DualDet: A Dual-Task Detection Benchmark for Stance and Bot Detection on Social Media

Fuqiang Niu[1*], Zini Chen[2,3*], Hu Huang[1], Genan Dai[3†], Bowen Zhang[3†]
[1]University of Science and Technology of China, Hefei, China
[2]Shenzhen University, Shenzhen, China
[3]Shenzhen Technology University, Shenzhen, China
daigenan@sztu.edu.cn, zhang_bo_wen@foxmail.com

Abstract

Social media manipulation is often driven by automated accounts that amplify polarized viewpoints, posing persistent challenges to online information integrity. Two core capabilities for understanding and mitigating such manipulation are stance detection and bot detection. Although usually studied separately, they are coupled in practice: automation can skew stance distributions, while stance-aligned communities can shape visibility and network signals used for bot detection. We introduce DualDet, a dual-task benchmark for stance and bot detection spanning election discourse (Biden, Trump) and vaccine-related discourse (Vaccine). DualDet integrates inherited bot labels and follower-graph context with expert-annotated user-level stance labels, achieving substantial inter-annotator agreement (Cohen's $\kappa > 0.9$). The dataset contains 124,802 users in total, of which 22,906 users are annotated with expert stance labels. We further provide dataset analyses and reproducible baselines for stance detection, bot detection, and joint evaluation, revealing measurable stance–bot dependencies and highlighting open challenges for coupling-aware modeling.

Introduction

Social media platforms have become a dominant infrastructure for information consumption and public debate (Gil de Zúñiga, Jung, and Valenzuela 2012). They enable individuals and organizations to disseminate news, opinions, and calls for action at unprecedented scale and speed, and they offer researchers a unique lens for observing public discourse in the wild (Vosoughi, Roy, and Aral 2018). At the same time, the openness and high-velocity diffusion of these platforms make them vulnerable to manipulation: coordinated actors can strategically amplify selective narratives, distort perceived consensus, and intensify polarization (Shao et al. 2018; Pacheco et al. 2021). Such dynamics have been repeatedly documented across major political and public-health events in recent years, highlighting a persistent challenge for online information integrity (Broniatowski et al. 2018; Laabar and Zaghouani 2024).

In practice, analyzing manipulation on social platforms typically requires jointly characterizing (i) the position expressed in content toward a specific target, and (ii) the authenticity of the accounts that generate and amplify such content. The former is commonly formalized as stance detection, which classifies whether a post is in favor of, against, or expresses no clear stance toward a given entity or issue (target) (Mohammad et al. 2016; Hosseinia, Dragut, and Mukherjee 2020; Zhang et al. 2024a). The latter is formalized as bot detection, which aims to identify automated or semi-automated accounts that can operate at scale and shape diffusion through systematic posting and interaction patterns (Feng et al. 2021a, 2022b). Although stance detection and bot detection are often benchmarked separately, they are intertwined in real-world settings: automation can skew the observed distribution of stances, while stance-aligned communities can affect which accounts become visible and which network signals are available for bot detection (Shi et al. 2025). This motivates evaluating both tasks under a unified benchmark rather than through isolated, task-specific datasets.

However, existing datasets and benchmarks largely decouple stance detection from bot detection, limiting our ability to study their interaction under controlled and reproducible protocols (Feng et al. 2022b; Rostami et al. 2025; Shi et al. 2025). Stance datasets often provide reliable stance annotations but lack account-level bot labels aligned to the same users and content, and they typically provide limited social-graph context, making it difficult to quantify how automation affects stance distributions or model robustness. Conversely, bot detection benchmarks provide account labels and network structure but seldom include reliable stance annotations aligned to the same content, preventing integrated analysis of stance behavior conditioned on automation. As a result, several practically important questions remain difficult to address systematically: (i) do stance classifiers behave differently on bot-authored versus human-authored content? (ii) do bot detectors exhibit community- or stance-dependent failure modes? and (iii) can simple joint learning or pipeline approaches leverage cross-task signals without amplifying errors?

To bridge this gap, we introduce **DualDet**: Dual-Task Detection Benchmark, a dataset and evaluation suite designed for integrated study of stance detection and bot de-

*These authors contributed equally.
†Corresponding author.

tection in two societally impactful domains. DualDet contains election-related discourse with targets including Trump and Biden, as well as vaccine-related public-health discourse with target Vaccine. DualDet is derived from TwiBot-22, from which we inherit bot/human account labels and the follower relation as the primary graph structure. We newly annotate stance labels (Favor, Against and None) via expert annotation with substantial inter-annotator agreement measured by Cohen's κ (> 0.9). The resulting resource contains 124,802 users in total, of which 22,906 users are annotated with expert stance labels, enabling both supervised benchmarking and broader observational analyses of stance–bot coupling.

Beyond releasing the data, DualDet is accompanied by comprehensive dataset characterization and reproducible baselines. We provide detailed analyses of bot–human interactions, posting behavior, and lexical patterns to quantify coupling phenomena, with particular attention to how bots and humans interact within stance-aligned contexts. We further establish standardized benchmarks for stance detection and bot detection on the labeled subset, and we report baseline results for representative model families, including text-based stance classifiers, feature-based bot detectors, and graph-aware approaches leveraging follower edges. Finally, we evaluate straightforward joint settings to expose where current methods fail to capture cross-task dependencies.

In summary, our contributions are as follows: (1) we introduce DualDet, a dual-task benchmark aligning expert stance annotations with inherited bot labels and follower-graph context in election and vaccine discourse; (2) we provide expert stance labels with documented guidelines and quality control, including inter-annotator agreement statistics; (3) we present systematic analyses that quantify stance–bot coupling in both content and interaction patterns; and (4) we establish reproducible benchmarks and baselines for stance detection, bot detection, and joint diagnostic evaluation, offering a standardized foundation for future research on integrated modeling for online information integrity.

Related Work

Stance Detection

Early research on stance detection predominantly targets the tweet level, where the goal is to classify a post's stance toward a given target. Representative benchmarks include SemEval-2016 (SEM16) (Mohammad et al. 2016), P-Stance (Li et al. 2021), and COVID-19 stance datasets (Glandt et al. 2021), which provide labeled tweets for target-specific stance classification. While these datasets have been foundational for model development, they typically focus on content-level supervision without aligning stance annotations to account authenticity signals.

Subsequent work extends stance detection to settings with conversational structure, aiming to model stance expressions across post–reply threads and local discourse context. Datasets such as SRQ (Villa-Cox et al. 2020), MT-CSD (Niu et al. 2024) and C-MTCSD (Niu et al. 2025) incorporate reply-thread context for modeling; however, their supervision remains primarily post-centric and does not explicitly integrate broader interaction networks into the labeling protocol. As a consequence, they are limited in diagnosing how stance formation and diffusion relate to account-level behavioral properties at scale.

