BotSSCL: Social Bot Detection with Self-Supervised Contrastive Learning

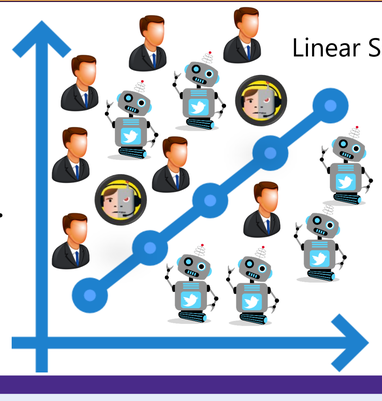
Mohammad Majid Akhtar†, Navid Bhuiyan†, Rahat Masood†, Muhammad Ikram‡ Salil S. Kanhere†
(†UNSW Sydney, ‡Macquarie University)

A Introduction and Problems

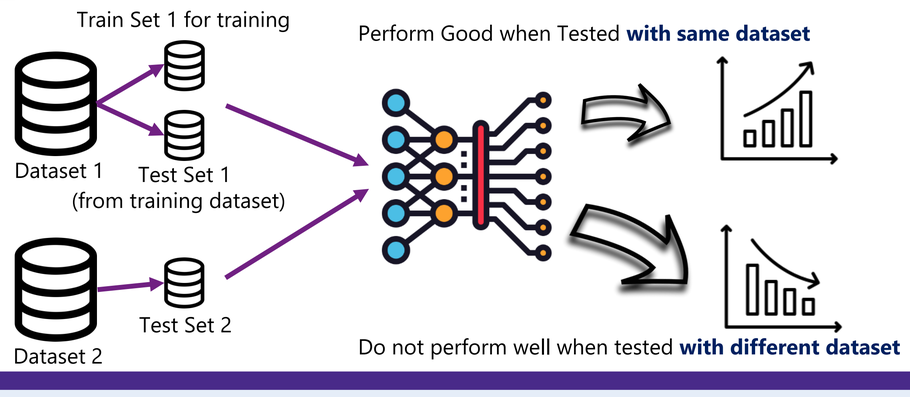
"'**Social Bot**' is an automated program that may spread false information and mimic genuine Online Social Network users to evade detection [1]."

Problem 1: Don't fairly detect sophisticated bots

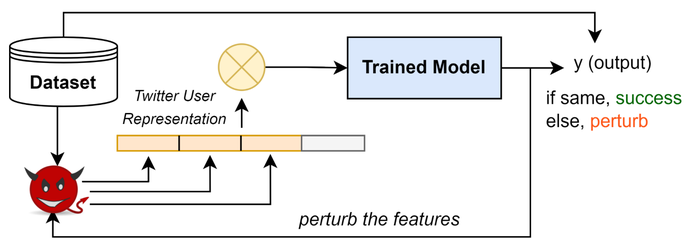
Existing Bot detectors have three issues [2].



Problem 2: Models suffer from Overfitting

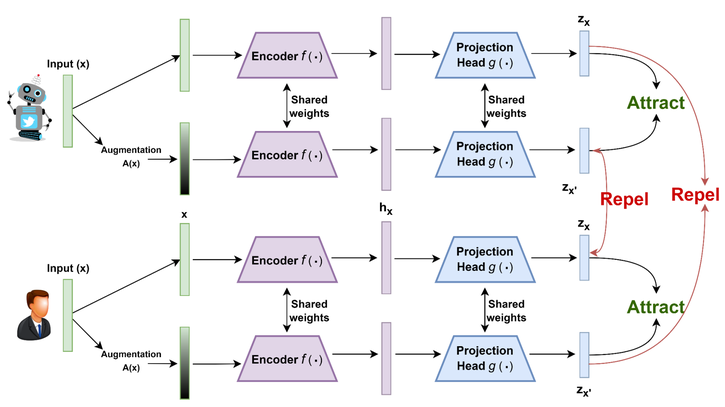


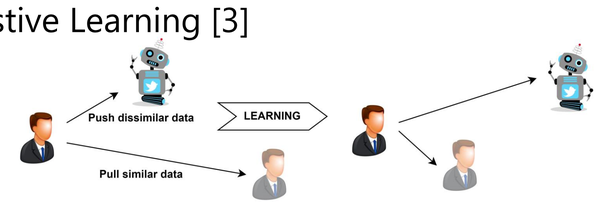
Problem 3: Vulnerable to Adversarial Attack



B Objective and Motivation

Motivation: To use Self-Supervised Contrastive Learning [3]





Problem Statement:
Given an account U and Tweets, we need to find $y(U) \in \{0,1\}$, where $y(U) = 1$ means bot.

