LLM-Empowered Stance Detection Based on the Expanded Stance Triangle Framework

Anonymous ACL submission

Abstract

Stance detection on social media refers to the task of predicting the attitudes (favor, against or neutral) of documents toward a specified target. Recently, there has been an increasing interest in employing Large Language Models (LLMs) to detect stance, demonstrating impressive performance without relying on labeled data. However, these models tend to be conservative and thus often classify documents as neutral, since users typically express their attitudes implicitly through other objects, rather than directly mentioning the target. In this paper, we present LLMTriStance, a novel LLM-empowered approach for stance detection in social media, integrating the expanded stance triangle framework from linguistics. Leveraging pseudo labels generated by LLMs and nouns extracted via syntactic tools, we apply pattern mining to actively discover the common objects associated with specific evaluations when expressing attitudes toward a target. These stance expression rules are then purified through conflict identification and resolving, enabling the generation of valuable prompts for LLMs across various cases. This process forms an iterative cycle, leading to progressive improvements in accuracy. Experimental results on multiple stance detection datasets show that our model outperforms state-of-the-art methods, providing interpretable object-attitude pairs as rationales for its predictions.

1 Introduction

With the widespread use of social media, it is significant to understand the public's perception of various social events. Stance detection is the task of automatically predicting the attitudes of documents toward a specified target (Wen and Hauptmann, 2023a), classifying them as favor, against and neutral. Early supervised methods (Mohammad et al., 2017; Dey et al., 2018) suffer from the lack of plentiful training data, as each target requires respective annotations. To this end, many zero-shot methods (Liang et al., 2021; Zou et al., 2022) were proposed for cross-target stance detection via transfer learning, but the discrepancy among targets severely limits their performances.

In the era of Large Language Models (LLMs), researchers have begun to leverage the strong understanding and generative capabilities of prompting LLMs to overcome the labeling issue of stance detection in an unsupervised manner, and achieve superior performance to supervised baselines on specific targets (Zhang et al., 2022; Cruickshank and Ng, 2023). Nevertheless, it exhibits limited accuracy on other targets, which can be explained by a fundamental mismatch between the inherent mechanism of LLMs and the requirements of stance detection tasks. LLMs are typically trained to maintain neutrality in order to avoid biases (Li and Zhang, 2024), which naturally leads them to classify a document as the incorrect neutral stance, especially on controversial topics such as atheism or feminist movements. Moreover, on social media, the stance is often not expressed with an explicitly mentioned target, but is instead conveyed implicitly through another object (Liu et al., 2023). This makes it particularly challenging for LLMs to accurately infer stance on sensitive topics in the absence of direct contextual cues.

Many studies were devoted to enhance the performance of LLMs specially for the stance detection task, by expanding the contexts and providing additional information (Cruickshank and Ng, 2023; Gatto et al., 2023; Liu et al., 2023; Li et al., 2023). However, these approaches neglect the intrinsic complexity of the stance detection task. It is deeply grounded in linguistic and discourse theories (Biber and Finegan, 1988; Du Bois and Kärkkäinen, 2012). In particular, the stance triangle framework (Du Bois, 2008) and its extensions (Liu et al., 2023) provide a comprehensive understanding of the essential elements involved in stance-taking, such as the stance holder, the explicit

object and the implicit target. These concepts have been used to enrich dataset annotations, but have not been effectively utilized in the computational detection approaches.

In this paper, we propose a novel approach named LLM-empowered Triangle-based Stance detection (LLMTriStance), building upon the extended stance triangle framework (Liu et al., 2023) to enhance both the accuracy and interpretability of stance detection. Our method leverages pseudo labels generated by LLMs and nouns extracted using syntactic tools to identify stance-representative objects, and mines frequent object-evaluation pairs to construct the expression rule of each stance. Then, intra-stance and inter-stance conflicts within these pairs are detected, enabling the identification of *stance-indicative objects* and *contradictory objects*. This process purifies the rules and further produces *aligned objects* (reflecting same attitudes as the specified target) and *opposite objects* (reflecting contrasting attitudes), which are used to generate informative prompts, helping the LLM adapt to different cases and refine its predictions. Within this LLM-empowered paradigm, initially inaccurate pseudo labels and imprecise stance expression rules mutually enhance each other through an iterative process, progressively improving the model's performance. Experimental results on multiple stance detection datasets show that LLMTriStance achieves superior performance compared to state-of-the-art methods, while also has the ability of providing interpretable object-attitude pairs as rationales for its predictions.

