SATAR: A Self-supervised Approach to Twitter Account Representation Learning and its Application in Bot Detection

Shangbin Feng
LUD Lab, Xi'an Jiaotong University
wind_binteng@stu.xjtu.edu.cn

September 26, 2021

Table of Contents

Twitter Bot

Definition

A **Twitter bot** is a type of automated programs, which are often operated to achieve malicious goals.

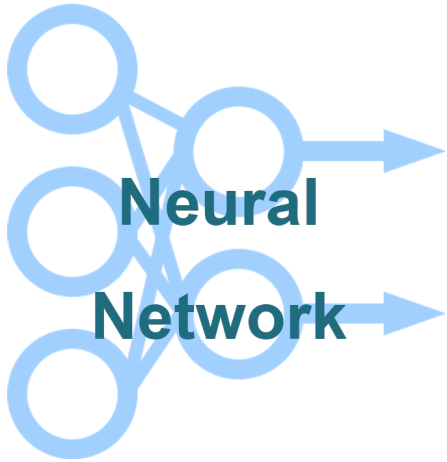
- Involve in election interference
  - Twitter Bots involve in the elections in the United States and Europe. WWW'19.
  - Are 'bots' manipulating the 2020 conversation? Here's what's changed since 2016.The Washington Post.
- Spread misinformation and propagate extreme ideology
  - Researchers: Nearly Half Of Accounts Tweeting About Coronavirus Are Likely Bots. NPR.
  - Berger et al., The Brookings project on US relations with the Islamic world.

What is bot detection

Definition

Problem: Twitter Bot Detection Given a Twitter user U and its information T, P and N, learn a bot detection function $f : f(U(T, P, N)) \rightarrow \hat{y}$, such that $\hat{y}$ approximates ground truth y to maximize prediction accuracy.

Related Work



- User profile, ICMSM 2011
- Social networks, ISONAM 2017
- Timeline of accounts, IEEE Intelligent Systems, 2016
- Redirection of URLs, TDSC 2013
- Classification of websites, S&P 2011

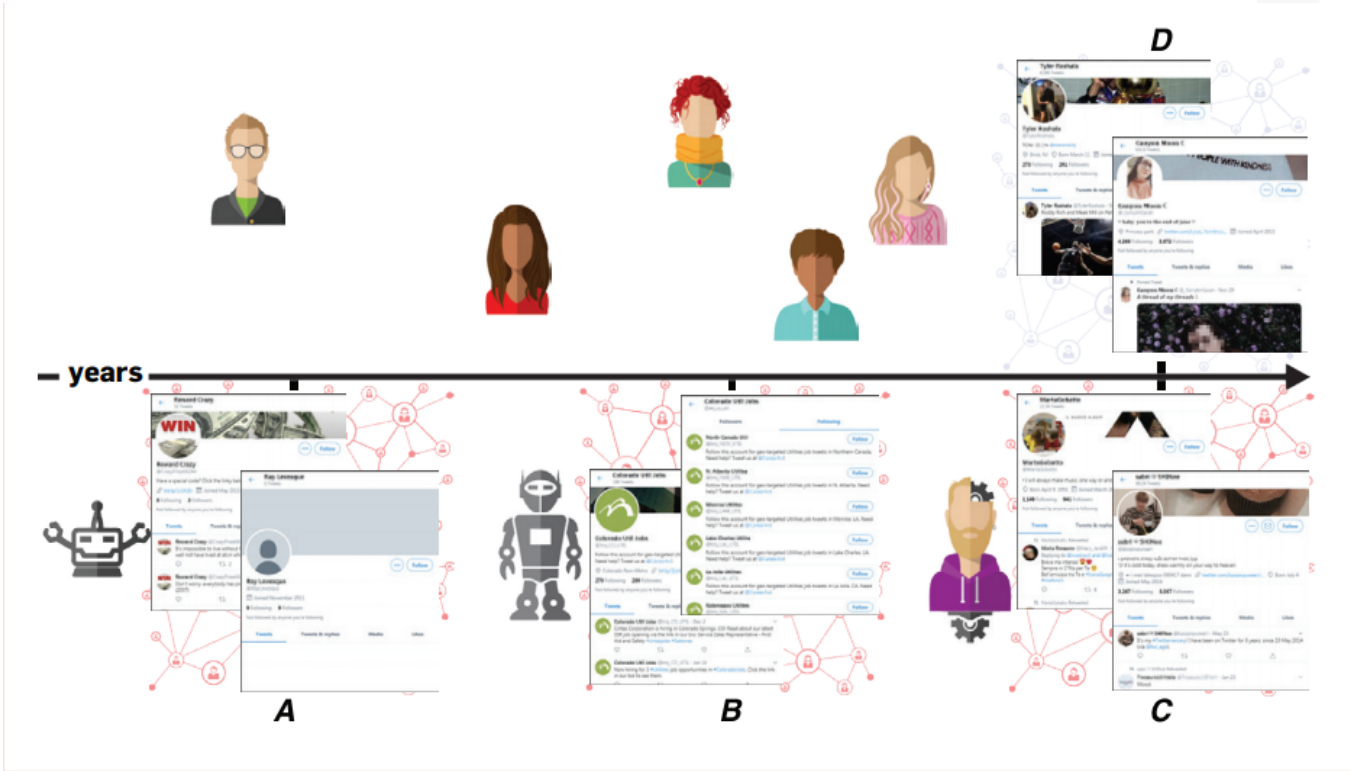
...