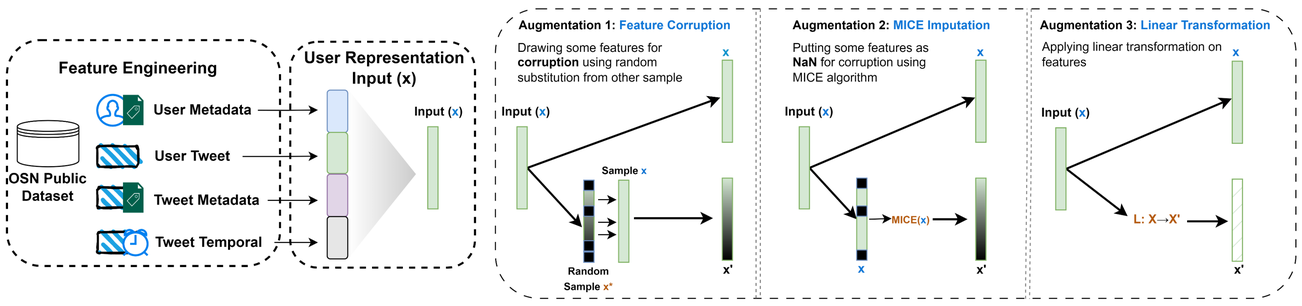
RQ1: Performance and Ablations

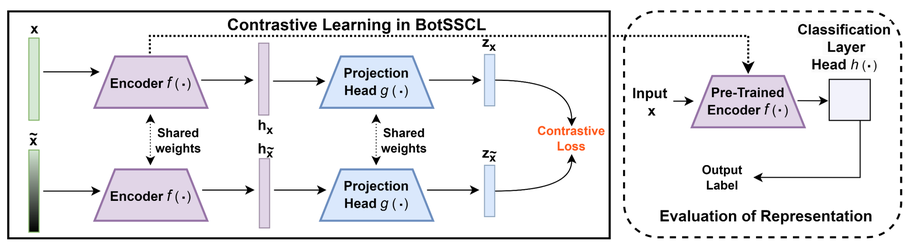
RQ2: Generalizability

RQ3: Adversarial Robustness

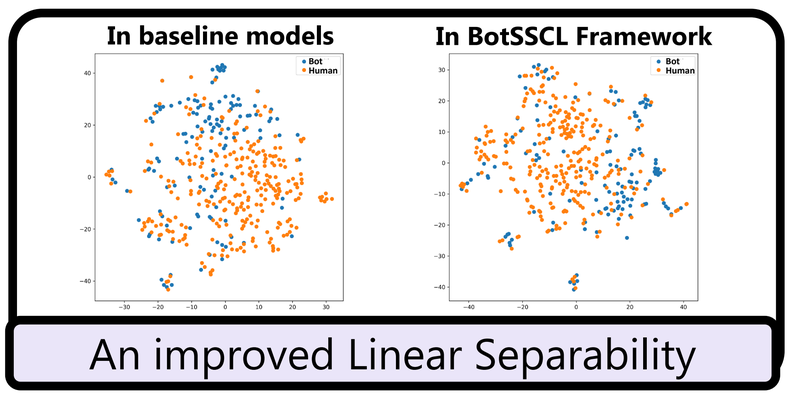
Perform tests on **Two Datasets (Varol and Gilani)** which includes these bots that mimic humans.

C Methodology



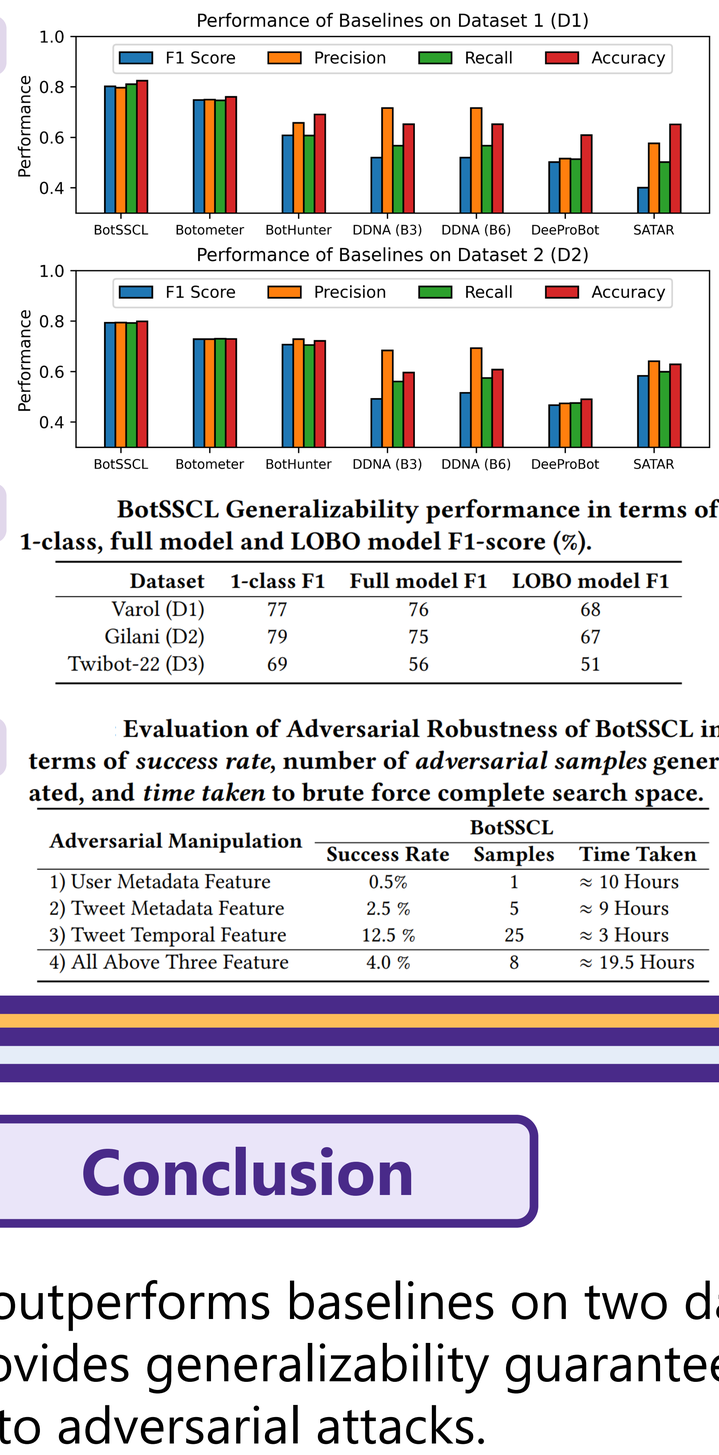


D Results



An improved Linear Separability

Result RQ1



Result RQ2

BotSSCL Generalizability performance in terms of 1-class, full model and LOBO model F1-score (%).

Dataset	1-class F1	Full model F1	LOBO model F1
Varol (D1)	77	76	68
Gilani (D2)	79	75	67
Twibot-22 (D3)	69	56	51

Result RQ3

Evaluation of Adversarial Robustness of BotSSCL in terms of *success rate*, number of *adversarial samples* generated, and *time taken* to brute force complete search space.

Adversarial Manipulation	BotSSCL		
	Success Rate	Samples	Time Taken
1) User Metadata Feature	0.5%	1	≈ 10 Hours
2) Tweet Metadata Feature	2.5 %	5	≈ 9 Hours
3) Tweet Temporal Feature	12.5 %	25	≈ 3 Hours
4) All Above Three Feature	4.0 %	8	≈ 19.5 Hours

E Analysis

- *≈6% and ≈8%* — We achieved ≈6% and ≈8% better than baseline models on both datasets.
- *≈67%* — BotSSCL is generalizable as it achieves similar performance when trained with any dataset and tested with other
- *≈4%* — Only allows 4% success to adversaries for evasion.

F Conclusion

BotSSCL outperforms baselines on two datasets. It also provides generalizability guarantees and is robust to adversarial attacks.

1. **Akhtar, M. M.**, Karunanayake, I., Sharma, B., Masood, R., Ikram, M., & Kanhere, S. S. (2023, October). Towards Automatic Annotation and Detection of Fake News. In 2023 IEEE 48th Conference on Local Computer Networks (LCN) (pp. 1-9). IEEE.
2. **Akhtar, M. M.**, Masood, R., Ikram, M., & Kanhere, S. S. (2023). False Information, Bots and Malicious Campaigns: Demystifying Elements of Social Media Manipulations. arXiv preprint arXiv:2308.12497.
3. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.