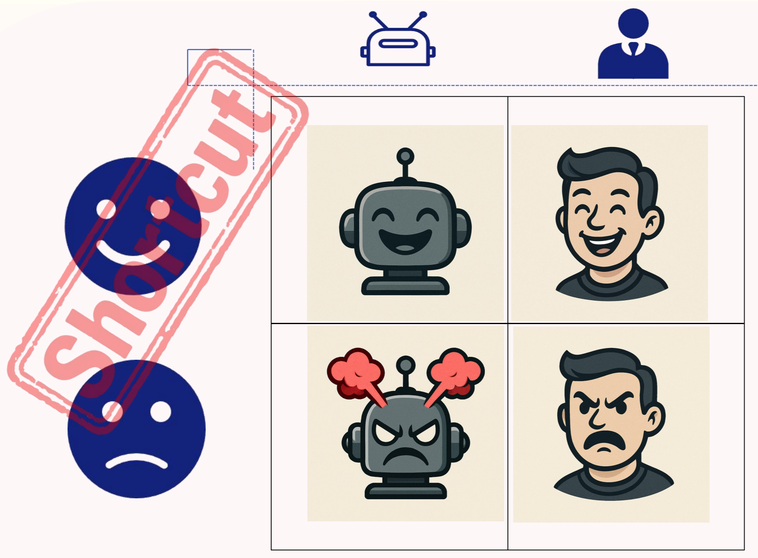
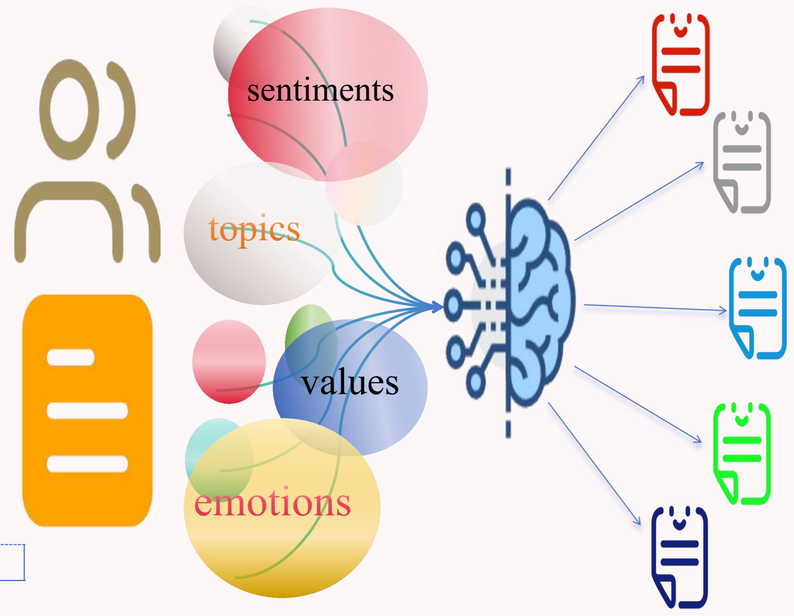
Bot Meets Shortcut: How Can LLMs Aid in Handling Unknown Invariance OOD Scenarios?

Shiyan Zheng[1], *Herun Wan*[1], *Minnan Luo*[1]*, *Junhang Huang*[2]