- RNN, TPS-ISA 2019
- RNN + Property, Information Sciences, 2018
- GAN, IJCAI 2019
- GCN, WWW 2019

...

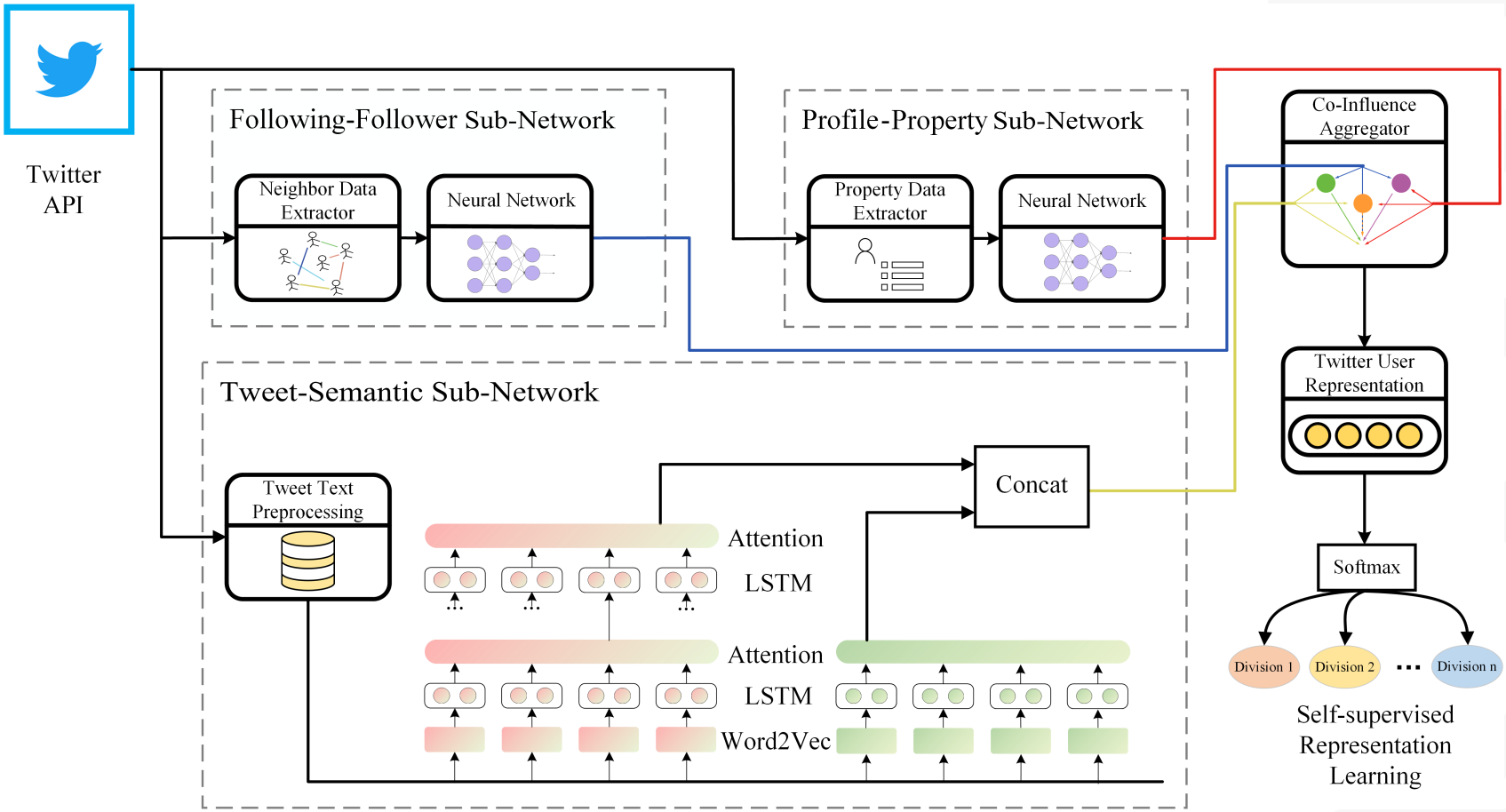
Task Challenges

- Generalization
  - Different kinds of Twitter Bots
- Adaptation
  - The evolution of Twitter Bots