In parallel, user-level stance detection seeks to infer a user's overall stance by aggregating diverse signals beyond a single post. To reduce annotation cost, many UserSD datasets rely on heuristic or distant supervision, such as retweet patterns (Darwish et al. 2020; Samih and Darwish 2021), stance-indicative hashtags (Zhang et al. 2024b), tweet aggregation (Gambini et al. 2023; Elzanfaly, Radwan, and Othman 2023), or following behaviors (Zhu, He, and Zhou 2020). These strategies can scale but often introduce label noise and may overlook the structural signals that influence how stances are expressed and propagated. More recent efforts attempt to improve scale or efficiency via heuristic construction (Semcovici and Paraboni 2025) or LLM-assisted labeling (Rostami et al. 2025); nevertheless, the resulting stance annotations are often content-centric and do not explicitly leverage interaction structure in the labeling process.

Table 1 highlights a structural gap in existing stance resources: most datasets lack at least one of the following capabilities—aligned bot labels, explicit social-graph context, or structure-aware labeling (i.e., interaction structure is explicitly used during labeling). Our dataset is designed to fill this gap by providing expert stance annotations aligned with bot labels and follower-graph context, under a structure-aware labeling protocol, while remaining publicly available for reproducible research.

Bot Detection

Bot detection aims to identify automated or semi-automated accounts that can operate at scale and shape diffusion dynamics. A large body of work develops feature-based and graph-based detectors, and progress is strongly tied to the availability of high-quality benchmarks. Classic Twitter bot datasets include the Cresci collection, such as Cresci-15 (fake followers) (Cresci et al. 2015) and Cresci-17 (social spambots) (Cresci et al. 2017), which have been widely used to evaluate detection methods and characterize evolving automation strategies.

More recent large-scale benchmarks emphasize diversity and richer context. TwiBot-20 provides a large benchmark with user profiles, tweets, and follow relationships to support both individual and community-aware bot detection (Feng et al. 2021a). TwiBot-22 further advances graph-based evaluation with a large-scale Twitter network and diversified relations, facilitating research on relational and graph neural methods for bot detection (Feng et al. 2022b). Despite these advances, bot detection datasets are typically optimized for account authenticity labeling and do not provide reliable stance annotations aligned to the same users and content, making it difficult to study *what* positions are amplified by automated accounts under a shared protocol.

A closely related dataset is MGTAB (Shi et al. 2025), which jointly annotates stance and bot labels and provides multi-relational user graphs. However, existing analyses note two practical limitations: (i) it is released in a feature-only/normalized form without access to the original

Dataset	Stance	Bot	Human Annotation	Struct.-Aware	Public Available
SEM16	✓	✗	✓	✗	✓
P-Stance	✓	✗	✓	✗	✓
SRQ	✓	✗	✓	✗	✓
MT-CSD	✓	✗	✓	✗	✓
DoubleH	✓	✗	✗	✗	✗
POLITISKY24	✓	✗	✗	✗	✓
Cresci-15	✗	✓	✗	✗	✓
Cresci-17	✗	✓	✗	✗	✓
TwiBot-20	✗	✓	✗	✗	✓
TwiBot-22	✗	✓	✗	✗	✓
MGTAB	✓	✓	✓	✗	✗
Our	✓	✓	✓	✓	✓

Table 1: Comparison of stance and bot detection datasets. *Struct.-Aware* indicates that interaction structure is explicitly used during labeling.

raw textual data, restricting reproducibility and flexibility for alternative modeling pipelines; and (ii) its stance labeling protocol remains largely content-centric without explicitly leveraging social relations, i.e., it is not structure-aware under the definition in Table 1. In addition, MGTAB's released resource emphasizes engineered user features, precomputed tweet embeddings, and a set of relation types, which may limit fine-grained textual analyses and end-to-end text modeling.

Motivated by these limitations, our dataset introduces a unified benchmark that jointly supports stance detection and bot detection with (1) expert stance labels aligned to bot labels, (2) explicit follower-graph context, (3) structure-aware labeling that leverages interaction structure during annotation, and (4) public availability to support reproducible evaluation and future method development.

DualDet Dataset

Data Collection

To mitigate the structural limitations and annotation constraints of existing resources for joint stance and bot detection, we build our dataset on top of the TwiBot-22 corpus (Feng et al. 2022b). TwiBot-22 provides large-scale user behavior histories together with an explicit social graph, enabling structure-aware analysis while maintaining broad coverage of both bot and human accounts.

We focus on three targets that frequently appear in high-stakes public discourse: *Joe Biden*, *Donald Trump*, and *Vaccine*. These targets span two highly contested domains—election-related discussions and vaccine-related public health debates—and thus allow us to capture diverse stance expressions as well as realistic cross-community interaction patterns.

To retrieve target-relevant content and facilitate reliable stance labeling, we apply a dual filtering strategy that combines explicit keyword matching with domain-specific hashtags, following common practices in topic-specific social media collection (Kawintiranon and Singh 2021; Liang et al. 2024; Poddar et al. 2023). Concretely, we first identify tweets that contain target keywords (e.g., *Biden*, *Trump*, and vaccine-related mentions), and then further refine the set by requiring the presence of corresponding domain hashtags.

Target	Bot	Bot Tweets	Human	Human Tweets
Biden	2,629	6,067	26,584	58,783
Trump	3,175	10,036	24,967	68,338
Vaccine	5,061	10,809	62,386	179,613

Table 2: Statistics of users and tweets in DualDet across targets, including the numbers of bot and human users and their associated tweets.

	Labeled Users and Class Distribution				
Target	Bot	%	Human	%	Total
Biden	568	7.62	6,884	92.38	7,452
Trump	624	6.27	9,327	93.73	9,951
Vaccine	1,406	25.55	4,097	74.45	5,503
Total	2,598	11.34	20,308	88.66	22,906

Table 3: Statistics of the labeled subset after preprocessing.

This keyword–hashtag pairing reduces topic drift, improves thematic coherence within each target, and yields cleaner stance signals for subsequent expert annotation. The resulting collection covers both bot and human users inherited from TwiBot-22, supporting joint benchmarking under consistent topical conditions.

Data Preprocessing

After data collection, our dataset contains 124,802 users and 333,646 tweets (Table 2), forming a multi-target corpus for subsequent joint study.

We further leverage the interaction signals provided by TwiBot-22 to quantify bot–human engagement patterns in the collected data, including *follow*, *mention*, *reply*, and *quote* interactions. We observe substantial cross-type connectivity across targets. For example, there are 17,816 bot–human follow links and 1,502 bot–human mention interactions, indicating frequent engagement between bot accounts and human users. Among these interaction types, *follow* is the most prevalent and relatively stable signal, and we therefore use it as the primary relational backbone in our benchmark.

To construct a structurally coherent subset for downstream labeling and evaluation, we filter users by posting activity and retain users with sufficient tweet history. We then build a directed social graph from *follow* relationships among the retained users, where an edge $u \rightarrow v$ indicates that user u follows user v, capturing explicit and stable social affiliations (Zhu, He, and Zhou 2020). All graph extraction (and any necessary neighborhood expansion within the filtered user set) is performed before sampling users for labeling. After preprocessing and sampling, we obtain 7,452, 9,951, and 5,503 users for the *Biden*, *Trump*, and *Vaccine* targets, respectively (22,906 users in total; Table 3). We additionally extract 32,603 directed follow edges among these users to support structure-aware modeling.

Data Annotation and Quality Assurance

For bot detection, we directly adopt the binary account labels (Bot/Human) provided by TwiBot-22. These labels are

Target	Samples and Proportion of Labels						
	Against	%	Favor	%	None	%	Total
Biden	1,504	20.18	4,462	59.88	1,486	19.94	7,452
Trump	6,550	65.82	1,604	16.12	1,797	18.06	9,951
Vaccine	1,050	19.08	3,804	69.13	649	11.79	5,503
Total	9,104	39.74	9,870	43.09	3,932	17.17	22,906

Table 4: Label distribution of the labeled subset.

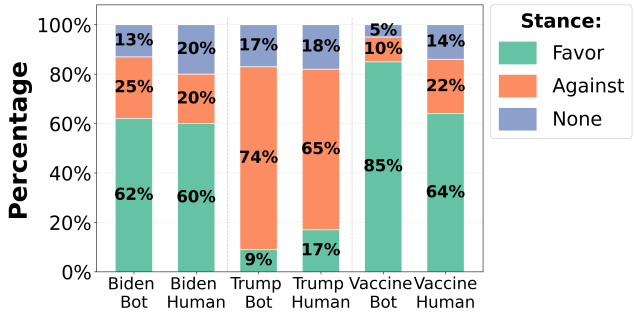


Figure 1: Stance distribution of bot and human users across three targets in the labeled subset.

produced through a weak-supervision pipeline that combines expert validation, ensemble model predictions, and probabilistic label refinement, yielding an expert-verified accuracy of 90.5%. This provides a reliable foundation for our bot detection benchmark.

For stance detection, we adopt a three-class scheme: *Favor*, *Against*, and *None*, indicating support, opposition, or no clear attitude toward the target, respectively. Different from prior datasets that annotate users without considering their social context, our annotation strategy explicitly leverages the underlying follow-network structure. Specifically, we partition users into three groups: *regular users*, *followed users* (users who are followed by others in the graph), and *isolated users* (users with no follow edges within the constructed subgraph). The labeling protocol is then tailored to each group. We label followed users and isolated users primarily based on manual inspection of their tweet histories and profile descriptions, reflecting their direct self-presentation and communication. For regular users, we consider both their own content and the stance signals implied by the set of accounts they follow, which better reflects how stance exposure and alignment may manifest under social affiliations.

We implement rigorous quality assurance throughout the annotation process: (1) Qualified annotators. Eight annotators with verified domain knowledge participated. Each annotator was required to complete a trial phase, which was reviewed by two senior adjudicators before joining the main annotation. (2) Blinded double annotation and multi-stage adjudication. Each instance was independently labeled by two annotators. Disagreements (16% of users) triggered a second-stage review, where a third adjudicator made the final decision. All annotations were conducted independently and blindly, without direct communication among annotators.

To quantify annotation reliability, we compute Cohen's κ (McHugh 2012) and overall inter-annotator agreement, following common practice of reporting agreement on the *Favor* and *Against* classes. The resulting κ scores for the *Biden*, *Trump*, and *Vaccine* targets are 0.90, 0.91, and 0.92, respectively, indicating near-expert agreement and high annotation quality. Compared with datasets relying on distant supervision or heuristics, our pipeline provides high-confidence stance labels that better support reliable benchmarking and analysis.

Data Analysis

After annotation, we obtain 22,906 users with both bot and stance labels. The distribution of stance labels across targets is reported in Table 4. In subsequent experiments, we conduct evaluation using a 70/15/15 train/validation/test split with a strict user-level partition. DualDet is publicly available at https://doi.org/10.5281/zenodo.19161528, which includes both the annotated subset and the full unannotated corpus.

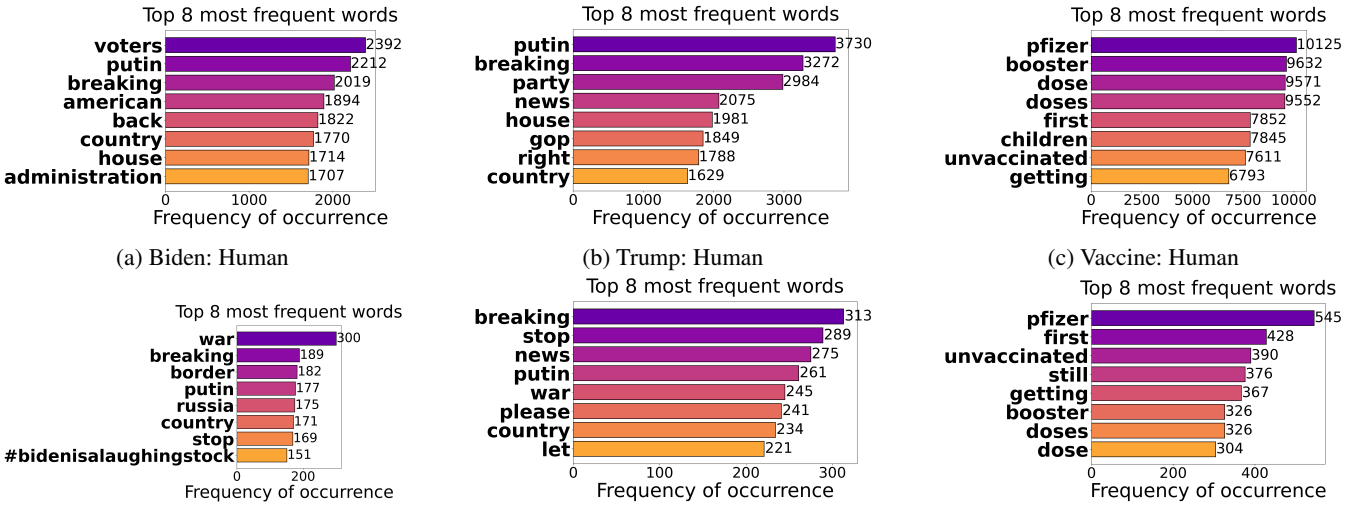
For the labeled subset, we first analyze the distributions of bot and stance labels. Beyond this subset, we further analyze the full collected corpus frvvom three additional perspectives: (i) lexical usage patterns, (ii) temporal posting dynamics, and (iii) interaction patterns between bot and human accounts.

Label Statistics and Stance Distribution Table 4 summarizes the stance label distribution for the three targets. Overall, the labeled subset is moderately imbalanced across targets and stance categories, reflecting real-world discourse skew. In total, *Favor* accounts for 43.09% (9,870 users), *Against* accounts for 39.74% (9,104 users), and *None* accounts for 17.17% (3,932 users).