[1]*Xi'an JiaoTong University* [2]*Beijing Institute of Technology*





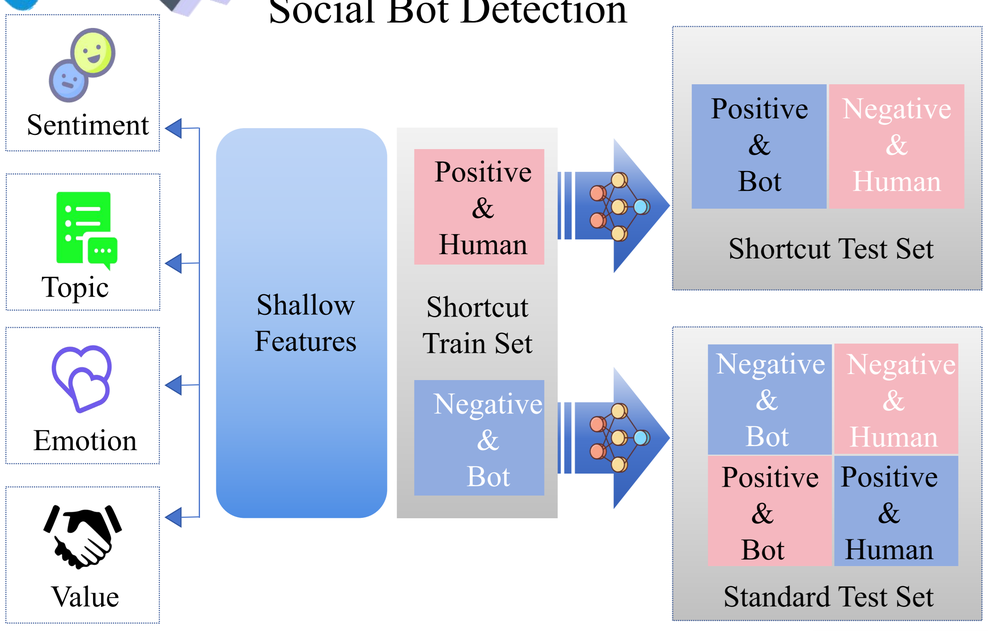
LLM enhancement angles on different features

Main Idea:

As social bots evolve, evaluating detectors under challenging shortcut-biased scenarios reveals vulnerabilities, motivating LLM-based strategies to uncover what truly makes an account a social bot.

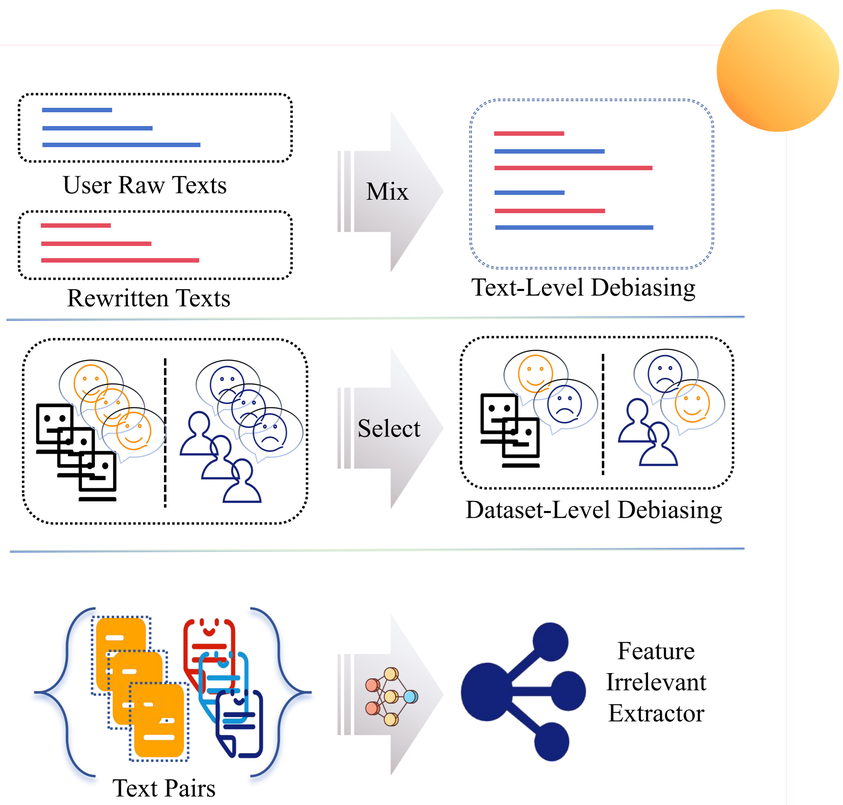
Work Overview

1. Shortcuts can affect social bot detection: Shortcut scenarios Setup for Social Bot Detection



RoBERTa		Cresci-2015-Data		Cresci-2017-Data		Twibot-20	
		$Shortcut_{te}$	$Standard_{te}$	$Shortcut_{te}$	$Standard_{te}$	$Shortcut_{te}$	$Standard_{te}$
Sentiments	$Standard_{tr}$	.945	.955	.884	.900	.682	.685
	$Shortcut_{tr}$	.565	.780	.287	.625	.051	.523
	Difference	40%↓	18%↓	67%↓	30%↓	92%↓	23%↓
Emotions	$Standard_{tr}$	.982	.979	.963	.940	.684	.687
	$Shortcut_{tr}$	.849	.917	.555	.766	.110	.520
	Difference	13%↓	6%↓	42%↓	18%↓	83%↓	24%↓
Topics	$Standard_{tr}$	.985	.973	.894	.885	.663	.684
	$Shortcut_{tr}$	.941	.962	.587	.757	.164	.535
	Difference	4%↓	1%↓	34%↓	14%↓	75%↓	21%↓
Values	$Standard_{tr}$	.915	.928	.890	.894	.691	.679
	$Shortcut_{tr}$	.659	.827	.514	.755	.180	.550
	Difference	28%↓	10%↓	42%↓	15%↓	73%↓	18%↓

