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

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

### Introduction

- Twitter bots are often operated to achieve malicious goals.
- Existing measures fail to:
  - generalize
  - adapt

### Model 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} |
|----------------------|----------------|-----------------|-----------------------|----------------|-------------------|-----------------|----------------|------------------------|----------|-----------|
| **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                  |          |           |

- SATAR consistently outperforms all state-of-the-art baselines on three Twitter bot detection data sets.
- SATAR\_{FT} outperforms SATAR\_{FC}, which demonstrates the efficacy of the pre-training and fine-tuning approach.
- SATAR is further proved to generalize to diversified real-world scenarios and adapt to the bot evolution.

### SATAR Overview

- Generalize semantic, property, neighborhood
- Adapt self-supervised representation learning

### Generalization Study

- SATAR could conduct cross-domain bot detection and successfully identifies different types of bots.

### Adaptation Study

- SATAR maintains steady detection accuracy for different generations of bots registered from 2007 to 2020.