Cresci et al., Communications of the ACM 2020

SATAR

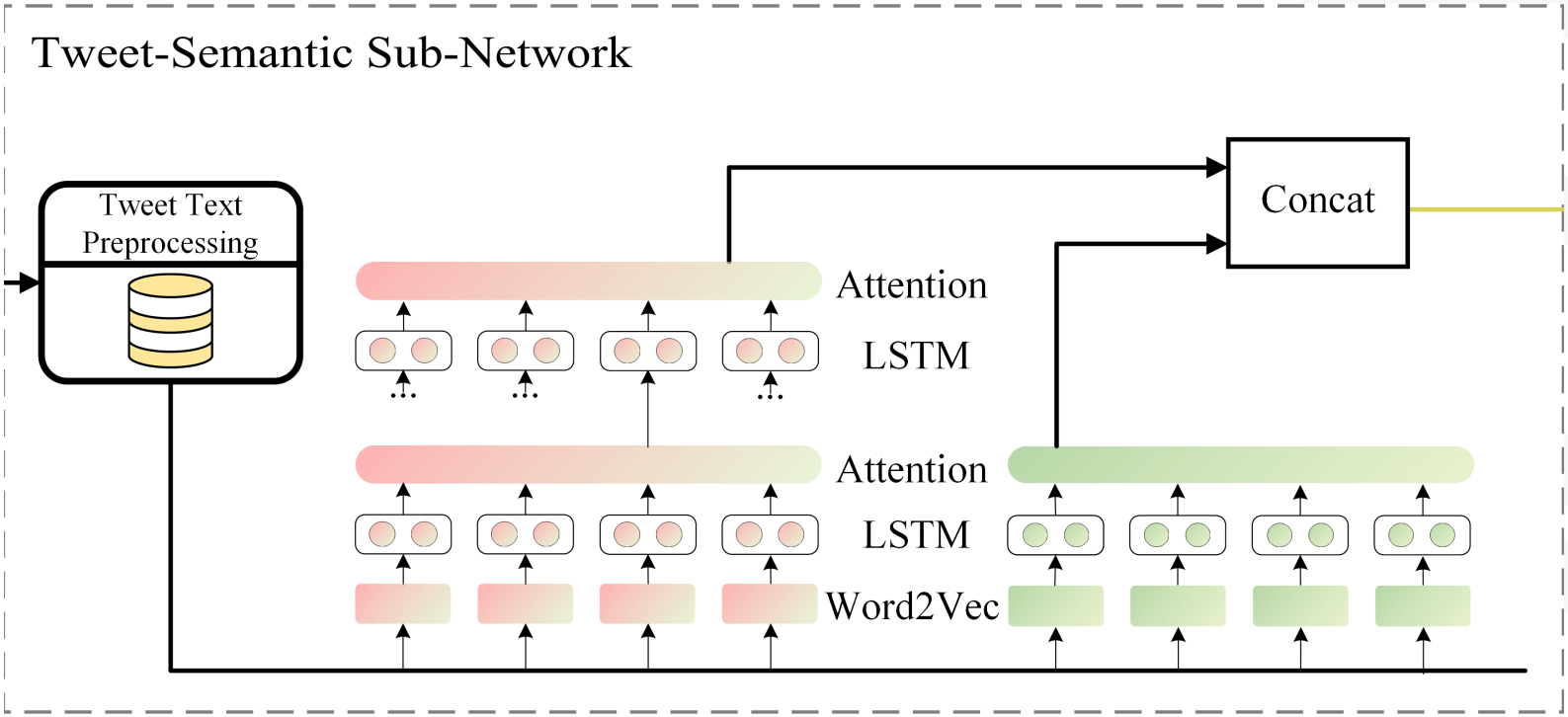


- Tweet-Semantic Sub-Network
- Profile-Property Sub-Network
- Following-Follower Sub-Network
- Co-Influence Aggregator

Tweet-Semantic Sub-Network

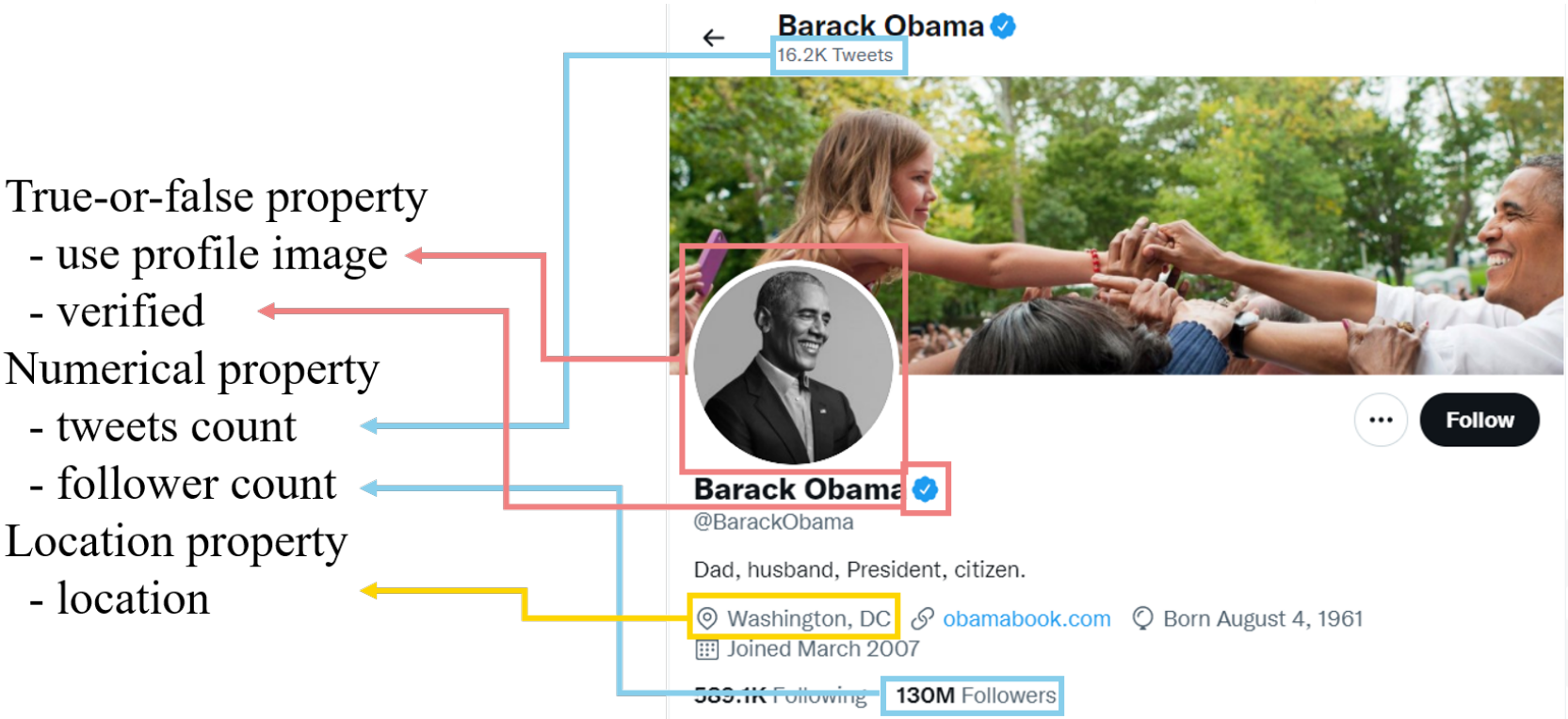
This sub-network exploits user information at two different semantic granularity.

- **Tweet-level**
- **Word-level**



Profile-Property Sub-Network

SATAR utilizes profile properties that could be directly retrieved from the Twitter API to avoid the undesirable bias.



Following-Follower Sub-Network

- For user followings,

$$u_n^f = \frac{1}{\sum_{u \in N^f} TF(u)} \sum_{u \in N^f} TF(u) r_s(u)$$

- For user followers,

$$u_n^t = \frac{1}{|N^t|} \sum_{u \in N^t} r_p(u)$$

Co-Influence Aggregator

Affinity index F_{sp}, F_{pn}, F_{ns}:

$$F_{sp} = tanh(r_s^T W_{sp} r_p)$$
$$F_{pn} = tanh(r_p^T W_{pn} r_n)$$
$$F_{ns} = tanh(r_n^T W_{ns} r_s)$$

Hidden representation h^s, h^p, h^n for each aspect:

$$h^s = tanh(W_s r_s + F_{sp}(W_p r_p) + F_{ns}(W_n r_n))$$
$$h^p = tanh(W_p r_p + F_{sp}(W_s r_s) + F_{pn}(W_n r_n))$$
$$h^n = tanh(W_n r_n + F_{ns}(W_s r_s) + F_{pn}(W_p r_p))$$

Twitter user representation r:

$$r = tanh(W_V \cdot concatenation(h^s; h^p; h^n))$$

Self-Supervised Learning

We believe that **follower count** would be an ideal self-supervised training signal.

- **Task-agnostic.** Follower count relates to all tasks on social media without being specific to any of them.

Self-Supervised Learning

We believe that **follower count** would be an ideal self-supervised training signal.

- **Task-agnostic.** Follower count relates to all tasks on social media without being specific to any of them.
- **Most Representative.** Follower count describes a Twitter user efficiently and accurately and involves the evaluation of other users.

Self-Supervised Learning

We believe that **follower count** would be an ideal self-supervised training signal.

- **Task-agnostic.** Follower count relates to all tasks on social media without being specific to any of them.
- **Most Representative.** Follower count describes a Twitter user efficiently and accurately and involves the evaluation of other users.
- **Robust to tamper.** An increase of 1,000 followers often costs from 13 to 19 U.S. dollars.
  S. Cresci et al. Decis. Support Syst 2015.

Experiment Settings

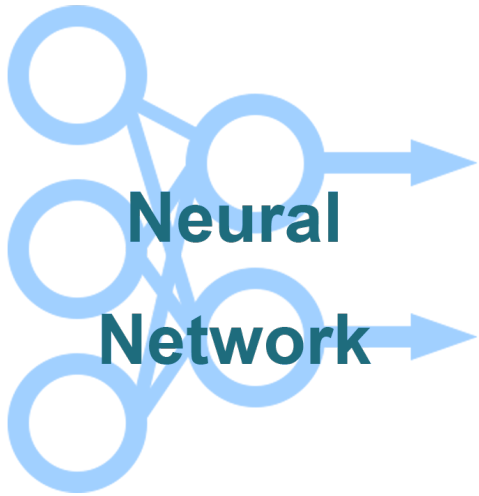
Dataset	User Count	Human Count	Bot Count
TwiBot-20	229,573	5,237	6,589
Cresci-17	9,813	2,764	7,049
PAN-19	11,378	5,765	5,613

- TwiBot-20 is a comprehensive benchmark. Feng *et al.*, CIKM 2021.
- Cresci-17 is a widely adopted dataset. Cresci *et al.*, IW3C2 2017.
- PAN-19 is a Bots and Gender Profiling shared task in the PAN workshop at CLEF 2019.

Experiment Settings

- Lee *et al.*, ICWSM 2011
- Yang et al., TIFS 2013
- Miller et al., Information Sciences
- Cresci et al., IEEE Intelligent Systems
- Botometer, WWW'16

...



- Kudugunta et al., Information Sciences
- Wei et al., TPS-ISA 2019
- Alhosseini et al., WWW'19

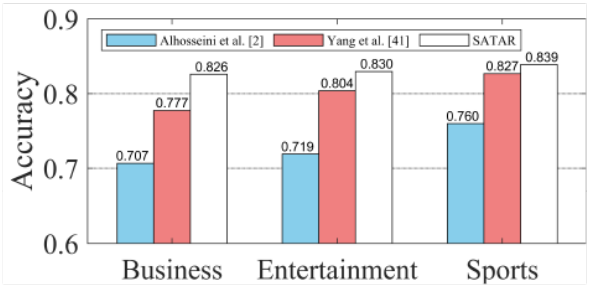
...

Bot Detection Performance

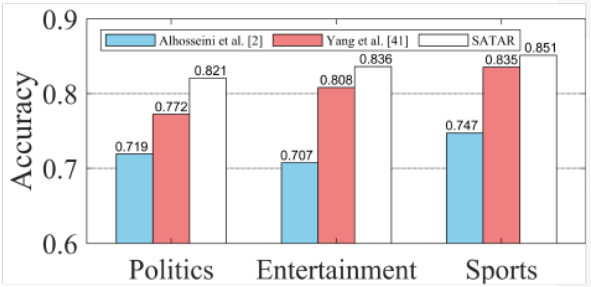
	Lee *et al.* [25]	Yang *et al.* [40]	Kudugunta *et al.* [23]	Wei *et al.* [38]	Miller *et al.* [30]	Cresci *et al.* [4]	Botometer [10]	Alhosseini *et al.* [1]	SATAR_{FC}	SATAR_{FT}
Semantic	✓		✓	✓	✓	✓	✓		✓	✓
Property	✓	✓	✓		✓		✓	✓	✓	✓
Neighbor							✓	✓	✓	✓