In summary, the main contributions include:

(1) A novel approach of integrating the extended stance triangle framework from linguistics with prompting LLMs is presented to solve the task of stance detection. To the best of our knowledge, it is the first successful interdisciplinary work that applies this theoretical framework to enhance both the accuracy and interpretability of computational stance detection in an unsupervised manner.

(2) Systematic methods are designed to identify common objects and their associated attitudes as stance expression rules, while also uncovering noteworthy objects through detecting various types of conflicts. This approach facilitates more robust and context-aware predictions by iteratively refining the LLM's responses through crafted prompts.

(3) Extensive experiments are conducted to show that our model outperforms state-of-the-art methods in both accuracy and transparency, with interpretable object-attitude pairs for the specified target as prediction rationales.

2 Related Work

2.1 Computational Stance Detection

Early studies on stance detection focused on various supervised machine learning models, including rule-based methods (Bøhler et al., 2016), feature-based methods (Tutek et al., 2016; Mohammad et al., 2017) and supervised deep learning methods (Wei et al., 2016; Zarrella and Marsh, 2016; Dey et al., 2018). However, since stance labels are target-specific, it is difficult to prepare labeled data in advance for different targets in practice. Additionally, the annotation process is expensive and time-consuming for domain experts.

For this reason, weakly supervised approaches (Ebrahimi et al., 2016) were proposed because they only require a small set of seed words and hashtags for each stance. This is much cheaper than labeling documents, but only favor and against stances can be accurately detected (Wei et al., 2019), as it is unlikely to represent the neutral stance with suitable seeds. Recently, zero-shot stance detection has emerged via transfer learning (Allaway and McKeown, 2020; Zhao et al., 2023; Liu et al., 2021; Liang et al., 2022a; Wen and Hauptmann, 2023b), building the connections between labeled target data and unseen target data. Nevertheless, the diverse scenarios of the task severely constrain the transferablity among different targets and datasets, which leads to sub-optimal performances compared to fully supervised methods.

2.2 LLMs for Stance Detection

As the advent of LLMs for natural language understanding, several work employed LLMs on the stance detection task. Zhang *et al.* (2022) proposed ChatGPT-based direct question-answering (DQA) model without labeled data and achieved better accuracy than supervised models. However, ChatGPT is a closed model with invisible training datasets, so the potential contamination of data makes the evaluation unreliable (Aiyappa et al., 2023). To avoid that, Cruickshank and Ng (2023) adopted open-sourced T5-based LLMs to detect stances. They found that using LLMs with appropriate instruction prompts can improve perfor-

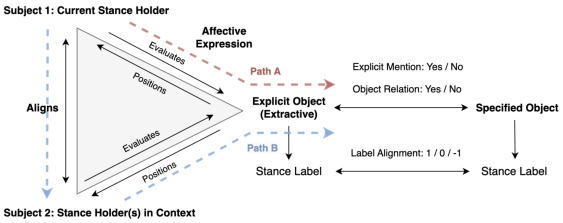

Figure 1: Expanded stance triangle framework.

mance effectively, but the results remain unsatisfactory when handling implicit stance expressions, especially for the frequent misidentification of neutral stances due to the conservative nature of LLMs.

Following the CoT prompting models which gained remarkable performances on complex task reasoning (Fei et al., 2023), many studies explore to utilize CoT techniques and multiple LLM-based agents to realize zero-shot stance detection on social media (Zhang et al., 2023; Gatto et al., 2023; Taranukhin et al., 2024; Lan et al., 2024). However, these methods employ the same prompt for different contexts, which results in unstable performance and undermines the generalization of the model facing various kinds of targets.

2.3 Linguistic Theoretical Frameworks for Stance Detection

As a natural language processing task, stance detection has been deeply affected by several theoretical foundations and frameworks. From a linguistic perspective, prior researches (Biber and Finegan, 1988; Du Bois and Kärkkäinen, 2012) have analyzed the stance expression through lexical patterns, syntactic constructions, and affective expressions. One influential framework is the Stance Triangle (Du Bois, 2008), which models stance-taking as a dynamic interaction between the stance holder, the object and other participants, providing a foundation for analyzing stances in both face-to-face conversations and online texts. Building on this, the Expanded Stance Triangle Framework (Liu et al., 2023) further characterizes the relationship between explicit and implicit objects, enabling more robust analysis of indirect references and implicit targets for social media texts. Although utilized to enrich the annotations of the training dataset, which enhances the out-of-domain cross-target performance, the essential ideas have not been integrated into the computational detection model itself. As a result, the accuracy improvement is limited at the cost of manual data preparation.

3 Expanded Stance Triangle Framework

The Stance Triangle (Du Bois, 2008) is a foundational framework for understanding stance-taking in communicative interactions. It consists of three key components: the current stance holder (Subject 1), the object of the stance (Object), and other stance holders in context (Subject 2). The framework also captures the dialogic nature of stance-taking through three expression acts between these components: evaluation, positioning, and alignment. What is more, an important challenge of stance detection on social media is stated in the triangle: **the author may express his attitude toward a specified target through an indirect reference in another document written by someone else (Path B)**. For instance, if a sentence from one chapter of the Bible is quoted, it can be deduced that the author is against the target "Atheism".

On this basis, Liu et al. (2023) proposed the Expanded Stance Triangle Framework, introducing two frequently occurring but previsously overlooked concepts: explicit objects and specified objects (implicit targets). That leads to the second challenge: **the author may express his attitude toward a specified target by implicitly mentioning another explicit object (latter part of Paths A&B)**. For example, when someone says "Are we so desperate in this country to seriously consider a 60+ woman as president?", the explicit object is "woman", while the implicit object could be "Hillary Clinton" as the actual target of the expression, assuming the context suggests a connection between woman and Hillary Clinton.

Moreover, this expanded framework also defines the relationships between explicit and implicit objects to assist stance reasoning through object relation and label alignment. The former determines whether the stance label of an implicit object (target) can be inferred from the extractive explicit object, while the latter further clarifies whether the attitudes on explicit and implicit objects are aligned, opposite or unrelated. These two relations together serve as crucial clues for detecting the hidden stance toward the specified target. For the example above, the disapproving attitude on "woman" implies against toward Hillary Clinton.

Therefore, to accommodate the complexity of social media texts and reduce reliance on extensive human annotations, a stance detection approach requires to go beyond focusing solely on the direct target. Instead, it should actively identify and con-

cretize the two implicit paths related to a specific target as prior knowledge, enabling more accurate and robust stance classification.

4 Proposed Approach

In this section, we first formulate the problem of stance detection and provide an overview of our proposed approach, LLMTriStance. We then elaborate the three core modules.

4.1 Problem Formulation

Given a corpus of unlabeled documents D and a target t, the stance detection task aims to assign a stance label $y_d \in Y$ to each document $d \in D$, where the stance label set Y consists of three categories: favor (F), against (A) and neutral (N).

4.2 Approach Overview

To better leverage the expanded stance triangle framework and effectively reason about implicit stance expressions, we at first design a triangle component identification module to recognize the key concepts within the stance triangle by the LLM for each document. Next, a triangle-based mining module is devised to discover object-evaluation pairs as stance expression rules, followed by detecting intra-stance and cross-stance conflicts to screen the rules and obtain different types of discriminative objects as clues. Finally, a triangle-based reflection module is developed to handle various scenarios. It constructs reflection prompts based on the object matching, and then inputs these prompts along with the original document into the LLM to guide stance re-assessment. These three modules collectively assist the LLM to refine its predictions in an iterative manner. The pseudo-code of the whole model and the prompt design are provided in Appendices A and B respectively.

4.3 Triangle Component Identification Module

Although directly attaining satisfactory accuracy is challenging, the LLM can generate an initial pseudo label $\hat{y}_d^{(0)} \in Y$ for each document d as the starting point for the detection model:

$$\hat{y}_d^{(0)} = \text{LLM}(\text{Label-prompt}(d)) \quad (1)$$

The set of documents currently assigned to the stance y can be denoted as $D^y = \{\, d \mid \hat{y}_d = y \,\}$. Besides, as illustrated in the expanded stance triangle framework, there may exist other stance holders in the text. The LLM also has the ability to generate the description of them, denoted as h_d:

$$h_d = \text{LLM}(\text{Holder-prompt}(d)) \quad (2)$$

Here, the stance holders encompass two types: the current stance holder (Subject 1) and the stance holders in context (Subject 2). In the case of Subject 1, the LLM explains and outputs the origin of the statement based on its content, e.g. *The statement appears to come from a social media post expressing faith or belief in God*. While for Subject 2, the LLM identifies and specifies the actual source of the statement, e.g., *The statement comes from the Bible, Matthew 23:12*. By uncovering indirect references to address the first challenge, this text extension lays the foundation for tackling the second challenge: locating implicit mentions.

Specifically, we extract nouns based on Part-of-Speech (POS) tagging from both the document d and the stance holder description h_d, and combine them to form the set $\tilde{O}_d$ of candidate objects for document d. However, some of these objects may appear incidentally and are not always relevant to the specified target, so should not be regarded as stable objects used to help stance detection. To select category-related objects that are more likely to reflect stance, we mine frequent items from the noun sets of documents labeled with either favor or against stances, denoted as a set of object words W, and the final set of objects O_d for each document d is then filtered based on W as follows:

$$\begin{aligned} W &= \{\, w \mid sup(w, D^{\mathsf{F}} \cup D^{\mathsf{A}}) \geq \epsilon_1 \} \\ O_d &= \{\, o \mid o \in \tilde{O}_d \cap W \} \end{aligned} \quad (3)$$

where ϵ_1 is the minimum support threshold for this first-step mining among individual objects.

Next, the evaluation $e_{d,i}$ representing the attitude on each object $o_{d,i} \in O_d$ $(i = 1..|O_d|)$ in the document is generated by the LLM as follows:

$$e_{d,i} = \text{LLM}(\text{Evaluation-prompt}(d, o_{d,i})) \quad (4)$$

Here, same as the overall attitude toward the specified target, each $e_{d,i}$ can take one of three categorical values: favor (F), against (A) or neutral (N). We treat each document as a transaction and each objet-evaluation pair within it as an item of the transaction. To construct a transaction database T^y for each stance $y \in Y$, we aggregate the object-evaluation pairs based on the pseudo labels:

$$\begin{aligned} T^y &= \{\, T_d \mid d \in D^y \,\}, \text{ where} \\ T_d &= \{(o_{d,i}, e_{d,i}) \mid i = 1..|O_d|\} \end{aligned} \quad (5)$$

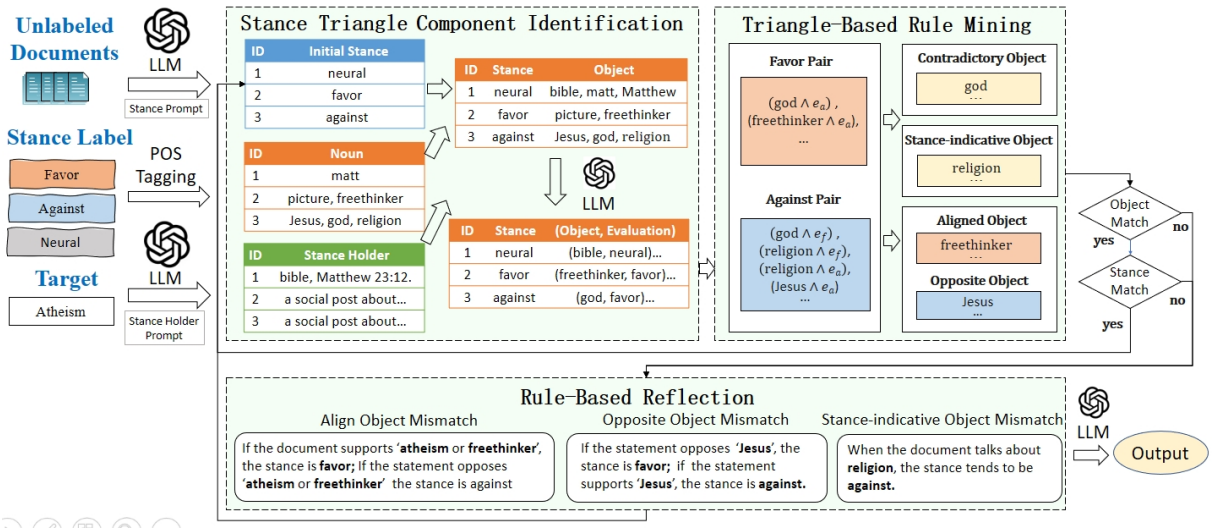

Figure 2: Overview of our approach LLMTriStance.

This stance transaction database serves as the foundation for the second-step mining of representative rules at the pair level, realized in the next module.

4.4 Triangle-Based Rule Mining Module

In order to discover the representative rules to express favor/against stances respectively, we mine frequent pairs of the form $p = (o, e)$ appearing in T^y for each stance $y \in \{\mathtt{F}, \mathtt{A}\}$. These pairs compose the candidate rule for each stance as follows.

$$\tilde{\mathcal{R}}^y = \{\, p \mid sup(p, T^y) \geq \epsilon_2 \} \tag{6}$$

where ϵ_2 is the minimum support threshold for pairs in the stance expression rules.

However, due to the inaccuracy in pseudo labels and the divergence of LLMs, the mined rules may contain conflicting pairs. We identify and resolve two main types of conflicts with strategies below:

- **Intra-stance conflicts**: Within the same stance, an object o may be associated with both favor and against evaluations. For example, in the rule of favor stance $\tilde{\mathcal{R}}^{\mathtt{F}}$, pairs $p_1 = (o, \mathtt{F})$ and $p_2 = (o, \mathtt{A})$ might appear simultaneously. This conflict suggests the object o itself is able to indicate the stance, i.e., no matter how the object is evaluated, a specific stance (favor here) is always conveyed toward the target. In such cases, we record these **stance-indicative objects** in S^y for each stance $y \in \{\mathtt{F}, \mathtt{A}\}$ as supplementary rules:

$$S^y = \{o \mid \exists p_1 = (o, \mathtt{F}) \in \tilde{\mathcal{R}}^y \wedge \exists p_2 = (o, \mathtt{A}) \in \tilde{\mathcal{R}}^y\} \tag{7}$$

  Additionally, the associated pairs are removed from the rule of the corresponding stance since the evaluation is considered inactive.

- **Cross-stance conflicts**: Certain pairs may occur frequently in both favor texts and against texts. For instance, a pair $p = (o, \mathtt{F})$ might belong to both $\mathcal{R}^{\mathtt{F}}$ and $\mathcal{R}^{\mathtt{A}}$. This conflict indicates that this evaluation is ambiguous and may reflect different attitudes depending on the context, making it difficult for the LLM to distinguish between stances. Therefore, we eliminate such non-discriminative pairs denoted as $P_c = \mathcal{R}^{\mathtt{F}} \cap \mathcal{R}^{\mathtt{A}}$ from both rules.

Through the conflict resolution above, the purified rule $\mathcal{R}^y$ with only valuable pairs for the stance $y \in \{\mathtt{F}, \mathtt{A}\}$ is obtained and denoted as:

$$\mathcal{R}^y = \{\, p = (o, e) \mid p \in \tilde{\mathcal{R}}^y - P^{\mathrm{c}} \wedge o \notin S^{\mathtt{F}} \cup S^{\mathtt{A}} \,\} \tag{8}$$

Based on the purified rule, we further extract **aligned objects** O^{ali} and **opposite objects** O^{opp} as follows, which correspond to the explicit objects with label alignment described in the second challenge:

$$\begin{aligned} O^{\mathrm{ali}} &= \{o \mid (o, \mathtt{F}) \in \mathcal{R}^{\mathtt{F}} \vee (o, \mathtt{A}) \in \mathcal{R}^{\mathtt{A}}\} \\ O^{\mathrm{opp}} &= \{o \mid (o, \mathtt{A}) \in \mathcal{R}^{\mathtt{F}} \vee (o, \mathtt{F}) \in \mathcal{R}^{\mathtt{A}}\} \end{aligned} \tag{9}$$

In this way, by fully leveraging the mined stance expression rules, common expressions when talking about a specified target are uncovered, at both the object-evaluation pair level and the object level. These important knowledge summarized from all texts on the target facilitates meaningful reflections by the LLM on each of its decisions.

4.5 Rule-Based Reflection Module

With the mined results above, we propose a rule-based reflection mechanism to adaptively assessing the reasonability of the current stance assignment for each document. If deemed unreasonable, the model provides the LLM with related common expressions as inference clues to correct the result. This is fulfilled depending on the matching of document contexts and derived rules.

For each document d, we examine each object $o \in O_d$ extracted from d to determine whether it belongs to the set of align objects O^{ali}, opposite objects O^{opp}, or stance-indicative objects S^y $(y \in \{\texttt{F}, \texttt{A}\})$. If any such object exists, and the stance inferred from it by the rule is same as the pseudo label, we consider the LLM's decision is well-grounded and no reflection is needed. Otherwise, we find all matched objects but yielding the inference result inconsistent with the pseudo label, and generate corresponding reflection prompts according to the rule, which compose a complete instruction. In cases when none of the extracted objects belong to the three defined object sets, we identify the most similar object in the document to any of the aligned objects and the opposite objects respectively based on word embeddings, and combine the generated prompts accordingly. The pseudo-code of this process containing the prompt design, is detailed in Algorithm 2 and put into Appendix A due to the page limit.

