# TwiUSD: A Benchmark Dataset and Structure-Aware LLM Framework for User Stance Detection

**Fuqiang Niu**[1]*, **Zini Chen**[2,3]*, **Zhiyu Xie**[3]*, **Hu Huang**[1]†
**Qing Liao**[4], **Qianlong Wang**[3], **Genan Dai**[3], **Bowen Zhang**[3]
[1]School of Cyber Science and Technology,
University of Science and Technology of China, Hefei, China
[2]College of Applied Science, Shenzhen University, Shenzhen, China
[3]School of Artificial Intelligence, Shenzhen Technology University, Shenzhen, China
[4]School of Computer Science and Technology,
Harbin Institute of Technology, Shenzhen, China

## Abstract

Political user-level stance detection is vital for analyzing polarization, yet progress is hindered by the scarcity of high-quality benchmarks integrating linguistic and social signals. Existing datasets, largely relying on noisy heuristic or distant supervision, limit model robustness and generalizability. To address this, we introduce TwiUSD, a large-scale, expert-annotated benchmark for political user-level stance detection with explicit social network structure. TwiUSD comprises 16,211 users and 47,757 tweets, labeled by domain experts using a protocol that integrates both user content and followee signals, ensuring high-quality annotations (kappa $> 0.9$). Building upon TwiUSD, we propose MRFG, a Multi-scale Relevance Filtering and Graph-aware framework that leverages large language models to filter stance-relevant followee content and adaptively routes features based on structural informativeness. This design enables robust stance prediction by jointly modeling semantic and relational cues. Extensive experiments show that MRFG significantly outperforms strong baselines, highlighting the importance of relevance filtering and structure-aware modeling.

## 1 Introduction

Political stance detection is the task of automatically identifying users' attitudes toward political-related targets in social media (AlDayel and Magdy, 2021). As social media platforms like *X* (formerly *Twitter*) increasingly shape public opinion and influence electoral outcomes, the ability to reliably infer users' political stances has become both academically significant and societally urgent (Küçük and Can, 2020; Li et al., 2021).

Research on stance detection has focused on two main levels: tweet-level stance detection, where individual posts are classified with respect to a target, and user-level stance detection (UserSD), where the objective is to infer a user's overall stance by integrating signals from their textual and behavioral activity (Zhang et al., 2024a; Rostami et al., 2025). In the political domain, UserSD is particularly valuable, as users' stances are rarely expressed in a single tweet but instead emerge from their cumulative posting behavior and broader social context.

Despite this urgency, progress in political UserSD has been substantially limited by the lack of high-quality, large-scale benchmarks that capture both the linguistic signals and the underlying social structure of online communities. Existing datasets overwhelmingly rely on distant supervision or heuristic labeling–such as hashtag or retweet patterns–which introduce pervasive noise, systematic bias, and unreliable annotations (Darwish et al., 2020; Zhu et al., 2020; Samih and Darwish, 2021; Gambini et al., 2023; Zhang et al., 2024b). Our analysis shows that widely used hashtag-based heuristic labeling methods can misclassify up to 56% of users at the user level (see Appendix D), revealing the substantial noise and bias such approaches introduce. Because downstream stance detection models critically rely on the fidelity and structure of benchmark datasets, these limitations directly lead to models that are brittle and poorly generalizable. This critical gap poses a substantial challenge to UserSD, which underpins key areas such as election analysis, political polarization research, and misinformation detection. The lack of trustworthy user-level benchmarks that reflect both content and network context may compromise progress in stance modeling and its real-world applications.

To address these urgent challenges, we introduce **TwiUSD**, the first publicly available, large-scale, manually annotated political UserSD benchmark that explicitly incorporates social network structure. TwiUSD contains 16,211 users and 47,757 tweets, each meticulously labeled by domain experts using

*These authors contributed equally.
†Corresponding authors: huanghu@ustc.edu.cn



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 45198–45212
July 2-7, 2026 

a protocol that integrates both users' own content and the stances of their followees. The resulting dataset not only achieves near-expert annotation quality (kappa $>$ 0.9), but also exhibits unique structural realism and diversity, enabling rigorous and representative evaluation of UserSD models.

While TwiUSD enables more realistic and rigorous evaluation, it also presents unique challenges. The complex and noisy social structures captured in TwiUSD reveal significant limitations of prior methods, which often fail to effectively filter context or exploit structural cues. Political discussions are particularly noisy: users follow accounts with diverse political leanings, and not all followee content reflects or influences a user's own stance. Overcoming these challenges requires a more adaptive, structure-aware approach. Therefore, we propose **MRFG**, a *Multi-scale Relevance Filtering and Graph-aware Stance Detection* framework that selectively identifies stance-relevant signals from a noisy social context. MRFG consists of two components: the *Multi-scale Relevance Filter* (MRF) module, which uses a large language model (LLM) to assess the semantic relevance of followee tweets and filter out noisy content. It also ranks feature dimensions based on structural informativeness. The *Graph-Sensitive Inference* (GSI) module splits features into structure-sensitive and structure-neutral groups processed via a relational graph convolutional network (RGCN) and a simple multilayer perceptron network (MLP), respectively. Unlike prior work (Lorge et al., 2024) that aggregates all features or applies simple model ensembles, MRFG explicitly separates structurally informative features from content-based features and routes them through specialized encoders respectively. This adaptive feature separation and routing mechanism is essential for effectively capturing nuanced stance information in noisy social contexts.

Our main contributions are as follows: (1) We present TwiUSD, the first publicly available manually annotated UserSD benchmark with explicit social network structure. With 16,211 users, it is also the largest of its kind, setting a new and realistic benchmark for the field. (2) We propose MRFG, a novel structure-aware framework that leverages LLM for multi-scale relevance filtering and adaptively routes features, enabling robust stance prediction by jointly modeling semantic and relational signals. (3) Comprehensive experiments on TwiUSD show that MRFG significantly outperforms strong baselines, highlighting the importance of relevance filtering and structure-aware modeling for robust stance detection.

| Dataset | Social Graph | Human Annotation | Struct.-Aware | Public Available |
|---|---|---|---|---|
| DoubleH | ✓ | ✗ | ✓ | ✗ |
| UUSDT | ✓ | ✗ | ✓ | ✗ |
| POLITISKY24 | ✓ | ✗ | ✗ | ✓ |
| MGTAB | ✓ | ✓ | ✗ | ✗ |
| **TwiUSD** | ✓ | ✓ | ✓ | ✓ |

Table 1: Comparison of existing UserSD datasets and TwiUSD. "Struct.-Aware" indicates that interaction structure is explicitly used during labeling.

## 2 Related Work

**Stance Detection Datasets.** Early work on stance detection mainly focuses on the tweet level, as in datasets such as SemEval-2016 (SEM16), P-Stance, and COVID-19-stance (COVID-19) (Mohammad et al., 2016; Li et al., 2021; Glandt et al., 2021). For UserSD, early datasets are primarily constructed using heuristic or distant supervision strategies to reduce annotation cost. These datasets infer user stance from signals such as retweet patterns (Darwish et al., 2020; Samih and Darwish, 2021), shared hashtags (Zhang et al., 2024b), aggregated tweet content (Gambini et al., 2023; Elzanfaly et al., 2023), or following behaviors (Zhu et al., 2020).

In contrast, TwiUSD advances political UserSD by providing expert-level manual stance labels, a complete follow graph, and a structure-aware annotation protocol, enabling realistic evaluation of structure-aware and context-adaptive stance models, as summarized in Table 1.

**Stance Detection Approaches.** Early stance detection methods focused on modeling target-specific textual cues using attention mechanisms or graph-based formulations (Dey et al., 2018; Li et al., 2022; Pick et al., 2022; Barel et al., 2025). With the emergence of pre-trained language models (PLMs) and, more recently, LLMs, stance detection has progressively shifted toward fine-tuning, prompting, and LLM-based reasoning paradigms (Devlin et al., 2019; Shin et al., 2020; Liang et al., 2022; Huang et al., 2023; Cai et al., 2023; Lan et al., 2024; Sun et al., 2025).

For UserSD, most existing work aggregates tweet-level predictions (Samih and Darwish, 2021), applies clustering (Darwish et al., 2020), or uses LLM-based tweet scoring (Gambini et al., 2023).

Some studies leverage graph-based models (Zhang et al., 2024b) to capture user-tweet and user-user interactions. Semcovici and Paraboni (2025) stack BERT and Llama on non-stance tweets from users and their neighbors to infer stance. Rostami et al. (2025) generate stance labels for hashtag-filtered political users using a retrieval-augmented LLM. However, these methods often rely on heuristics, lack principled relevance filtering, and underutilize social structural information, leading to brittle performance in noisy, real-world contexts.

Despite recent advances, UserSD still struggles to robustly integrate semantic content with noisy and complex social relations, highlighting the need for structure-aware and context-adaptive benchmarks and frameworks.

| Stage | Target | User | Tweet |
|---|---|---|---|
| Stage-1 | Biden | 26,584 | 58,783 |
| | Trump | 24,967 | 68,338 |
| Stage-2 | Biden | 8,348 | 19,084 |
| | Trump | 10,837 | 31,253 |

Table 2: User and tweet counts by stage.

## 3 TwiUSD Dataset

**Data Collection.** To address the lack of structural realism and annotation quality in prior UserSD datasets, TwiUSD is constructed atop the TwiBot-22 corpus (Feng et al., 2022), which uniquely provides large-scale user histories and explicit social graphs (see Appendix A). Focusing on the most structurally and semantically polarized theme–namely, the U.S. presidential election–we systematically identified the two most salient stance targets: *Joe Biden* and *Donald Trump*. This selection ensures broad stance divergence and maximal network heterogeneity. Moreover, the 2020 election is among the most recent political events with publicly accessible user histories and social graphs, making it well suited for structure-aware user-level stance modeling. To maximize data authenticity and minimize annotation bias, we employed a two-stage filtering strategy: (1) all non-human (bot) accounts were excluded using the detection protocol in TwiBot-22; (2) we retained only tweets that both explicitly mention the target (*Biden* or *Trump*) and contain ideologically neutral, election-related hashtags (following the approach of Kawintiranon and Singh (2021); Liang et al. (2024)). This targeted selection is designed to reduce topical drift and improve labeling clarity. Full hashtag lists and filtering steps are detailed in Appendix B. The resulting subsets (see Table 2, Stage-1) provide a structurally diverse and thematically focused foundation for user-level stance annotation, enabling rigorous downstream modeling and analysis.

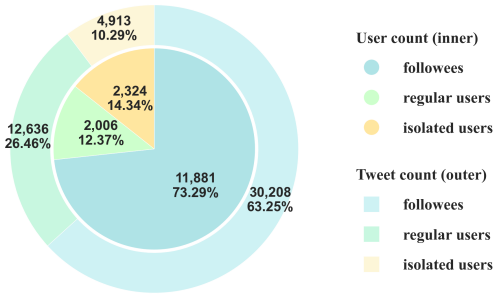


Figure 1: User and tweet distributions in TwiUSD.

**Data Preprocessing.** To enable effective stance modeling, we construct a directed follow graph from TwiBot-22. We first extract all *follow* relations among the filtered users to capture persistent social ties (Zhu et al., 2020), and identify *followees* as the accounts being followed and *regular users* as accounts that follow at least one followee. To enhance graph connectivity and reduce sparsity, we further expand the graph by adding two-hop followers of the followees, namely users who follow these regular users. All preprocessing steps are conducted prior to annotation, with strict separation of training and test users to prevent data leakage. After preprocessing, the dataset contains 1,010 and 1,113 followees for the *Biden* and *Trump* targets, respectively, and a total of 20,468 follow links, forming a structurally rich user graph for stance modeling (see Table 2, Stage-2).

**Data Annotation and Quality Assurance.** We adopted a three-class annotation scheme for user stance labels: *favor*, *against*, and *none*, indicating support, opposition, or no clear attitude toward the target, respectively. Unlike prior datasets that treat all users uniformly, our annotation strategy explicitly uses the underlying follow network structure. Users are divided into regular users, followees, and isolated users, and annotation protocols are designed accordingly. First, followees and isolated users are labeled based on manual inspection of their tweets and profile descriptions, reflecting their direct communication and self-presentation. Then, for regular users, stance labels are inferred by jointly considering their own content and con-

| Target | Samples and Proportion of Labels | | | | | | |
|---|---|---|---|---|---|---|---|
| | against | % | favor | % | none | % | Total |
| **Biden** | 1,360 | 19.76 | 4,110 | 59.70 | 1,414 | 20.54 | 6,884 |
| **Trump** | 6,089 | 65.28 | 1,546 | 16.58 | 1,692 | 18.14 | 9,327 |
| **Total** | 7,449 | 45.95 | 5,656 | 34.89 | 3,106 | 19.16 | 16,211 |

Table 3: Label distribution of the TwiUSD dataset.

textual information from their followees, mirroring realistic stance propagation patterns in social networks. In pilot annotation, we observed that, among users with available followee context, structure-aware judgments diverge from content-only judgments in approximately 25% of cases. This finding supports the use of followee context during annotation, especially when users express their stances only implicitly.

Rigorous quality assurance procedures are implemented throughout: (1) Qualified annotators: Eight annotators with verified domain knowledge participated, each required to pass a trial annotation phase reviewed by two senior adjudicators. (2) Blinded double annotation and multi-stage adjudication: Each instance was independently labeled by two annotators. Disagreements between the two annotators occurred in 16% of all annotated instances, which triggered a second-stage adjudication by an independent third annotator. All annotation was performed independently and blindly, with no direct communication among annotators.

To quantify reliability, we computed Cohen's kappa statistic (McHugh, 2012) and overall inter-annotator agreement, using the "favor" and "against" classes as in (Li et al., 2021). The resulting kappa scores for the *Biden* and *Trump* targets were 0.90 and 0.91, respectively, indicating near-expert agreement and high annotation quality. Compared to prior datasets relying on distant supervision or heuristics, our annotation pipeline ensures reliable and realistic stance labels, laying a solid foundation for downstream modeling. Detailed annotation guidelines and a representative example are provided in Appendix C.

**Data Analysis.** The TwiUSD[1] dataset comprises 16,211 users and 47,757 tweets. As shown in Figure 1, the majority of users are regular users (73.29%), with followees and isolated users representing 12.37% and 14.34%, respectively. Despite their smaller numbers, followees contribute over a quarter of all tweets (26.46%), indicating significantly higher activity levels and central roles in information dissemination. Table 3 details the stance label distribution. This multi-faceted user composition and activity heterogeneity provide a realistic testbed for stance modeling, and surpass the diversity of most existing user-level stance datasets.

To assess the reliability of hashtag-based unsupervised stance labeling, we applied two representative methods that infer user stance from stance-indicative hashtags. Our analysis reveals that these approaches suffer from substantial misclassification rates, with 56.16% of users incorrectly labeled with the opposite stance. This finding underscores the limitations and potential biases of hashtag-based supervision, and highlights the necessity of manual, context-sensitive annotation for robust user-level stance detection. Detailed comparison results are provided in Appendix D.

For supervised experiments, we partitioned the dataset for each target into training, validation, and test sets with a 70/15/15 split, ensuring robust evaluation and future reproducibility. All data collection and annotation procedures comply with Twitter's terms of service and institutional ethical standards.

## 4 Methods

### 4.1 Task Definition

Given a user $u$, our goal is to predict the user-level stance label $y_u \in \{favor, against, none\}$ toward a target entity (*Trump* or *Biden*). Each user is associated with multiple information sources, including the user's own tweet set $T_u$, the tweets posted by users that $u$ follows $T_{F(u)}$ where $F(u)$ denotes the set of users followed by $u$, the profile description $d_u$, and the directed follow relations in the user graph.

### 4.2 Framework Overview

As shown in Figure 2, MRFG is a relevance-driven and structure-aware framework composed of two tightly coupled modules: MRF and GSI. MRF serves as a relevance-aware controller that filters noisy social context and identifies structure-sensitive features by combining LLM-based semantic relevance estimation with TFI-based structural ranking. Based on these signals, GSI performs dual-path inference, routing structure-sensitive features through a RGCN and structure-neutral features through a MLP, whose outputs are fused for stance prediction. This relevance-guided separation and routing mechanism enables MRFG to selectively exploit social structure while avoiding indiscriminate aggregation of noisy context.

[1]https://github.com/nfq729/TwiUSD

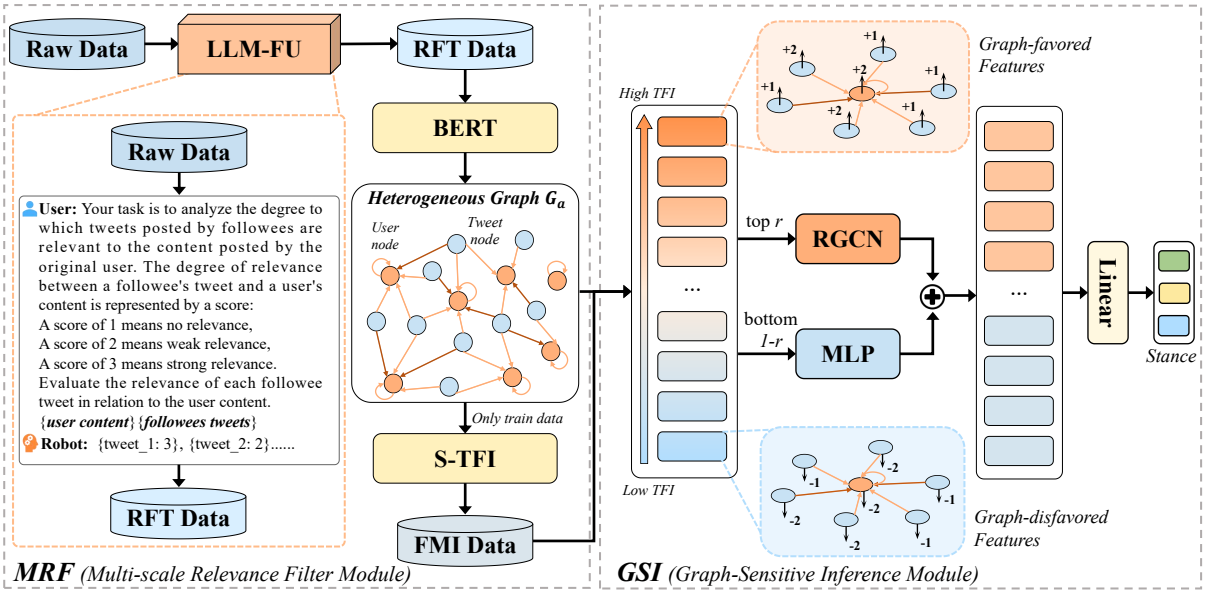


Figure 2: The architecture of our MRFG framework.

### 4.3 MRF Module

To mitigate the noise inherent in raw social contexts, we propose a two-stage filtering framework starting with an LLM-based Filtering Unit (LLM-FU). For each followee tweet $T_{F(u)}$, an LLM evaluates its semantic relevance to the target user's content on a scale of 1 (none) to 3 (strong). Only tweets with scores $\geq 2$ are retained as Relevance-Filtered Tweets (RFT), ensuring downstream modules focus on informative signals. (See Appendix E for prompts).

In the second stage, given the RFT data, we construct user representations and a relevance-aware graph to support feature-level informativeness estimation. We encode each user by concatenating their profile description $d_u$ and tweet set $T_u = \{T_u^1, \ldots, T_u^n\}$ into a single sequence separated by special tokens, which is fed into BERT with mean pooling to obtain the user representation $E_u$. Each retained followee tweet is encoded independently using the same encoder.

We then construct a directed heterogeneous graph $G_a$ with user and tweet nodes, where each retained followee tweet is linked to the target user by a directed edge typed by its LLM-FU relevance score. User nodes include self-loops for information propagation.

To identify which feature dimensions are most beneficial for stance inference in this graph, we adopt the Structural Topological Feature Informativeness (S-TFI) metric, which quantifies the mutual information between each feature dimension and the stance labels after graph-based feature smoothing. We choose S-TFI because it explicitly considers both the semantic content and the structural relationships in the social graph, making it particularly suited for our scenario where stance is influenced by both user-generated content and social connections. Formally, for user node features $X \in \mathbb{R}^{n \times d}$ and node labels $Y$, the informativeness of the $m$-th feature is defined as:

$$\mathrm{TFI}_m = I(Y; \tilde{X}_{:,m}), \quad \text{with } \tilde{X} = \hat{A}X \quad (1)$$

where $\hat{A}$ denotes the normalized adjacency matrix of the graph, and $\tilde{X}$ represents the smoothed features after propagation. This design ensures that features prioritized by S-TFI are not only semantically meaningful but also structurally aligned with the label distribution on the social graph.

Feature dimensions are ranked in descending order of their TFI scores, yielding a prioritized feature set referred to as feature mutual information (FMI) Data. All S-TFI computations are performed exclusively on the training set to prevent data leakage. Together, these two stages produce relevance-filtered and structure-aware feature representations that enable robust downstream stance inference.

### 4.4 GSI Module

Building on the relevance-filtered features produced by MRF, we propose the Graph-Sensitive Inference (GSI) module, which performs dual-path stance inference by separately encoding structure-sensitive and structure-neutral features. User fea-

tures are partitioned based on their S-TFI scores: the top ratio ($r$) features are selected as graph-favored features ($X_G$), while the remaining are treated as graph-disfavored features ($X_{\neg G}$). The hyperparameter $r$ is tuned on the validation set and fixed to 0.3 in all main experiments. The two feature subsets are processed by separate encoders.

**(1) Graph-Favored Encoding via RGCN.** We use a RGCN to process the graph-favored features from user nodes over the heterogeneous user-tweet graph $G = G_a = (V, E, \mathcal{R})$, where $\mathcal{R}$ denotes relation types (e.g., tweet-to-user, self-loops). This path captures higher-order structural dependencies that are essential for stance inference in social networks. Specifically, we select the top-$r$ TFI-ranked feature dimensions from user node embeddings as *graph-favored features*, denoted by $X_G$. These features are used as the initial input to the RGCN, capturing structural dependencies among users and their related content. Formally, given initial feature representations $h^{(0)} = X_G$, the hidden representation at layer $l+1$ for user node $i$ is computed as:

$$h_i^{(l+1)} = \sigma\left(\sum_{\zeta \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^{\zeta}} \frac{1}{c_{i,\zeta}} W_{\zeta}^{(l)} h_j^{(l)} + W_0^{(l)} h_i^{(l)}\right) \quad (2)$$

where $\mathcal{N}_i^{\zeta}$ denotes the set of neighbors of user node $i$ under relation $\zeta$, $c_{i,\zeta} = |\mathcal{N}_i^{\zeta}|$ is a normalization constant, and $W_{\zeta}^{(l)}$, $W_0^{(l)}$ are trainable parameters. $\sigma(\cdot)$ is a non-linear activation function. We use a 2-layer RGCN to obtain the final structure-aware representation $Z_G$ for each user node.

**(2) Graph-Disfavored Encoding via MLP.** The graph-disfavored features ($X_{\neg G}$) are processed independently using a two-layer MLP with ReLU activations. This pathway preserves semantic information that is less dependent on social structure and prevents the introduction of noise from unnecessary graph propagation. The output is denoted as $Z_{\neg G}$.

**(3) Feature Fusion and Stance Prediction.** The outputs of the two encoding paths are concatenated and passed to a final linear classifier:

$$Z = [Z_G \,||\, Z_{\neg G}], \quad \hat{y} = \text{Linear}(Z) \quad (3)$$

We use the standard cross-entropy loss for stance classification. All modules, including the text encoder and graph-based components, are trained jointly in an end-to-end fashion.

## 5 Experimental Setup

**Evaluation Metrics.** We use $F_1$ score and accuracy ($Acc$) to evaluate model performance, following Li et al. (2021) and Mohammad et al. (2017). Specifically, we report $F_{\text{favor}}$ and $F_{\text{against}}$ for the "*favor*" and "*against*" classes, their average $F_{\text{avg}}$, and overall accuracy.

**Baseline Methods.** We compare our method with representative baselines from four categories: *UserSD methods:* UUSDT(Darwish et al., 2020), Tweets2Stance(Gambini et al., 2023); *Supervised Neural Models:* TAN(Du et al., 2017), CrossNet(Xu et al., 2018); *Fine-Tuned PLMs:* BERT(Devlin et al., 2019), RoBERTa(Liu et al., 2019), BERT-GCN(Liu et al., 2021), TPDG(Liang et al., 2021), JointCL(Liang et al., 2022), KEPrompt(Huang et al., 2023); *LLM-Based Methods:* Llama 2/3, ChatGPT (3.5/4), COLA(Lan et al., 2024), GraphICL(Sun et al., 2025). Baseline model descriptions are provided in Appendix F.

**Implementation Details.** We use Claude-3.5, DeepSeek-R1, and GPT-4o for relevance scoring. The stance encoder fine-tunes bert-base-uncased with AdamW ($1 \times 10^{-5}$). Top-$r$ features ($r = 0.3$) are encoded by a two-layer RGCN (384 hidden units), with remaining features processed by a two-layer MLP. Models are trained for up to 10 epochs (batch size 32, early stopping patience 3) and evaluated by averaging results over three seeds on A100 GPUs. The LLM-based relevance filtering stage is a one-time preprocessing step that can be executed offline in batches. Model training on a single NVIDIA A100 40GB GPU takes approximately 1.5 hours in our main setting, indicating that the additional computational overhead is mainly introduced during preprocessing rather than iterative training.

## 6 Experimental Results

**Main Results on TwiUSD.** Table 4 reports the main stance detection results on our TwiUSD dataset. All models are evaluated under the same-target setting, with results averaged over three random seeds to ensure robustness.

From these results, we draw the following key observations. First, heuristic and unsupervised methods (UUSDT, Tweets2Stance) perform overwhelmingly poorly, with $F_{avg}$ and accuracy often less than half that of supervised baselines. This gap exposes the severe limitations of prior automatic

| METHOD | Biden | | | | Trump | | | |
|---|---|---|---|---|---|---|---|---|
| | $F_{favor}$ | $F_{against}$ | $F_{avg}$ | $Acc$ | $F_{favor}$ | $F_{against}$ | $F_{avg}$ | $Acc$ |
| UUSDT | 31.63 | 38.96 | 35.29 | 36.14 | 12.15 | 64.50 | 38.32 | 48.53 |
| Tweets2Stance | 12.18 | 47.37 | 29.77 | 32.78 | 39.76 | 88.43 | 64.09 | 79.72 |
| TAN | 66.45 | 12.54 | 39.49 | 50.24 | 47.63 | 84.19 | 65.91 | 74.18 |
| CrossNet | 68.55 | 35.59 | 52.07 | 55.39 | 33.90 | 79.97 | 53.60 | 66.63 |
| BERT | 82.41 | 69.49 | 75.95 | 75.92 | 62.95 | 87.52 | 75.23 | 79.57 |
| RoBERTa | 83.72 | 74.69 | 79.21 | 77.72 | 64.66 | 88.03 | 76.34 | 79.61 |
| BERT-GCN | 65.94 | 41.91 | 53.92 | 52.39 | 30.67 | 79.42 | 55.05 | 63.40 |
| TPDG | 80.48 | 67.90 | 74.19 | 74.17 | 58.70 | 86.86 | 72.78 | 78.35 |
| JointCL | 81.12 | 70.83 | 75.97 | 74.89 | 56.21 | 86.40 | 71.31 | 77.68 |
| KEPrompt | 80.31 | 65.61 | 72.96 | 72.55 | 57.08 | 87.29 | 72.19 | 78.14 |
| Llama2-70B | 77.52 | 65.15 | 71.34 | 65.21 | 42.95 | 77.11 | 60.03 | 59.19 |
| Llama3-70B | 66.30 | 56.58 | 61.44 | 52.36 | 58.77 | 73.08 | 65.92 | 59.34 |
| GPT-3.5 | 72.36 | 58.85 | 65.61 | 60.37 | 55.63 | 84.64 | 70.13 | 72.18 |
| GPT-4 | 61.16 | 65.65 | 63.41 | 57.27 | 59.17 | 70.52 | 64.85 | 61.25 |
| COLA | 75.48 | 60.78 | 68.13 | 62.94 | 55.50 | 84.16 | 69.83 | 72.10 |
| GraphICL | 75.53 | 55.99 | 65.76 | 63.18 | 46.38 | 89.12 | 67.75 | 81.27 |
| MRFG (DeepSeek-R1) | 86.57 | 78.81 | 82.69 | 81.02 | 66.80 | 92.40 | 79.60 | 85.53 |
| MRFG (Claude-3.5) | 87.62 | 77.47 | 82.55 | 82.54 | 67.21 | 92.75 | 79.98 | 86.15 |
| **MRFG (GPT-4o)** | **88.80** | **79.57** | **84.19** | **83.04** | **69.74** | **92.80** | **81.27** | **87.47** |
| *w/o* LLM-FU | 86.58 | 76.08 | 81.33 | 80.78 | 63.67 | 92.41 | 78.04 | 85.53 |
| *w/o* $S\text{-}TFI_R$ | 88.48 | 79.29 | 83.88 | 82.96 | 67.57 | 92.80 | 80.19 | 86.20 |
| *w/o* $S\text{-}TFI_m$ | 86.39 | 77.18 | 81.78 | 81.24 | 66.67 | 91.72 | 79.20 | 87.03 |

Table 4: Main experimental results (%) on the TwiUSD dataset. Bold numbers indicate the best overall performance. Underlined numbers denote the best results among all non-MRFG baselines.

labeling approaches, and highlights the unique challenge posed by TwiUSD. Second, while traditional supervised methods (TAN, CrossNet) and advanced PLMs (BERT, RoBERTa) achieve substantial gains, they remain limited in capturing the multifaceted, context-dependent reasoning required for robust stance inference. Notably, structure-augmented models such as TPDG and JointCL do not consistently outperform their content-based counterparts, suggesting that naive or shallow integration of graph signals is insufficient for this task. Third, LLMs, despite their remarkable general language abilities, underperform compared with task-specific supervised models, underscoring the importance of targeted, fine-grained social reasoning that goes beyond generic knowledge. Finally, MRFG establishes a new state-of-the-art across all metrics and stance targets, substantially outperforming both the strongest content-based and structure-augmented models. For example, it improves over RoBERTa on Biden by +4.98 $F_{avg}$ and +5.32 $Acc$. This validates our hypothesis that explicit, relevance-aware context filtering and dual-path feature routing are crucial for accurate and robust user stance detection, especially in ambiguous or noisy social environments.

To further validate the robustness and scalability of our framework, we evaluate MRFG with different LLM backbones for relevance filtering, including DeepSeek-R1, Claude-3.5, and GPT-4o. Consistent gains across targets indicate that MRFG is robust to the choice of LLM backbone.

**Impact of Incorporating Followee Information.** To verify the effectiveness of incorporating followee information, we compare three representative baselines (CrossNet, TPDG, and GraphICL) with and without followee tweets. As shown in Figure 3, adding followee context consistently improves performance across all models, with average $F_{avg}$ gains of 3.62 points. This robust, paradigm-agnostic benefit demonstrates that social context provides essential disambiguating signals–especially for users whose own stance expression is ambiguous or implicit.

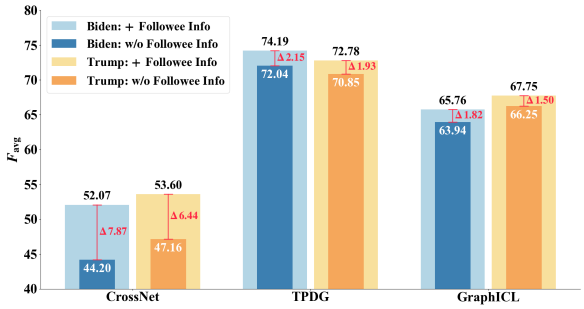


Figure 3: Performance with/without followee information.

**Ablation Study.** To evaluate the impact of each core component in MRFG, we conduct ablation

experiments with three variants: (1) **w/o** LLM-FU, which removes the relevance estimation module; (2) **w/o** $S\text{-}TFI_R$ disables the TFI ranking-and-split step and feeds all features to the RGCN branch. (3) **w/o** $S\text{-}TFI_m$ disables the TFI ranking-and-split step and feeds all features to the MLP branch. As shown in Table 4, removing the LLM-FU module consistently reduces $F_{avg}$ and accuracy, confirming the value of filtering irrelevant followee tweets. Likewise, replacing the S-TFI-based dual-path design with a single encoder also degrades performance, with the w/o $S\text{-}TFI_m$ setting performing worst. The relatively small performance drop of w/o $S\text{-}TFI_R$ suggests that, even without explicit feature splitting, the RGCN branch can still exploit part of the structural information; S-TFI mainly provides a more selective routing mechanism that helps suppress less informative features. This suggests that graph-based reasoning is especially effective for structure-sensitive features, while unfiltered processing weakens informativeness. Overall, these results validate the necessity of both selective relevance filtering and TFI-guided feature separation, and highlight the complementary strengths of graph-based and content-based reasoning within MRFG.

**Why use LLM for relevance estimation?** To evaluate our LLM-FU, we compare it to a baseline that uses *cosine similarity* between BERT embeddings of the user's content and each followee tweet, where tweets with similarity in $[0.7, 0.85)$ are deemed weakly relevant, and $[0.85, 1]$ as strongly relevant (others discarded). As Figure 4 shows, LLM-FU consistently outperforms this baseline across all feature selection ratios $r$ for both targets. This performance gap highlights the limitations of cosine similarity, which relies solely on shallow embedding similarity and cannot fully capture the nuanced semantic or pragmatic relevance between users and tweets. In contrast, LLMs offer stronger context understanding and reasoning capabilities. They assess relevance not just based on textual similarity but also by interpreting stance-related intent and implicit references. This enables more accurate filtering of useful contextual tweets, ultimately leading to better stance prediction.

**Impact of Feature Selection Ratio $r$.** We analyze the effect of the feature selection ratio $r$ used for splitting features between the GNN and MLP paths. Figure 4 presents $F_{avg}$ scores for both the LLM-based and cosine-based filtering approaches

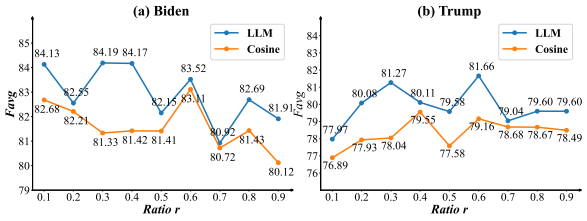


Figure 4: LLM-based and Cosine-based relevance estimation under different feature selection ratios $r$ on the TwiUSD dataset.

across varying $r$ values on the *Biden* and *Trump* subsets. Across both targets and strategies, performance generally peaks when $r$ is in the range of $0.3$–$0.4$, indicating that a moderate allocation of structural features yields the most effective balance between content and relational signals. When $r$ is too small, the model lacks sufficient structural context; when $r$ is too large, overly relying on potentially noisy graph signals can degrade performance. Both filtering strategies follow a similar trend, though with slightly different sensitivities to extreme $r$ values.

**Error Analysis.** We inspected the model's misclassifications and found two consistent failure modes. First, sparse author signal: users who have posted only one tweet provide almost no textual evidence for stance inference, and MRFG achieves only about 79% accuracy on this subset–well below the overall 85.26% attained on the full test set. These accounts seldom contain explicit stance cues, leaving both the language encoder and the graph branch without reliable signals. Second, conflicting social context: when a user follows accounts with heterogeneous or ambiguous political leanings, the graph pathway receives mixed messages that even the relevance filter cannot fully resolve. In such cases, the model often over-weights the prevailing followee opinion and fails to reconcile contradictory evidence, leading to incorrect predictions despite neutral or limited textual content. Detailed error statistics for both cases are provided in Appendix G. These findings highlight two avenues for future work: augmenting sparse user representations and introducing uncertainty-aware message passing to address conflicting followee signals.

We further examine model errors across different user types. We observe that regular users have a higher error rate than followees, suggesting that stance prediction is more challenging when users provide weaker or more implicit textual evidence.

One possible reason is that regular users more often require contextual interpretation from their social connections, whereas followees tend to express clearer stance cues in their own content.

## 7 Conclusion

In this paper, we introduce TwiUSD, a large-scale expert-annotated benchmark for political UserSD with an explicit follow graph, and show that widely used heuristic supervision can be highly noisy. To effectively leverage this data, we proposed the MRFG, which performs LLM-based relevance filtering and TFI-guided dual-path inference to jointly exploit semantic and relational cues in noisy social contexts. Extensive experiments demonstrate consistent improvements over strong baselines, highlighting the importance of relevance-aware context selection and structure-aware modeling.

## Acknowledgments

This work has been supported by the Guangdong S&T Program (2025B0101130002) and the National Key R&D Program of China (No. U25B2042).

## Limitations

Our approach relies exclusively on textual content and structural relations derived from user tweets and followee interactions. Accordingly, TwiUSD is limited to text-based information and does not include other modalities that may provide additional stance cues. In addition, our model does not incorporate external data or knowledge sources for augmentation, and instead relies solely on user-generated content without leveraging background knowledge, entity linking, or retrieval-enhanced mechanisms.

TwiUSD also exhibits notable label imbalance across stance classes (Table 3), where one category dominates for each target. Although we report macro-averaged F1 to mitigate skewed distributions during evaluation, the imbalance may still bias training and reduce performance on minority classes. Addressing severe label imbalance remains an important direction for future work.

## Ethical considerations

This work introduces TwiUSD, a benchmark dataset for political UserSD with an explicit follow graph, constructed on top of TwiBot-22, a publicly available Twitter corpus released for academic research. All data used in this study are obtained in accordance with the TwiBot-22 (Feng et al., 2022) license and comply with the platform's terms of service and privacy policies.

To annotate the dataset, we recruited 8 qualified annotators with verified domain knowledge and enforced a trial annotation phase reviewed by senior adjudicators, followed by blinded double annotation and multi-stage adjudication to ensure label reliability. Each annotator is paid $6.5 per hour (above the average local payment of similar jobs). The entire annotation process lasted 4 months, with approximately 400 total annotator-hours. During data processing and annotation, we discarded samples containing personally identifiable information when encountered. For data sharing, we will follow platform policy and the original dataset license, and only release permissible identifiers and human-annotated labels rather than redistributing raw platform content.

Because the dataset involves political content, examples shown in the paper are sampled for scientific illustration and do not represent the authors' personal viewpoints. We used LLM services, including ChatGPT and other comparable models, to assist with writing and parts of the data processing pipeline, and followed the relevant terms and policies.

TwiUSD and MRFG are intended solely for academic research and benchmark evaluation, and should not be used for real-world deployment, political profiling, targeted persuasion, surveillance, or other high-stakes decision-making applications. To support responsible release, access to the public version of the dataset will follow the original platform and source-dataset constraints, and downstream use should be limited to non-commercial academic research. Moreover, because the source data do not provide reliable demographic attributes such as age, gender, or race, we cannot conduct a comprehensive fairness evaluation across demographic groups in this work.

## References

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Guy Barel, Oren Tsur, and Dan Vilenchik. 2025. Acquired TASTE: Multimodal stance detection with textual and structural embeddings. In *Proceedings of*

*the 31st International Conference on Computational Linguistics*, pages 6492–6504, Abu Dhabi, UAE. Association for Computational Linguistics.

Zefan Cai, Baobao Chang, and Wenjuan Han. 2023. Human-in-the-loop through chain-of-thought. *arXiv preprint arXiv:2306.07932*.

Kareem Darwish, Peter Stefanov, Michaël J. Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*, pages 141–152.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2018. Topical stance detection for twitter: A two-phase lstm model using attention. In *European Conference on Information Retrieval*, pages 529–536. Springer.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3988–3994.

Doaa S. Elzanfaly, Zeyad Radwan, and Nermin Abdelhakim Othman. 2023. User stance detection and prediction considering most frequent interactions. In *Artificial Intelligence and Online Engineering*, pages 421–433, Cham. Springer International Publishing.

Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, Xinshun Feng, Qingyue Zhang, Hongrui Wang, Yuhan Liu, Yuyang Bai, Heng Wang, Zijian Cai, Yanbo Wang, Lijing Zheng, and 3 others. 2022. Twibot-22: Towards graph-based twitter bot detection. In *Advances in Neural Information Processing Systems*, pages 35254–35269.

Margherita Gambini, Caterina Senette, Tiziano Fagni, and Maurizio Tesconi. 2023. From tweets to stance: An unsupervised framework for user stance detection on twitter. In *Discovery Science*, pages 96–110, Cham. Springer Nature Switzerland.

Joseph Gatto, Omar Sharif, and Sarah Preum. 2023. Chain-of-thought embeddings for stance detection on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4154–4161, Singapore. Association for Computational Linguistics.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 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 (Long Papers)*, volume 1.

Hu Huang, Bowen Zhang, Yangyang Li, Baoquan Zhang, Yuxi Sun, Chuyao Luo, and Cheng Peng. 2023. Knowledge-enhanced prompt-tuning for stance detection. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–20.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735, Online. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the international AAAI conference on web and social media*, volume 18, pages 891–903.

Chen Li, Hao Peng, Jianxin Li, Lichao Sun, Lingjuan Lyu, Lihong Wang, Philip S. Yu, and Lifang He. 2022. Joint stance and rumor detection in hierarchical heterogeneous graph. *IEEE Trans. Neural Networks Learn. Syst.*, 33(6):2530–2542.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 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*.

Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. 2021. Target-adaptive graph for cross-target stance detection. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23*, pages 3453–3464.

Bin Liang, Ang Li, Jingqian Zhao, Lin Gui, Min Yang, Yue Yu, Kam-Fai Wong, and Ruifeng Xu. 2024. Multi-modal stance detection: New datasets and model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12373–12387, Bangkok, Thailand. Association for Computational Linguistics.

Bin Liang, Qinlin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 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)*, volume 1, pages 81–91. Association for Computational Linguistics.

Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3152–3157.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Isabelle Lorge, Li Zhang, Xiaowen Dong, and Janet Pierrehumbert. 2024. STEntConv: Predicting disagreement between Reddit users with stance detection and a signed graph convolutional network. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15273–15284, Torino, Italia. ELRA and ICCL.

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

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 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*, pages 31–41.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Technol.*, 17(3).

Ron Korenblum Pick, Vladyslav Kozhukhov, Dan Vilenchik, and Oren Tsur. 2022. Stem: unsupervised structural embedding for stance detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11174–11182.

Peyman Rostami, Vahid Rahimzadeh, Ali Adibi, and Azadeh Shakery. 2025. PolitiSky24: U.S. political bluesky dataset with user stance labels. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Younes Samih and Kareem Darwish. 2021. A few topical tweets are enough for effective user stance detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2637–2646, Online. Association for Computational Linguistics.

Pedro Semcovici and Ivandré Paraboni. 2025. Social media user stance detection without stance text. In *Proceedings of the 21st Brazilian Symposium on Information Systems, SBSI 2025, Recife, Brazil, May 19-23, 2025*, pages 1–8. SBC.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Yuanfu Sun, Zhengnan Ma, Yi Fang, Jing Ma, and Qiaoyu Tan. 2025. GraphICL: Unlocking graph learning potential in LLMs through structured prompt design. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2440–2459, Albuquerque, New Mexico. Association for Computational Linguistics.

Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783.

Bowen Zhang, Genan Dai, Fuqiang Niu, Nan Yin, Xiaomao Fan, Senzhang Wang, Xiaochun Cao, and Hu Huang. 2024a. A survey of stance detection on social media: New directions and perspectives. *arXiv preprint arXiv:2409.15690*.

Chong Zhang, Zhenkun Zhou, Xingyu Peng, and Ke Xu. 2024b. Doubleh: Twitter user stance detection via bipartite graph neural networks. In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media*, pages 1766–1778.

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

## A TwiBot-22 Information

TwiBot-22 (Feng et al., 2022) is a large-scale, high-quality annotated graph dataset, containing one million account information and tens of millions of tweets, as well as hundreds of millions of edges used to describe various types of edges. In addition, this dataset contains multiple social network relationships of users (such as following, like, forwarding, comments, etc.). This rich graph structure information can be used to analyze the relationships between users and the user's stance and tendency in social networks. The specific dataset details are shown in Table 5. The dataset is large in scale and contains rich social network information, so it can be used to construct a social media account-level stance detection dataset and used for account-level stance detection tasks.

## B TwiUSD Keywords

To identify tweets relevant to the U.S. presidential election, we followed prior work (Kawintiranon and Singh, 2021; Liang et al., 2024) and selected a set of politically salient yet ideologically neutral hashtags. This approach avoids introducing bias during data collection while preserving political relevance.

| Item | Value | Item | Value | Item | Value |
|---|---|---|---|---|---|
| entity type | 4 | post | 88,217,457 | following | 2,626,979 |
| relation type | 14 | pin | 347,131 | follower | 1,116,655 |
| user | 1,000,000 | like | 595,794 | contain | 1,998,788 |
| hashtag | 5,146,289 | mention | 4,759,388 | discuss | 66,000,633 |
| list | 21,870 | retweet | 1,580,643 | bot | 139,943 |
| tweet | 88,217,457 | quote | 289,476 | human | 860,057 |
| user metadata | 17 | reply | 1,114,980 | entity | 92,932,326 |
| hashtag metadata | 2 | own | 21,870 | relation | 170,185,937 |
| list metadata | 8 | member | 1,022,587 | max degree | 270,344 |
| tweet metadata | 20 | follow | 493,556 | verified user | 95,398 |

Table 5: Statistics of TwiBot-22.

| Method | Target | Predict Stance | Samples and Proportion of True Labels | | | | | | | Error Predict | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | against | % | favor | % | none | % | Total | Count | % |
| **UUSDT** | **Trump** | **favor** | 128 | 79.50 | 26 | 16.15 | 7 | 4.35 | 161 | 135 | 83.85 |
| | | **against** | 176 | 95.66 | 4 | 2.17 | 4 | 2.17 | 184 | 8 | 4.34 |
| **DoubleH** | **Biden** | **favor** | 18 | 6.41 | 239 | 85.05 | 24 | 8.54 | 281 | 42 | 14.95 |
| | | **against** | 276 | 15.36 | 1,198 | 66.67 | 323 | 17.97 | 1,797 | 1,521 | 84.64 |
| | **Trump** | **favor** | 2,602 | 91.17 | 191 | 6.69 | 61 | 2.14 | 2,854 | 2,663 | 93.31 |
| | | **against** | 199 | 96.13 | 3 | 1.45 | 5 | 2.42 | 207 | 8 | 3.86 |

Table 6: Analysis of hashtag-based stance prediction from unsupervised methods. We compare UUSDT (Darwish et al., 2020) and DoubleH (Zhang et al., 2024b), which infer user stance from hashtag occurrences in tweets. The results show the distribution of true labels per predicted stance and the corresponding error rates.

We retained tweets that explicitly mentioned either "*Biden*" or "*Trump*" and contained at least one of the following hashtags:

#Vote, #Debates2020, #USElection2020, #Presiden-tialDebate2020, #2020Election, #voter-suppression, #GetOuttheVote, #2020elections, #Trump2020LandSlide, #TrumpCrimeFamily, #DonaldTrump, #Republican, #BidenForPresident, #SleepyJoe, #JoeBiden, #Democrats

This keyword set was used to construct the initial tweet pool before user-level filtering and stance annotation.

## C Annotation Guidelines

Annotators assign one of three stance labels: *favor*, *against*, or *none*. A user is labeled as *favor* if the available evidence indicates support for the target, *against* if the available evidence indicates opposition, and *none* if no clear stance can be reliably identified. For followees and isolated users, annotators determine the stance label based on the user's tweets and profile description. For regular users, annotators first examine the user's own tweets and profile information. When the user's stance is not clearly expressed from their own content, annotators further consider contextual information from their followees as supplementary evidence. If the available evidence is weak, ambiguous, or conflicting, annotators assign the label *none* unless a clear stance can be supported by the combined context. All instances are independently labeled by two annotators, and disagreements are resolved through additional adjudication.

As a representative example, when a regular user posts election-related content without expressing a clear attitude toward the target, annotators first examine the user's own tweets and profile. If the stance remains unclear, they further consider contextual information from the user's followees as supplementary evidence. If no clear stance can still be supported, the user is labeled as *none*.

## D Comparison with unsupervised datasets

Existing user-level stance detection methods often rely on unsupervised strategies, using stance-indicative hashtags to construct weakly labeled datasets.For example, UUSDT (Darwish et al., 2020) labeled users who used the hashtag #MAGA as pro Trump and users who used any of the hashtags #resist, #resistance, #impeachTrump, #theResistance, or #neverTrump as anti.

| Role | Input |
|---|---|
| Instruction | Your task is to analyze the degree to which tweets posted by followees are relevant to the content posted by the original user. The degree of relevance between a followee's tweet and a user's content is represented by a score: A score of 1 means no relevance, A score of 2 means weak relevance, A score of 3 means strong relevance. Evaluate the relevance of each followee tweet in relation to the user content. |
| User's Tweets | user_tweet1; user_tweet2; user_tweet3; user_tweet4. |
| Followees' Tweets | followee_tweet1; followee_tweet2; followee_tweet3; followee_tweet4. |
| Output format | Use the score to indicate the degree to which each followee tweet is related to the user's tweets. Output the followee tweet identifier and its corresponding score in the format "(followee_tweet#:score)" following the given order. |
| Output data | (followee_tweet1:1), (followee_tweet2:2), (followee_tweet3:1), (followee_tweet4:2) |

Table 7: Input structure of the LLM and the corresponding output examples.

Besides,DoubleH (Zhang et al., 2024b) used labeled users who used the hashtag #trumpvirus, #bidenharris, #biden2020, #votebidenharris2020, #resistant, #votebluetoendthisnightmare, #iamtulsi, #trumpisanationaldisgrace, #dumptrump, #trumpmeltdown, #voteblue as pro Biden, and #gop, #kag, #hunterbiden, #maga, #trump2020, #tcot, #burisma, #votered, #republican, #americafirst, #sleepyjoe, #mc2020, #fourmoreyears, #bidencrimefamily, #bidenriots as pro Trump.

To assess the generalizability of these hashtag-based heuristics, we applied both UUSDT and DoubleH's labeling strategies to the TwiUSD dataset. The results, shown in Table 6, indicate that these rules do not transfer well. In particular, users labeled by hashtag presence alone often hold the opposite stance, with error rates reaching up to 93.31%. For example, when DoubleH predicted a user as "favor Trump," 91.17% of those users were actually "against." Similarly, UUSDT misclassified 83.85% of users predicted as pro-Trump.

These findings suggest that stance-indicative hashtags, while useful in certain contexts, may be highly topic- and time-sensitive, and often fail to reflect the user's actual stance in broader datasets. This highlights the need for more robust, content- and structure-aware methods like ours for reliable user-level stance detection.

## E LLM-FU Prompt

We use an LLM-based filter to determine the correlation between tweets, which are used to select followee tweets highly relevant to the target user.Specifically, we divide the tweet relevance into three categories: unrelated, weak correlation and strong correlation, which are expressed by scores 1, 2, and 3 respectively. For each tweet of the followees, a score is output to indicate the relevance to the target user's tweets. In our implementation, we use the GPT-4o model (via OpenAI API) as the underlying LLM for relevance assessment due to its strong contextual understanding and reasoning capabilities. The input format and representative output examples for the GPT-4o-based filtering process are shown in Table 7.

## F Baseline Methods

**UserSD Methods.** **UUSDT** (Darwish et al., 2020) performs unsupervised stance detection by extracting user features and applying dimensionality reduction and clustering techniques. **Tweets2Stance** (Gambini et al., 2023) leverages LLMs to identify tweet topics and estimate relevance, from which tweet-level stances are inferred and aggregated into user-level predictions.

**Supervised Neural Models.** **TAN** (Du et al., 2017) models the relationship between target and tweet using attention to learn target-specific stance representations. **CrossNet** (Xu et al., 2018) performs cross-target stance detection by encoding tweets and targets via BiLSTM and applying aspect-level attention to extract stance-bearing components.

**Fine-Tuned Pre-trained Models.** **BERT** (Devlin et al., 2019) is fine-tuned directly on stance-labeled training data to serve as a strong contextualized baseline. **RoBERTa** (Liu et al., 2019) improves upon BERT by leveraging more training data, larger batch sizes, and optimized pre-training strategies. **BERT-GCN** (Liu et al., 2021) enhances BERT with commonsense and structural knowledge using a graph convolutional layer for better generalization in low-resource settings. **TPDG** (Liang et al., 2021) dynamically disentangles target-dependent and target-independent components in stance expressions using structured modeling. **JointCL** (Liang et al., 2022) combines stance contrastive learning with target-aware graph contrastive learning to improve generalization to

| Case | User's data | Followee's data | Predicted_label | True_label |
|---|---|---|---|---|
| 1 | tweet_1: House Democrats watering down Biden's tax proposals: which to be fair, may be needed to get them passed... | user1_tweet1:10,000+ excess deaths a day caused by UK, Switzerland @EU_Commission & Germany blocking the #TRIPSWaiver.<br>stance:[user1:'none'] | none | against |
| 2 | tweet_1: "Putin may encircle Kiev with tanks, but he'll never gain the hearts and souls of the Iranian people." #JoeBiden #UkraineWar #UkraineKrieg. | user1_tweet1: US Congressman asks President @JoeBiden to reject Pakistan's effort to install 'jihadist' envoy<br>Reported by: Sidhant Sibal (@sidh...)<br>stance:[user1:'against'] | against | none |
| 3 | tweet_1: BREAKING: #SCOTUS will hear Biden v. Texas. This case will determine if the will of millions of voters who rejected Trump,<br>tweet_2: #DayWithoutImmigrants is trending at No.10 in US! #Immigration #WhiteHouse #ImmigrationReformNow #JoeBiden #ValentinesDay... | user1_tweet1: Pres.-elect @JoeBiden nominated Merrick Garland as AG of @TheJusticeDept and other civil rights experts to DOJ. We look forward to working with the Biden/Harris DOJ to restore and enforce civil rights of America's vulnerable farmworkers.<br>user1_tweet2: President-elect @JoeBiden has nominated Tom Vilsack as the next @USDA Secretary. We are hopeful Vilsack will make the...<br>user2_tweet1: If a so-called parole status is all that can get to President Biden's desk, then Democrats need to stop playing games wi...<br>user2_tweet2: HAPPENING NOW: TPS Families & Community Faith Leaders in front of the White House for an action denouncing @joebiden, Dem...<br>...<br>stance:[user1:'favor', user2:'against'] | favor | none |
| 4 | tweet_1: BREAKING NEWS: Trump's Steve Bannon says that Americans should be supporting Russia instead of President Biden. RT IF... | user1_tweet1: If Democrats are worried about Biden looking weak, this is the kind of thing that actually makes him look weak because eve...<br>user1_tweet2: Biden and Democrats getting the war with Russia they been begging for and trying to start since 2016. What's wild is Ukrain...<br>user2_tweet1: Both Republicans & Democrats blast #facialrecognition's threat to civil rights, so where's the ban?...<br>user3_tweet1: I am profoundly honored to be the Principal Deputy Press Secretary for @JoeBiden. I am especially thrilled to work alongs...<br>user3_tweet2: JUST IN: President-elect Joe Biden has officially been certified the winner of Georgia's 16 electoral votes following a statewide...<br>...<br>stance:[user1:'against', user2:'none', user3:'favor'] | favor | against |

Table 8: Example of Error Case of Target "Biden".

unseen targets. **KEPrompt** (Huang et al., 2023) uses an automatic verbalizer to define label words and reformulates stance detection as a prompt-based classification task.

**LLM-Based Prompting Methods.** We implement LLM-based prompting baselines following the prompting strategy proposed by Gatto et al. (2023). We use **LLaMA 2-70B**[2], **LLaMA 3-70B**[3], and **ChatGPT** (GPT-3.5 and GPT-4)[4] to generate zero-shot or few-shot stance predictions via in-context learning. **COLA** (Lan et al., 2024) introduces collaborative role-aware agents to model stance from multiple perspectives through LLM-based reasoning. **GraphICL** (Sun et al., 2025) incorporates graph-structured prompts into LLM input to enable relational stance reasoning over structured user-content graphs.

[2]https://huggingface.co/meta-llama/Llama-2-70b-chat-hf
[3]https://huggingface.co/meta-llama/Meta-Llama-3-70B
[4]https://platform.openai.com/docs/overview

## G Error Case

To better understand the limitations of our model, we manually analyze several misclassified cases from the test set. We observe two major sources of error.

First, users with only a single tweet tend to be more difficult to classify. Specifically, we find that the prediction accuracy for users with only one tweet is 79%, which is noticeably lower than the overall model accuracy of 85.26%. These users often lack explicit stance cues, making both textual and structural signals insufficient for reliable inference (Table 8 Case 1).

Second, users whose followees express conflicting or ambiguous stance signals are more difficult to classify. As shown in Case 3 and Case 4, the model sometimes over-relies on the dominant stance among followees, failing to reconcile contradictory signals—particularly when the user's

own tweets are neutral or implicitly opinionated. In such scenarios, even LLM-based relevance filtering may preserve noisy or inconsistent inputs, leading to misclassification.

Table 8 presents several illustrative error cases. In Case 1, the user has a single tweet implying skepticism toward Biden's tax proposal, noting that it may need revision to pass. Although this expresses indirect opposition, the limited content offers insufficient cues, leading the model to predict a neutral stance. In Case 2, the user's tweet focuses on geopolitical issues unrelated to Biden, and no clear stance is expressed. However, all followees oppose Biden, which biases the model toward an incorrect "against" label. In Cases 3 and 4, while the users' own tweets reveal their stance, the followees hold mixed and conflicting opinions. This inconsistency confuses the model's aggregation mechanism and leads to incorrect final predictions.

These findings suggest that future improvements should focus on disambiguating stance under sparse supervision and filtering conflicting social signals more robustly.