We further compare stance distributions between bot and human accounts (Figure 1). Across all three targets, bot accounts exhibit a more polarized stance profile, with a higher proportion assigned to *Favor* or *Against* and a lower proportion assigned to *None*. For example, in the *Vaccine* target, bots are predominantly *Favor* (85%) with only 5% *None*, whereas humans show a higher *None* proportion (14%) and a less concentrated stance distribution. A similar pattern holds for *Trump*, where bots are more concentrated in *Against* (74%) compared to humans (65%). These results suggest that automation is associated with more stance-extreme participation, motivating joint evaluation of stance and bot detection.

Lexical Analysis via Word Frequency To understand how bots and humans discuss each target, we conduct a word-frequency analysis for each target and account type. Figure 2 visualizes the most frequent words for *Biden*, *Trump*, and *Vaccine*.

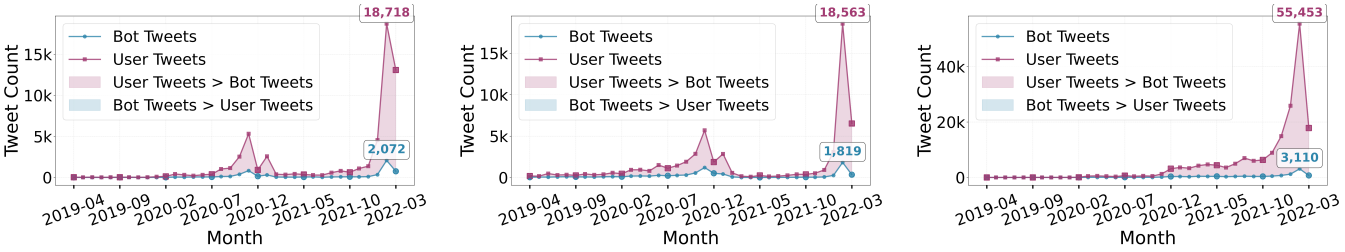
Overall, we observe consistent lexical differences between bot-generated and human-generated content across targets. Bot content tends to concentrate on a smaller set of highly repeated terms, suggesting more templated or campaign-like messaging, whereas human content exhibits greater lexical diversity and a broader range of target-related discussion. Across targets, bot-associated vocabularies more



(a) Biden: Human (b) Trump: Human (c) Vaccine: Human

(d) Biden: Bot (e) Trump: Bot (f) Vaccine: Bot

Figure 2: Word frequency analysis for the Biden, Trump, and Vaccine targets



(a) Biden (b) Trump (c) Vaccine

Figure 3: Comparative trends in monthly post volume between bot and human accounts for the Biden, Trump, and Vaccine targets.

frequently include attention-triggering or mobilizing terms such as *war* and *stop*. In the *Vaccine* target, bot content also shows stronger repetition of salient vaccination-related terms such as *unvaccinated*. These observations indicate that lexical concentration and repetition provide complementary cues for distinguishing automated activity and help explain why stance and bot signals are intertwined in practice. We leave more fine-grained lexical comparisons to future work.

Temporal Posting Dynamics We analyze monthly posting volumes for bot and human accounts in each target (Figure 3). Across targets, bots and humans exhibit broadly aligned temporal patterns, with major peaks occurring in similar time periods. This suggests that both account types are responsive to the same external events and news cycles, while bots may amplify and propagate related content during these high-attention windows.

We also observe substantial differences in absolute volume, with human accounts contributing more total tweets than bots. This gap is expected because the collected data contains more human users than bot users, and because active human participation dominates baseline discourse. Nevertheless, the temporal co-movement indicates that event-driven bursts may provide critical contexts where bot activity and stance polarization interact most strongly.

Interaction Patterns Between Bots and Humans We analyze interaction patterns between bot and human accounts using the available relation signals (e.g., *follow*, *mention*, *reply*, and *quote*). Across all three targets, bot accounts are not isolated; instead, they are embedded in the same interaction space as human users. The degree and cross-type interaction distributions further suggest that interactions are predominantly directed toward human accounts, while bot–bot interactions are comparatively rare. Detailed per-target distributions (in-/out-degree, cross-type interaction counts, and interaction ratios) are provided in Appendix Interaction Statistics.

Experimental Setup

Evaluation Metrics

We evaluate stance detection and bot detection using standard classification metrics. For *stance detection*, we follow

Task	Category	Methods	Biden		Trump		Vaccine		Avg.	
			F	Acc	F	Acc	F	Acc	F	Acc
Bot	Task-Specific	BotRGCN	28.21	92.28	74.68	85.30	57.07	77.42	53.32	85.00
		BotDGT	32.76	94.50	68.34	83.05	58.98	81.38	53.36	86.31
		RGT	35.29	94.08	**79.51**	89.50	63.02	**84.97**	59.27	**89.52**
		LMBot	2.53	91.99	47.34	73.49	27.35	73.67	25.74	79.72
		GCN	15.95	93.16	61.11	80.64	49.10	76.87	42.05	83.56
		GAT	27.33	93.35	63.45	81.29	50.47	78.65	47.08	84.43
		SimpleHGN	36.45	**95.41**	70.38	85.43	67.72	83.92	58.18	88.25
		HGT	33.69	95.05	77.66	88.50	**68.42**	84.80	**59.92**	89.45
	LLM	GPT3.5	**50.62**	88.64	41.10	54.45	50.55	53.38	47.42	65.49
		GraphICL	14.52	12.43	16.90	11.05	32.01	24.97	21.14	16.15
		COLA	48.18	70.21	49.71	73.88	49.90	57.58	49.26	67.22
		DeepSeek	8.18	8.59	6.89	7.17	20.60	25.45	11.89	13.74
	Multi-Task	MG-SIN	48.50	91.42	48.72	**95.02**	38.35	62.19	45.19	82.88
		MTIN	48.01	92.42	48.67	90.09	52.62	70.42	49.77	84.31
Stance	Task-Specific	JointCL	35.54	35.95	28.48	28.64	27.90	27.03	30.64	30.54
		CrossNet	39.42	61.65	47.89	68.55	38.02	69.32	41.78	66.51
		TPDG	47.82	46.35	52.37	53.44	53.84	64.77	51.34	54.85
	LLM	GPT3.5	67.75	63.38	**72.72**	**73.21**	74.27	69.78	71.58	68.79
		GraphICL	**75.01**	60.55	61.29	46.89	76.45	55.59	70.92	54.34
		COLA	74.48	**69.86**	72.43	72.94	**83.37**	**80.30**	**76.76**	**74.37**
		DeepSeek	74.13	69.50	66.57	67.11	74.75	74.62	71.82	70.41
	Multi-Task	MG-SIN	65.87	59.66	36.43	46.95	22.34	21.65	41.55	42.75
		MTIN	37.46	59.66	68.12	65.54	72.21	74.32	59.26	66.51

Table 5: Results(%) on DualDet for bot detection and stance detection across the three targets. *Avg.* denotes the average performance over the three targets (Biden, Trump, and Vaccine). **Bold** indicates the best performance and underline indicates the second-best performance for each task. For bot detection, F denotes macro-F_1; for stance detection, F denotes the average of F_1 on Favor/Against

prior work Li et al. (2021); Mohammad, Sobhani, and Kiritchenko (2017) and report class-wise F_1 scores for the *Favor* and *Against* classes, denoted as F_{favor} and F_{against}, together with their average $F = \frac{1}{2}\left(F_{\text{favor}} + F_{\text{against}}\right)$ and overall accuracy (Acc). For *bot detection*, we report overall accuracy (Acc) and macro-F_1, over the two classes (Bot/Human), which is robust to class imbalance.

Baseline Methods

To evaluate the capability of representative models on stance detection and bot account detection, we conduct a comprehensive benchmark on DualDet. Our evaluation covers three major methodological paradigms.

Bot Detection Models (Task-Specific): **BotRGCN** (Feng et al. 2021b), **BotDGT** (He et al. 2024), **RGT** (Feng et al. 2022a), **LMBot** (Cai et al. 2024), **GCN** (Kipf 2016), **GAT** (Velickovic et al. 2017), **SimpleHGN** (Lv et al. 2021), **HGT** (Hu et al. 2020);

Stance Detection Models (Task-Specific): **JointCL** (Liang et al. 2022), **CrossNet** (Xu et al. 2018), **TPDG** (Liang et al. 2021);

LLM-Based Methods: **GPT-3.5**[1], **GraphICL** (Sun et al. 2025), **COLA** (Lan et al. 2024), **DeepSeek**[2];

Multi-Task Learning Baselines: **MG-SIN** (Chai et al. 2023), **MTIN** (Chai et al. 2022).

[1]https://openai.com/

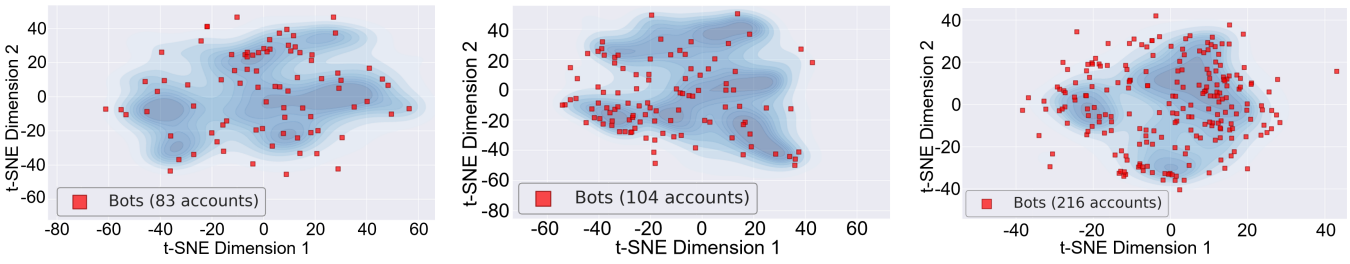
[2]https://www.deepseek.com/

Experimental Results

Main Results

Table 5 reports the main benchmark results on DualDet for bot detection and stance detection across three targets, together with the averaged scores (Avg.) over *Biden*, *Trump*, and *Vaccine*. Overall, we observe two consistent trends: (i) bot detection benefits substantially from structure-aware graph modeling, while (ii) stance detection is better handled by strong text-centric models, especially LLM-based approaches. Meanwhile, existing multi-task baselines do not consistently outperform strong single-task models, suggesting that effective joint modeling remains non-trivial.

Bot detection results For bot detection, task-specific graph models dominate the averaged performance. In particular, HGT achieves the best Avg. F (59.92), while RGT attains the best Avg. Acc (89.52), indicating that explicitly modeling relational structure is crucial for robustness across targets. Compared to simple homogeneous GNNs (e.g., GCN/GAT), heterogeneous or relation-aware models (SimpleHGN/HGT/RGT) generally perform better, especially on *Trump* and *Vaccine*, where bot-related interaction patterns are more distinguishable.



(a) Biden (b) Trump (c) Vaccine

Figure 4: Bot distribution over human density in the t-SNE embedding space for the *Biden*, *Trump*, and *Vaccine* targets. Red squares denote bot user, while the blue background indicates the kernel density of human user.

A notable phenomenon is that *Biden* yields high accuracies but relatively low F for many graph-based methods. This is consistent with the strong class imbalance in the *Biden* subset (few bot accounts), where accuracy can remain high even when the minority class is poorly recovered; therefore, macro-F_1 provides a more diagnostic signal. In contrast, *Vaccine* has a higher bot prevalence, and the best methods (e.g., HGT/RGT/SimpleHGN) achieve stronger and more balanced F–Acc trade-offs.

LLM-based baselines show limited effectiveness for bot detection overall (e.g., GPT-3.5 Avg. Acc 65.49), indicating that account authenticity is difficult to infer reliably from text alone without leveraging structural cues. Similarly, existing multi-task baselines improve over some weak text-only settings on certain targets but remain below the best task-specific graph models in averaged performance, highlighting the need for stronger coupling-aware designs.

Stance detection results For stance detection, LLM-based methods consistently outperform task-specific stance models on averaged scores. COLA achieves the best Avg. performance (Avg. F 76.76, Avg. Acc 74.37), and the second-best averaged results are obtained by DeepSeek (Avg. F 71.82, Avg. Acc 70.41). In contrast, conventional task-specific baselines (JointCL/CrossNet/TPDG) perform substantially worse, suggesting that user-level stance inference in DualDet is challenging for models primarily designed for tweet-level stance signals or limited context aggregation.

Across targets, stance prediction is most accurate on *Vaccine* for strong LLM-based methods (e.g., COLA reaches F 83.37 and Acc 80.30), while *Trump* exhibits relatively stronger polarization and thus benefits from models with robust semantic understanding (e.g., GPT-3.5 achieves competitive F/Acc on *Trump*). These differences indicate that target-specific discourse characteristics (e.g., polarization level and topical framing) can meaningfully affect stance modeling difficulty.

Implications for joint evaluation Taken together, the results provide two implications. First, bot detection and stance detection favor different dominant modeling signals (structure vs. semantics), which partially explains why naive multi-task learning does not consistently yield gains. Second, strong performance on one task does not directly translate to the other, motivating DualDet as a unified benchmark to evaluate whether future methods can capture the stance–bot coupling more effectively and improve robustness across targets.

Bot–Human Coupling in Representation Space

To examine whether bot accounts exhibit distributional coupling with human users, we project user representations into a shared embedding space and visualize their distributions. Specifically, we encode each user using a BERT-based text encoder applied to the user's historical tweets, and then reduce the representation dimensionality with t-SNE for visualization(Maaten and Hinton 2008). Figure 4 overlays bot accounts (red squares) on top of the density map of human accounts (blue background) for each target.

Across all three targets, bot accounts are not uniformly scattered in the embedding space; instead, they concentrate in regions with high human density. This pattern indicates that bots tend to occupy representation neighborhoods dominated by human users, suggesting strong behavioral and semantic coupling between bots and humans in discourse. Such coupling also helps explain why modeling stance and bot detection independently can be suboptimal: bots may deliberately mimic human-like content patterns to blend into human communities, while simultaneously amplifying stance-aligned narratives.

Discussion: Research Applications and Uses

DualDet enables research that jointly studies stance detection and bot detection under a unified benchmark with aligned labels and follower-graph context. It supports (i) diagnostic analyses of stance–bot coupling, such as how stance distributions differ between bot and human users and how model performance varies across stance-aligned communities; (ii) structure-aware modeling, including text–graph methods that leverage follower relations; and (iii) multi-task and joint learning to examine when cross-task signals help or hurt generalization across targets (Biden, Trump, and Vaccine). Overall, DualDet facilitates reproducible evaluation and measurement of coupled phenomena that are difficult to study with datasets covering only one task.

Limitations and ethical considerations DualDet has several limitations and ethical considerations. First, it covers only three targets (*Biden*, *Trump*, and *Vaccine*) and is derived from a Twitter benchmark, so its findings may not generalize to other topics, languages, platforms, or time periods. Second, bot labels are inherited from the source corpus and, although high-quality, may still contain noise. User-level stance labels rely on available tweet histories and profile information, and may therefore be affected by ambiguous, incomplete, or evolving user opinions. Third, our structure-aware protocol incorporates follow-network context, which may introduce homophily-related artifacts and make some users easier to classify, thereby affecting how models exploit neighborhood signals. Finally, the dataset exhibits class imbalance and may also be vulnerable to distribution shift as online discourse evolves over time.

From an ethical perspective, DualDet is derived from publicly available benchmark datasets and is intended solely for research purposes, such as studying stance detection, bot detection, and their interactions in online social media. The data is licensed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Although the resource is designed for research use, user-level identifiers and derived features may still involve privacy and re-identification risks and should therefore be handled with appropriate care. Potential misuse includes political profiling, harassment, or unjustified account enforcement. Misclassification costs are non-trivial, especially when human users are falsely labeled as bots. We therefore emphasize research-only use and encourage robustness, fairness, and bias auditing in downstream applications.

Conclusion

In this paper, we introduced DualDet, a dual-task benchmark for jointly studying stance detection and bot detection with aligned labels and explicit follower-graph context across three targets: *Biden*, *Trump*, and *Vaccine*. DualDet contains a large-scale corpus of 124,802 users and 333,646 tweets, together with a labeled subset of 22,906 users that includes bot/human labels and expert-annotated user stance labels under a structure-aware construction.

Our analyses provide evidence of stance–bot coupling: bot accounts exhibit more polarized stance distributions than human users, and their representations concentrate in human-dense regions, suggesting strong behavioral and semantic overlap. We further establish comprehensive baselines spanning task-specific models, LLM-based methods, and multi-task learning, offering a reproducible benchmark for future work. Overall, DualDet bridges a practical gap in existing resources by unifying stance, bot labels, and interaction structure in a single benchmark, and it enables future research on coupling-aware modeling and evaluation for online discourse integrity.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (No. 62306184); the Natural Science Foundation for Top Talents of SZTU (Nos. GDRC202518 and GDRC202320); the Shenzhen Science and Technology Program (Nos. RCBS20231211090548077 and JCYJ20240813113218025); the Guangdong Basic and Applied Basic Research Foundation (2026A1515010133); and the Project for Improving Scientific Research Capabilities of Key Construction Disciplines in Guangdong Province (No. 2025ZDJS039).

References

Broniatowski, D. A.; Jamison, A. M.; Qi, S.; AlKulaib, L.; Chen, T.; Benton, A.; Quinn, S. C.; and Dredze, M. 2018. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. *American journal of public health*, 108(10): 1378–1384.

Cai, Z.; Tan, Z.; Lei, Z.; Zhu, Z.; Wang, H.; Zheng, Q.; and Luo, M. 2024. LMbot: distilling graph knowledge into language model for graph-less deployment in twitter bot detection. In *Proceedings of the 17th ACM international conference on web search and data mining*, 57–66.

Chai, H.; Cui, J.; Tang, S.; Ding, Y.; Liu, X.; Fang, B.; and Liao, Q. 2023. Mg-sin: Multigraph sparse interaction network for multitask stance detection. *IEEE Transactions on Neural Networks and Learning Systems*.

Chai, H.; Tang, S.; Cui, J.; Ding, Y.; Fang, B.; and Liao, Q. 2022. Improving multi-task stance detection with multi-task interaction network. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2990–3000.

Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; and Tesconi, M. 2015. Fame for sale: Efficient detection of fake Twitter followers. *Decision Support Systems*, 80: 56–71.

Cresci, S.; Di Pietro, R.; Petrocchi, M.; Spognardi, A.; and Tesconi, M. 2017. The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In *Proceedings of the 26th international conference on world wide web companion*, 963–972.

Darwish, K.; Stefanov, P.; Aupetit, M. J.; and Nakov, P. 2020. Unsupervised User Stance Detection on Twitter. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, 141–152.

Elzanfaly, D. S.; Radwan, Z.; and Othman, N. A. 2023. User Stance Detection and Prediction Considering Most Frequent Interactions. In Auer, M. E.; El-Seoud, S. A.; and Karam, O. H., eds., *Artificial Intelligence and Online Engineering*, 421–433. Springer International Publishing.

Feng, S.; Tan, Z.; Li, R.; and Luo, M. 2022a. Heterogeneity-aware twitter bot detection with relational graph transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3977–3985.

Feng, S.; Tan, Z.; Wan, H.; Wang, N.; Chen, Z.; Zhang, B.; Zheng, Q.; Zhang, W.; Lei, Z.; Yang, S.; et al. 2022b. Twibot-22: Towards graph-based twitter bot detection. *Advances in Neural Information Processing Systems*, 35: 35254–35269.

Feng, S.; Wan, H.; Wang, N.; Li, J.; and Luo, M. 2021a. Twibot-20: A comprehensive twitter bot detection bench-

mark. In *Proceedings of the 30th ACM international conference on information & knowledge management*, 4485–4494.

Feng, S.; Wan, H.; Wang, N.; and Luo, M. 2021b. BotRGCN: Twitter bot detection with relational graph convolutional networks. In *Proceedings of the 2021 IEEE/ACM international conference on advances in social networks analysis and mining*, 236–239.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gambini, M.; Senette, C.; Fagni, T.; and Tesconi, M. 2023. From Tweets to Stance: An Unsupervised Framework for User Stance Detection on Twitter. In Bifet, A.; Lorena, A. C.; Ribeiro, R. P.; Gama, J.; and Abreu, P. H., eds., *Discovery Science*, 96–110.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Gil de Zúñiga, H.; Jung, N.; and Valenzuela, S. 2012. Social media use for news and individuals' social capital, civic engagement and political participation. *Journal of computer-mediated communication*, 17(3): 319–336.

Glandt, K.; Khanal, S.; Li, Y.; Caragea, D.; and Caragea, C. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, volume 1.

He, B.; Yang, Y.; Wu, Q.; Liu, H.; Yang, R.; Peng, H.; Wang, X.; Liao, Y.; and Zhou, P. 2024. Botdgt: Dynamicity-aware social bot detection with dynamic graph transformers. *arXiv preprint arXiv:2404.15070*.

Hosseinia, M.; Dragut, E.; and Mukherjee, A. 2020. Stance Prediction for Contemporary Issues: Data and Experiments. In Ku, L.-W.; and Li, C.-T., eds., *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, 32–40. Online: Association for Computational Linguistics.

Hu, Z.; Dong, Y.; Wang, K.; and Sun, Y. 2020. Heterogeneous graph transformer. In *Proceedings of the web conference 2020*, 2704–2710.

Kawintiranon, K.; and Singh, L. 2021. Knowledge Enhanced Masked Language Model for Stance Detection. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4725–4735. Online: Association for Computational Linguistics.

Kipf, T. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Laabar, S.; and Zaghouani, W. 2024. Multi-Dimensional Insights: Annotated Dataset of Stance, Sentiment, and Emotion in Facebook Comments on Tunisia's July 25 Measures. In Afli, H.; Bouamor, H.; Casagran, C. B.; and Ghannay, S., eds., *Proceedings of the Second Workshop on Natural Language Processing for Political Sciences @ LREC-COLING 2024*, 22–32. Torino, Italia: ELRA and ICCL.

Lan, X.; Gao, C.; Jin, D.; and Li, Y. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the international AAAI conference on web and social media*, volume 18, 891–903.

Li, Y.; Sosea, T.; Sawant, A.; Nair, A. J.; Inkpen, D.; and Caragea, C. 2021. P-Stance: A Large Dataset for Stance Detection in Political Domain. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP Online Event, August 1-6*.

Liang, B.; Fu, Y.; Gui, L.; Yang, M.; Du, J.; He, Y.; and Xu, R. 2021. Target-adaptive Graph for Cross-target Stance Detection. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23*, 3453–3464.

Liang, B.; Li, A.; Zhao, J.; Gui, L.; Yang, M.; Yu, Y.; Wong, K.-F.; and Xu, R. 2024. Multi-modal Stance Detection: New Datasets and Model. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 12373–12387. Bangkok, Thailand: Association for Computational Linguistics.

Liang, B.; Zhu, Q.; Li, X.; Yang, M.; Gui, L.; He, Y.; and Xu, R. 2022. Jointcl: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, 81–91.

Lv, Q.; Ding, M.; Liu, Q.; Chen, Y.; Feng, W.; He, S.; Zhou, C.; Jiang, J.; Dong, Y.; and Tang, J. 2021. Are we really making much progress? revisiting, benchmarking and refining heterogeneous graph neural networks. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 1150–1160.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

McHugh, M. L. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3): 276–282.

Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; and Cherry, C. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval NAACL-HLT, San Diego, CA, USA, June 16-17*, 31–41.

Mohammad, S. M.; Sobhani, P.; and Kiritchenko, S. 2017. Stance and Sentiment in Tweets. *ACM Trans. Internet Technol.*, 17(3).

Niu, F.; Yang, M.; Li, A.; Zhang, B.; Peng, X.; and Zhang, B. 2024. A Challenge Dataset and Effective Models for Conversational Stance Detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 122–132.

Niu, F.; Yang, Y.; Fu, X.; Dai, G.; and Zhang, B. 2025. C-mtcsd: A chinese multi-turn conversational stance detection dataset. In *Companion Proceedings of the ACM on Web Conference 2025*, 769–772.

Pacheco, D.; Hui, P.-M.; Torres-Lugo, C.; Truong, B. T.; Flammini, A.; and Menczer, F. 2021. Uncovering coordinated networks on social media: methods and case studies.

In *Proceedings of the international AAAI conference on web and social media*, volume 15, 455–466.

Poddar, S.; Basu, M.; Ghosh, K.; and Ghosh, S. 2023. Overview of the FIRE 2022 track: Information Retrieval from Microblogs during Disasters (IRMiDis). FIRE '22, 12–14. Association for Computing Machinery.

Rostami, P.; Rahimzadeh, V.; Adibi, A.; and Shakery, A. 2025. PolitiSky24: US Political Bluesky Dataset with User Stance Labels. *arXiv preprint arXiv:2506.07606*.

Samih, Y.; and Darwish, K. 2021. A Few Topical Tweets are Enough for Effective User Stance Detection. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2637–2646.

Semcovici, P.; and Paraboni, I. 2025. Social media user stance detection without stance text. In da Cunha, M. X. C.; Viana, D.; Maciel, R. S. P.; and Araújo, A. A., eds., *Proceedings of the 21st Brazilian Symposium on Information Systems, SBSI, 2025*, 1–8.

Shao, C.; Ciampaglia, G. L.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2018. The spread of low-credibility content by social bots. *Nature communications*, 9(1): 4787.

Shi, S.; Qiao, K.; Liu, Z.; Yang, J.; Chen, C.; Chen, J.; and Yan, B. 2025. Mgtab: A multi-relational graph-based twitter account detection benchmark. *Neurocomputing*, 130490.

Sun, Y.; Ma, Z.; Fang, Y.; Ma, J.; and Tan, Q. 2025. Graph-ICL: Unlocking Graph Learning Potential in LLMs through Structured Prompt Design. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Findings of the Association for Computational Linguistics: NAACL 2025*.

Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y.; et al. 2017. Graph attention networks. *stat*, 1050(20): 10–48550.

Villa-Cox, R.; Kumar, S.; Babcock, M.; and Carley, K. M. 2020. Stance in replies and quotes (srq): A new dataset for learning stance in twitter conversations. *arXiv preprint arXiv:2006.00691*.

Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *science*, 359(6380): 1146–1151.

Xu, C.; Paris, C.; Nepal, S.; and Sparks, R. 2018. Cross-Target Stance Classification with Self-Attention Networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 778–783.

Zhang, B.; Dai, G.; Niu, F.; Yin, N.; Fan, X.; Wang, S.; Cao, X.; and Huang, H. 2024a. A Survey of Stance Detection on Social Media: New Directions and Perspectives. *arXiv preprint arXiv:2409.15690*.

Zhang, C.; Zhou, Z.; Peng, X.; and Xu, K. 2024b. DoubleH: Twitter User Stance Detection via Bipartite Graph Neural Networks. In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media*, 1766–1778.

Zhu, L.; He, Y.; and Zhou, D. 2020. Neural opinion dynamics model for the prediction of user-level stance dynamics. *Information Processing & Management*, 57(2): 102031.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes. The data are derived from publicly available resources; we follow platform/dataset usage policies and adopt privacy-preserving practices (see Section Discussion: Research Applications and Uses).
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes. The paper contributes a dataset and a benchmark suite with comprehensive analyses and baselines, rather than proposing a new method.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes. We discuss potential artifacts such as class imbalance and distributional skew (see Section Discussion: Research Applications and Uses).
   (e) Did you describe the limitations of your work? Yes. See Section Discussion: Research Applications and Uses.
   (f) Did you discuss any potential negative societal impacts of your work? Yes. See Section Discussion: Research Applications and Uses.
   (g) Did you discuss any potential misuse of your work? Yes. See Section Discussion: Research Applications and Uses.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes. We provide documentation and a responsible release plan, and we support reproducibility by releasing data artifacts (see Section Discussion: Research Applications and Uses and the release link in Section DualDet Dataset).
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes.
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? N/A. This work does not involve hypothesis testing with theoretical results.
   (b) Have you provided justifications for all theoretical results? N/A.
   (c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? N/A.
   (d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? N/A.