RoBERTa's accuracy drops when moving from standard to shortcut-based.



RoBERTa		Cresci-2015-Data		Cresci-2017-Data		Twibot-20	
		$Shortcut_{te}$	$Standard_{te}$	$Shortcut_{te}$	$Standard_{te}$	$Shortcut_{te}$	$Standard_{te}$
Sentiments	$Shortcut_{tr}$	.565	.780	.287	.625	.051	.523
	AMR+CIGA	.678 (19%↑)	.838 (7%↑)	.271 (5%↓)	.671 (7%↑)	.112 (120%↑)	.531 (1%↑)
	Text-Level*	.840 (48%↑)	.910 (16%↑)	.597 (108%↑)	.757 (21%↑)	.229 (349%↑)	.559 (6%↑)
	Dataset-Level*	.835 (47%↑)	.912 (16%↑)	.667 (132%↑)	.794 (27%↑)	.288 (466%↑)	.571 (9%↑)
	Model-Level*	.864 (52%↑)	.907 (16%↑)	.690 (140%↑)	.836 (33%↑)	.415 (715%↑)	.580 (10%↑)
Emotions	$Shortcut_{tr}$	.849	.917	.555	.766	.110	.520
	AMR+CIGA	.770 (9%↓)	.870 (5%↓)	.382 (31%↓)	.679 (11%↓)	.232 (109%↑)	.538 (3%↑)
	Text-Level*	.936 (10%↑)	.964 (5%↑)	.670 (20%↑)	.819 (6%↑)	.235 (113%↑)	.569 (9%↑)
	Dataset-Level*	.936 (10%↑)	.971 (5%↑)	.610 (9%↑)	.782 (2%↑)	.288 (160%↑)	.569 (9%↑)
	Model-Level*	.923 (8%↑)	.933 (1%↑)	.628 (13%↑)	.803 (4%↑)	.400 (262%↑)	.585 (12%↑)
Topics	$Shortcut_{tr}$	.941	.962	.587	.757	.164	.535
	AMR+CIGA	.774 (17%↓)	.850 (11%↓)	.405 (30%↓)	.656 (13%↓)	.189 (15%↑)	.530 (0%↓)
	Text-Level*	.979 (4%↑)	.972 (1%↑)	.780 (32%↑)	.833 (9%↑)	.296 (80%↑)	.577 (7%↑)
	Dataset-Level*	.992 (5%↑)	.962 (0%↓)	.807 (37%↑)	.807 (6%↑)	.607 (270%↑)	.567 (5%↑)
	Model-Level*	.954 (1%↑)	.953 (0%↓)	.839 (42%↑)	.867 (14%↑)	.522 (218%↑)	.635 (18%↑)
Values	$Shortcut_{tr}$	.659	.827	.514	.755	.180	.550
	AMR+CIGA	.592 (10%↓)	.785 (5%↓)	.572 (11%↑)	.739 (2%↓)	.225 (24%↑)	.536 (2%↓)
	Text-Level*	.762 (15%↑)	.873 (5%↑)	.711 (38%↑)	.826 (9%↑)	.295 (63%↑)	.567 (3%↑)
	Dataset-Level*	.633 (3%↓)	.818 (1%↓)	.771 (50%↑)	.828 (9%↑)	.407 (126%↑)	.595 (8%↑)
	Model-Level*	.615 (6%↓)	.812 (1%↓)	.830 (61%↑)	.869 (15%↑)	.488 (171%↑)	.614 (11%↑)

LLM-based strategies greatly reduce shortcut effects.

2. LLM-based methods help mitigate shortcut: LLM-Based Mitigation Strategies for Shortcuts Social Bot Detection

Thank you for your interest!
Contact:
2223515385@stu.xjtu.edu.cn

Full Paper:

Code: