accuracy and report the average value of each heatmap (avg), which serves as an overall indicator of generalization ability.

- Graph-based methods are better at generalizing to new communities.
- Good performance does not guarantee good generalization.

The TwiBot-22 Evaluation Framework

We consolidate all implemented codes and datasets into the TwiBot-22 evaluation framework, which provides a one-stop shop for future research in Twitter bot detection.

Paper

Website

Code