		Lee *et al.* [25]	Yang *et al.* [40]	Kudugunta *et al.* [23]	Wei *et al.* [38]	Miller *et al.* [30]	Cresci *et al.* [4]	Botometer [10]	Alhosseini *et al.* [1]	SATAR_{FC}	SATAR_{FT}
TwiBot-20	Acc	0.7456	0.8191	0.8174	0.7126	0.4801	0.4793	0.5584	0.6813	0.7838	**0.8412**
	F1	0.7823	0.8546	0.7517	0.7533	0.6266	0.1072	0.4892	0.7318	0.8084	**0.8642**
	MCC	0.4879	0.6643	0.6710	0.4193	-0.1372	0.0839	0.1558	0.3543	0.5637	**0.6863**
Cresci-17	Acc	0.9750	0.9847	0.9799	0.9670	0.5204	0.4029	0.9597	/	0.9622	**0.9871**
	F1	0.9826	0.9893	0.9641	0.9768	0.4737	0.2923	0.9731	/	0.9737	**0.9910**
	MCC	0.9387	0.9625	0.9501	0.9200	0.1573	0.2255	0.8926	/	0.9069	**0.9685**
PAN-19	Acc	/	/	/	0.9464	/	0.8797	/	/	0.8728	**0.9509**
	F1	/	/	/	0.9448	/	0.8701	/	/	0.8729	**0.9510**
	MCC	/	/	/	0.8948	/	0.7685	/	/	0.7456	**0.9018**

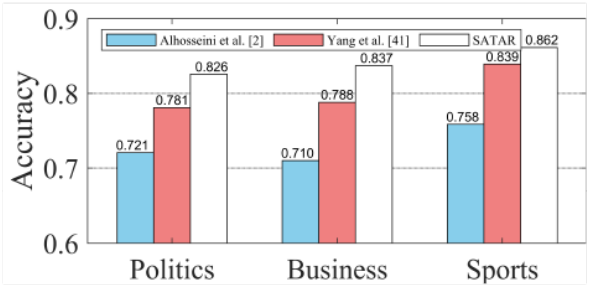
Generalization Study



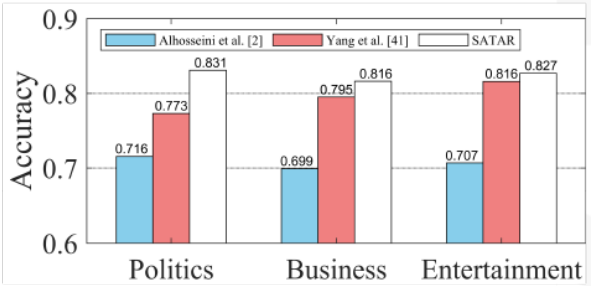
(a) train on politics domain



(b) train on business domain



(c) train on entertainment domain



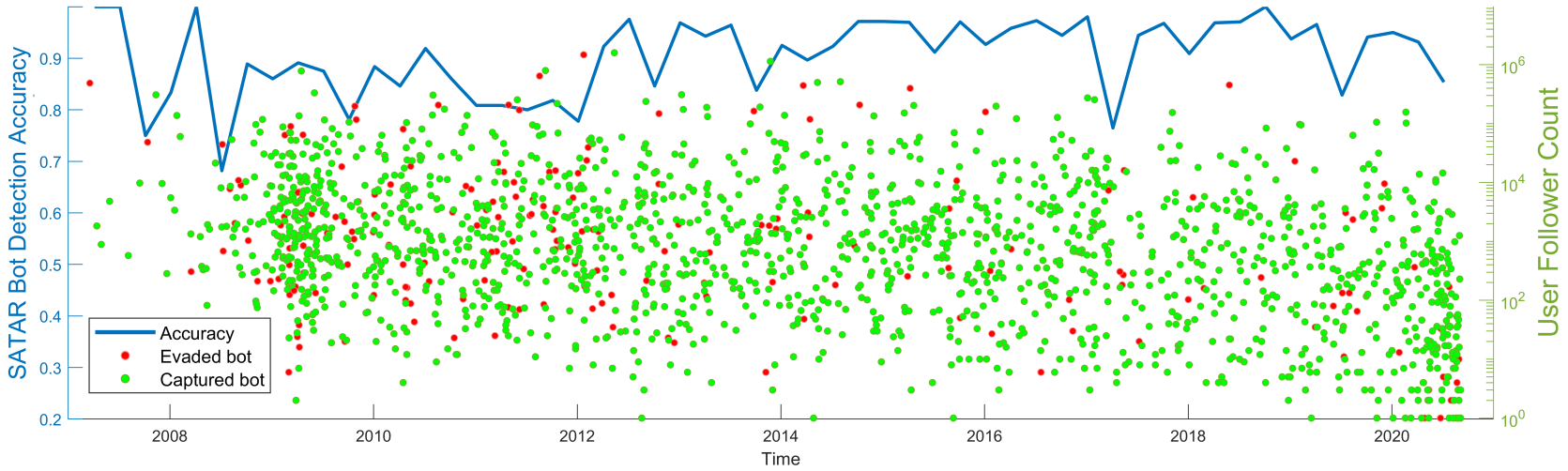
(d) train on sports domain

cross-domain bot detection

Generalization Study

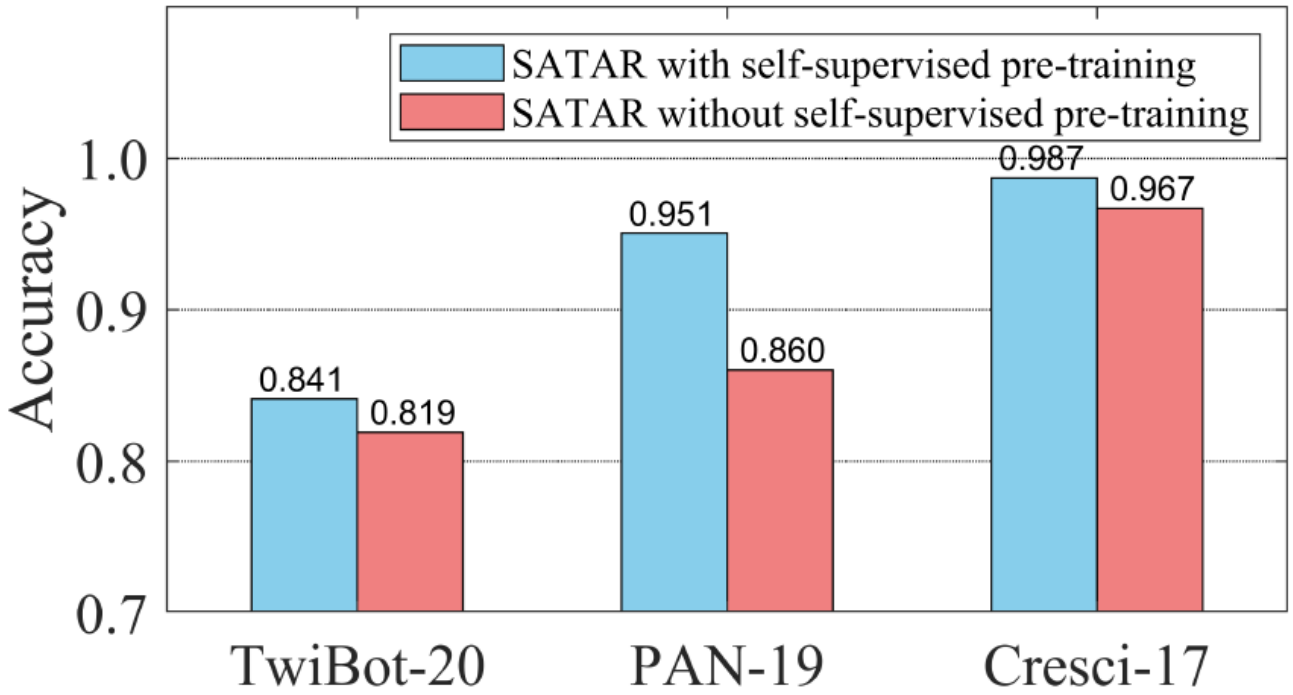
All semantic, property and neighborhood information are essential.

Adaptation Study



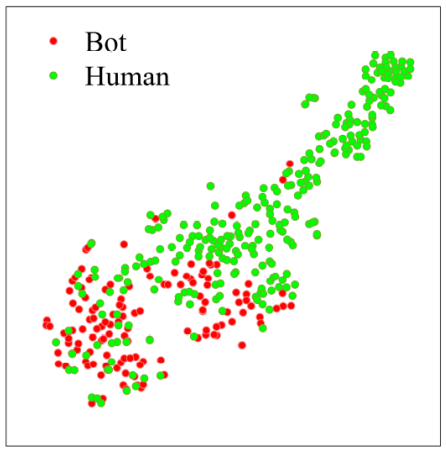
SATAR's performance is consistent over time.

Adaptation Study

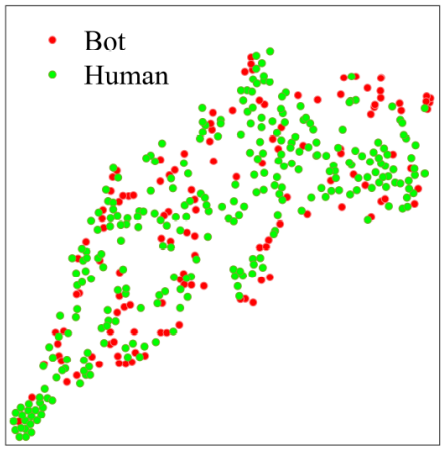


Self-supervised pre-training is essential.