(e) Did you address potential biases or limitations in your theoretical framework? N/A.

(f) Have you related your theoretical results to the existing literature in social science? N/A.

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? N/A.

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? N/A. This paper does not include theoretical proofs.

(b) Did you include complete proofs of all theoretical results? N/A.

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? Yes. We release data at https://doi.org/10.5281/zenodo.19161528.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? Yes. We describe data splits in the data analysis/experimental setup sections, and we follow the original papers' hyperparameter settings for baselines.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? Yes. We report results averaged over three random seeds.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? Yes. Experiments were run on an NVIDIA RTX 4090 GPU (24GB).

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? Yes. The benchmark covers multiple model families and targets, and the evaluation protocol matches the dataset's intended joint setting.

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? Yes. See Section DualDet Dataset.

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes. We cite the original datasets and baseline methods used.

(b) Did you mention the license of the assets? Yes. See Section Discussion: Research Applications and Uses.

(c) Did you include any new assets in the supplemental material or as a URL? Yes. https://doi.org/10.5281/zenodo.19161528.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? Yes. See Section Discussion: Research Applications and Uses.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Yes.Our data is built from publicly available social media content (see Section Discussion: Research Applications and Uses).

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes. Our dataset is publicly available at https://doi.org/10.5281/zenodo.19161528 with full FAIR compliance. Findable: Our dataset is available at https://doi.org/10.5281/zenodo.19161528; Accessible: Available under Creative Commons Attribution 4.0 International license; Interoperable: Standardized schemas aligned with TwiBot-22; Reusable: Complete documentation and usage examples.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? Yes. The dataset (with its Datasheet) is published on Zenodo, an open-access repository that assigns a persistent DOI to ensure findability and citability. The metadata is fully open, and we apply a Creative Commons license to facilitate reuse.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? Yes. See the Zenodo link https://doi.org/10.5281/zenodo.19161528.

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? Yes. This study utilizes the existing, publicly available TwiBot-22 dataset. Our analysis did not involve direct interaction with human subjects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? Yes. Participant compensation was calculated in accordance with our institution's standard remuneration guidelines. The estimated average hourly wage was 7$ (is higher than the local standard wage for similar jobs).

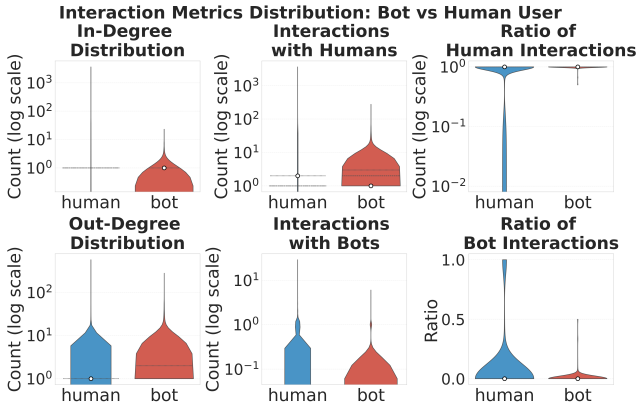
(d) Did you discuss how data is stored, shared, and deidentified? Yes. See the Zenodo link https://doi.org/10.5281/zenodo.19161528.

Appendix

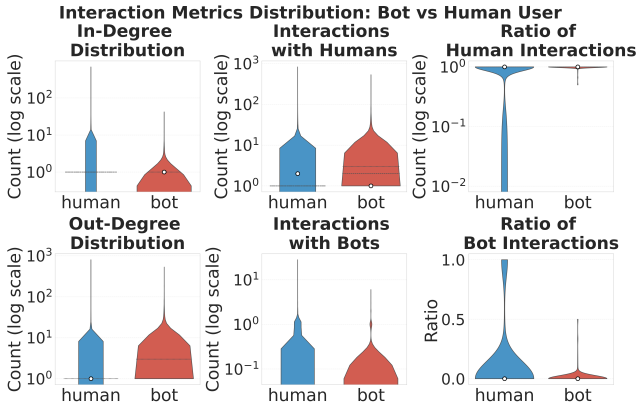
Interaction Statistics

Figures 5 visualize the distributions of interaction metrics for bot and human accounts under the three targets. For each target, we summarize user connectivity using (i) in-degree and out-degree on the induced interaction graph and (ii) cross-type interaction statistics, including the number of interactions with human accounts, the number of interactions with bot accounts, and the corresponding ratios.

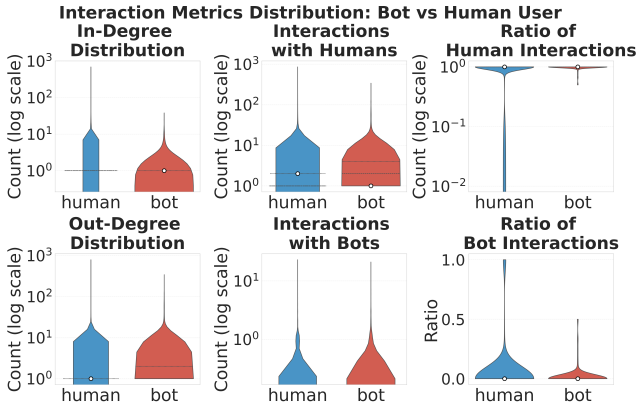
Degree distributions. Across targets, both in-degree and out-degree exhibit heavy-tailed behavior, indicating that a small fraction of accounts attract or generate substantially



(a) *Biden*



(b) *Trump*



(c) *Vaccine*

Figure 5: Visualizing human-bot interaction across three targets: (a) *Biden*, (b) *Trump*, and (c) *Vaccine*.

more interactions than the majority. We also observe systematic differences between account types: bot accounts tend to show stronger outgoing connectivity (higher out-degree), consistent with more active engagement or outreach behavior, while human accounts exhibit more pronounced variation in incoming connectivity (in-degree), reflecting the presence of highly visible human users.

Cross-type interaction counts. For all three targets, interactions with human accounts occur more frequently than interactions with bot accounts for both account types. This implies that bot accounts primarily engage with human users rather than forming dense bot-only clusters, and that human users are also exposed to non-trivial bot activity through routine interactions.

Interaction ratios. The ratio plots further confirm that the majority of interactions are directed toward humans: both bot and human accounts exhibit high human-interaction ratios, while the bot-interaction ratio is generally low with a long tail. This indicates that although bot–bot interactions exist, they are comparatively uncommon, and most observed coupling is manifested through bot–human engagement.