TwiBot-20: A Comprehensive Twitter Bot Detection Benchmark

Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, Minnan Luo

Introduction

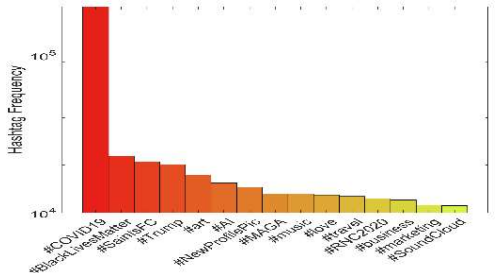
- Existing data sets:
  - low user diversity
  - limited user information
  - data scarcity

Benchmarking Performance

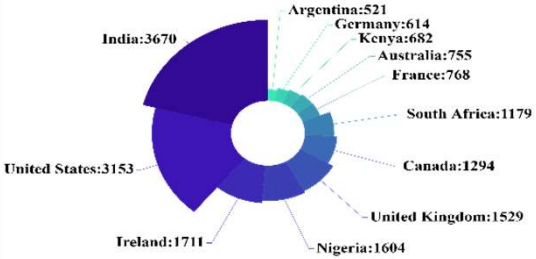
		Lee *et al.* [16]	Yang *et al.* [32]	Kudugunta *et al.* [14]	Wei *et al.* [29]	Miller *et al.* [21]	Cresci *et al.* [5]	Botometer [9]	Alhosseini *et al.* [1]
TwiBot-20	Acc	0.7456	**0.8191**	0.8174	0.7126	0.4801	0.4793	0.5584	0.6813
	F1	0.7823	**0.8546**	0.7517	0.7533	0.6266	0.1072	0.4892	0.7318
	MCC	0.4879	0.6643	**0.6710**	0.4193	-0.1372	0.0839	0.1558	0.3543
Cresci-17	Acc	0.9750	**0.9847**	0.9799	0.9670	0.5204	0.4029	0.9597	/
	F1	0.9826	**0.9893**	0.9641	0.9768	0.4737	0.2923	0.9731	/
	MCC	0.9387	**0.9625**	0.9501	0.9200	0.1573	0.2255	0.8926	/
PAN-19[3]	Acc	/	/	/	**0.9464**	/	0.8797	/	/
	F1	/	/	/	**0.9448**	/	0.8701	/	/
	MCC	/	/	/	**0.8948**	/	0.7685	/	/

- All bot detection baselines achieves significantly lower performance on TwiBot-20.
- TwiBot-20 provide neighborhood information and the two other dataset fall short.
- Bot detectors need to leverage more user information in order to perform well.
- The real-world Twittersphere has shifted from 2013.

User Diversity

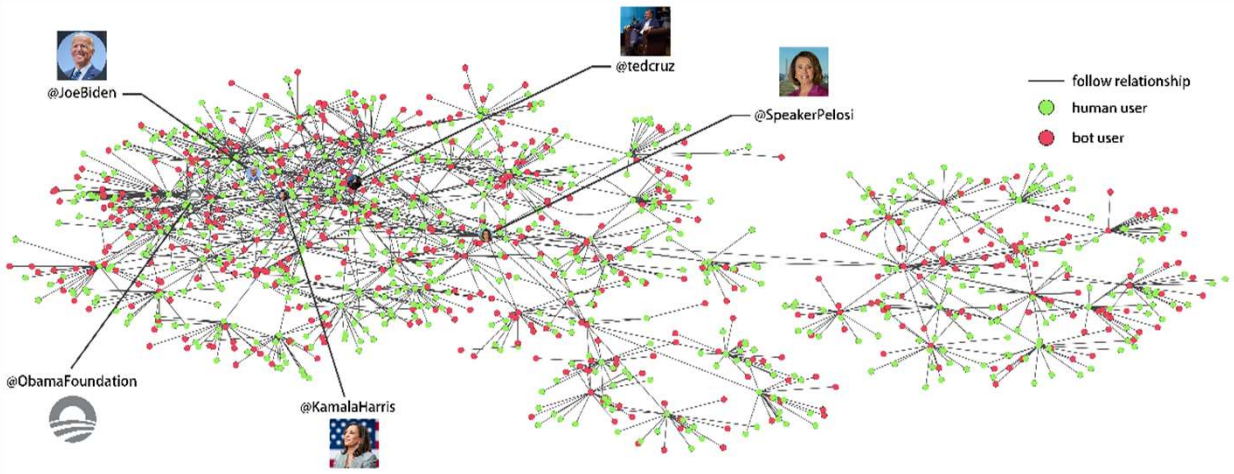


- User interest diversity



- Geographic diversity

User Information Completeness



- A user cluster in TwiBot-20 with @SpeakerPelosi as the seed user is illustrated.
- The follow relationship in TwiBot-20 provides neighborhood information and forms a dense graph structure to enable community-based bot detection measures.

Methodology

Algorithm 1: TwiBot-20 User Selection Strategy

Input: initial seed user u_0 in a user cluster
Output: user information set F

$u_0.layer \leftarrow 0$; // designate seed user as layer 0
$S \leftarrow \{u_0\}$; // set of users to expand
$u_0.expanded \leftarrow False$;
$F \leftarrow \emptyset$;
while $S \neq \emptyset$ **do**
  $u \leftarrow S.pop()$; // expand with user u
  $T(u) \leftarrow get\_tweet(u)$;
  $P(u) \leftarrow get\_property(u)$;
  if $u.layer \geq 3$ **or** $u.expanded == True$ **then**
    $F \leftarrow F \cup u(T, P, N = \emptyset)$;
    continue; // three layers max
  $u.expanded \leftarrow True$;
  $N^f(u) \leftarrow get\_friend(u)$;
  $N^t(u) \leftarrow get\_follower(u)$;
  $N(u) \leftarrow \{N^f(u), N^t(u)\}$;
  $F \leftarrow F \cup u(T, P, N)$;
  $S \leftarrow S \cup N^f(u) \cup N^t(u)$;
  for $y \in N^f(u) \cup N^t(u)$ **do**
    $y.expanded \leftarrow False$;
    $y.layer \leftarrow u.layer + 1$;
Return F; // obtained one cluster of user information

Data Scarcity

Dataset	#User	#Property	#Tweet	#Follow
varol-icwsm [27]	2,573	0	0	0
pronbots [31]	21,965	750,991	0	0
celebrity [31]	5,971	879,954	0	0
gilani-17 [13]	2,653	104,515	0	0
cresci-rtbust [19]	693	28,968	0	0
cresci-stock [7]	13,276	551,603	0	0
midterm-18 [32]	50,538	909,684	0	0
botwiki [32]	698	29,082	0	0
verified [32]	1,987	83,383	0	0
PAN-19[3]	11,568	0	369,246	0
caverlee [15]	22,224	155,568	5,613,166	0
cresci-17 [6]	14,398	547,124	18,179,186	0
TwiBot-20	**229,573**	**8,723,736**	**33,488,192**	**455,958**

- three aspects of user information
- establish the largest benchmark

Xi'an Jiaotong University

CIKM 2021 1-5 NOVEMBER