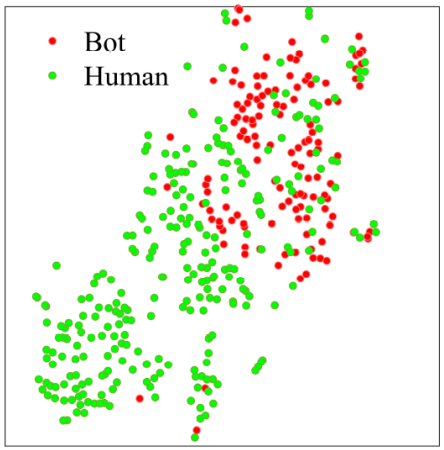
Representation Learning Study



(a) SATAR (Homogeneity Score: 1.577×10^{-1})



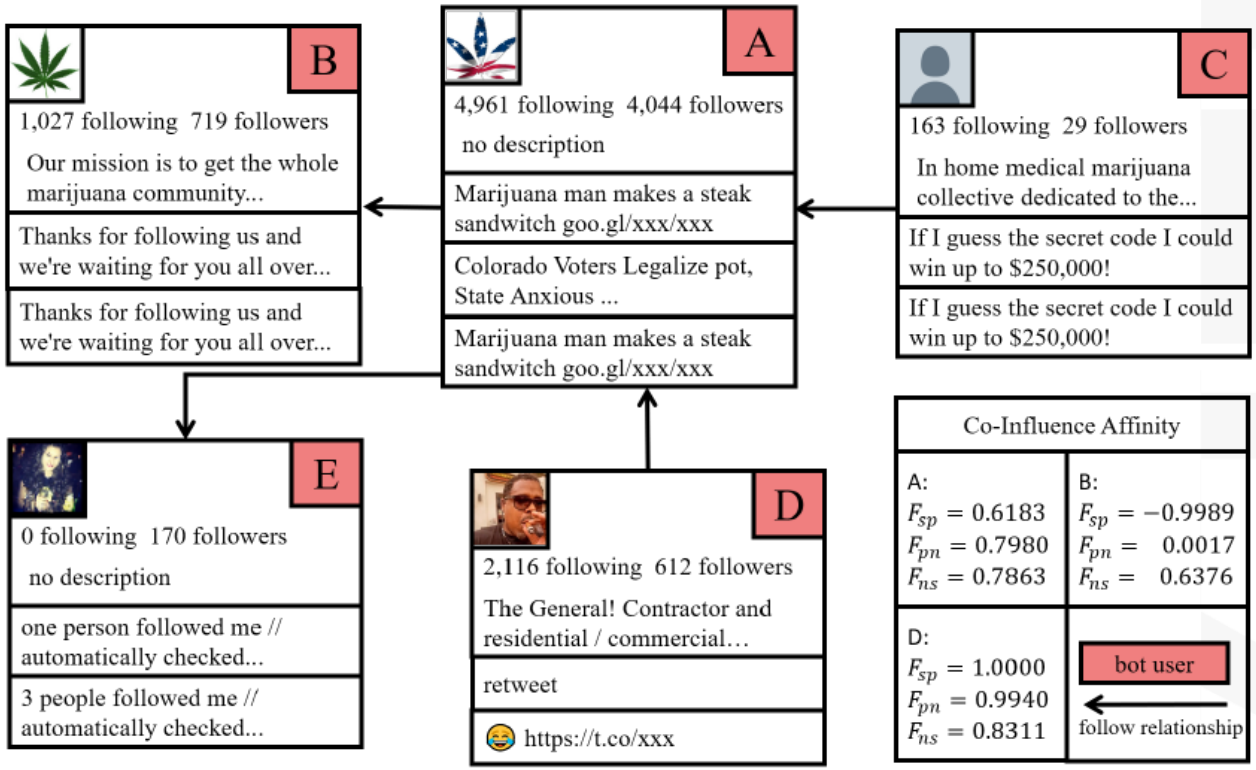
(b) Alhosseini et al. [2] (Homogeneity Score: 9.741×10^{-5})



(c) Yang et al. [41] (Homogeneity Score: 5.570×10^{-2})

SATAR learns high-quality user representation vectors.

Case Study



T, P and N all play a role in SATAR's evaluation of twitter users.

Conclusion

We proposed SATAR, a self-supervised approach to Twitter account representation learning and applied it to the task of bot detection that:

- jointly uses semantic, property and neighborhood information of users without feature engineering
- is the first work to introduce self-supervised representation learning to improve the performance of bot detection
- outperforms baselines on all three datasets and is proved to **generalize** and **adapt** through further exploration

Resorces

We make the code and model of SATAR available at

- `https://github.com/BunsenFeng/SATAR`

For the datasets we used to train and test SATAR:

- TwiBot-20: `https://github.com/BunsenFeng/TwiBot-20`
- Cresci-17: `http://mib.projects.iit.cnr.it/dataset.html`
- Pan-19: `https://zenodo.org/record/3692340`

Thank You !

Shangbin Feng
LUD Lab, Xi'an Jiaotong University
wind_binteng@stu.xjtu.edu.cn
September 26, 2021