Finally, the constructed reflection prompts, together with the original document, are fed into the LLM to guide the re-determination of the document's stance as follows:

$$\hat{y}'_d = \text{LLM}(\text{Reflect-prompt}(d)) \qquad (10)$$

The entire process is conducted iteratively. After the new stances of documents are predicted, the representative rules can be recalculated, and the process is repeated. During each cycle, the rules gradually incorporate new stance-aware insights based on the stance re-determination, leading to refined classification results with each iteration.

5 Experiments

5.1 Datasets

We evaluate our approach on two commonly used Twitter datasets: SemEval-2016 Task 6 (Mohammad et al., 2016) and P-Stance (Li et al., 2021). Each tweet in these datasets is associated with a target and assigned a manually annotated stance label toward the target. To pursue a fair comparison with other models, we only use the test data from those datasets designed for supervised stance detection. The dataset statistics are shown in Table 1.

Dataset	Target	Favor	Against	Neutral
SEM16	DT	148	299	260
	HC	163	565	256
	FM	268	511	170
	LA	167	544	222
	A	124	464	145
	CC	335	26	203
P-Stance	Biden	3217	4079	-
	Sanders	3551	2774	-
	Trump	3663	4290	-

Table 1: Statistics of datasets in our experiment.

SemEval-2016 (Mohammad et al., 2016)[1] consists of six targets, such as Atheism (AT), Climate Change is a real Concern (CC), Feminist Movement (FM), Hillary Clinton (HC), Legalization of Abortion (LA) and contains Donald Trump (DT).

P-Stance (Li et al., 2021)[2] focuses on the political domain and is composed of three targets: Donald Trump (Trump), Joe Biden (Biden) and Bernie Sanders (Sanders). As noted in (Li et al., 2021), documents labeled as "None" exhibit low annotation consistency, so following prior work, we exclude these documents from our analysis.

5.2 Baselines

We compare our model with state-of-the-art methods in stance detection, including Bert-based method: BERT (Devlin et al., 2019) ; Graph-based methods: ASGCN (Zhang et al., 2019) and TPDG (Liang et al., 2021); adversarial learning method: TOAD (Allaway et al., 2021); contrastive learning methods: JointCL (Liang et al., 2022b); LLM-based methods: GPT-3.5 (Zhang et al., 2022), GPT-3.5+COT (Zhang et al., 2023), KASD-ChatGPT (Li et al., 2023) and COLA (Lan et al., 2024). Among them, BERT, ASGCN and TPDG are fully-supervised methods, relying on labeled training data for each target; TOAD and JointCL are zero-shot methods, trained on data from other targets and transferred to the current task without additional training; The LLM-based methods (GPT-3.5, GPT-3.5+COT and COLA) do

[1]https://alt.qcri.org/semeval2016/task6
[2]https://github.com/chuchun8/PStance

Method	SemEval-2016(%)							P-Stance(%)			
	DT	HC	FM	LA	AT	CC	Avg	Trump	Biden	Sanders	Avg
BERT	57.9	61.3	59.0	63.1	60.7	38.8	56.8	67.7	73.1	68.2	69.7
ASGCN	58.7	61.0	58.7	63.2	59.5	40.6	56.9	77.0	78.4	70.8	75.4
TPDG	63.0	73.4	67.3	74.7	64.7	42.3	64.2	76.8	78.1	71.0	75.3
JointCL	50.5	54.8	53.8	49.5	54.5	39.7	50.5	62.0	59.0	73.0	64.7
GPT-3.5	62.5	68.7	44.7	51.5	9.1	31.1	44.6	62.9	80.0	71.5	71.5
GPT-3.5+COT	63.3	70.9	47.7	53.4	13.3	34.0	47.1	63.9	81.2	73.2	72.8
KASD-ChatGPT	64.2	80.9	70.4	63.2	30.5	43.4	58.7	85.1	84.6	80.0	83.2
COLA	68.5	81.7	63.4	71.0	**70.8**	65.5	70.2	86.6	84.0	79.7	83.4
Qwen2.5-14B	69.7	84.3	73.8	62.4	53.0	67.5	69.6	80.1	86.2	79.0	81.8
LLMTriStance (Qwen2.5-14B)	**71.3**	84.2	75.9	67.1	66.4	69.2	72.3	81.9	85.9	79.9	82.6
Qwen2.5-32B	66.6	81.4	76.6	68.4	64.0	66.9	70.7	81.2	81.2	77.6	80.0
LLMTriStance (Qwen2.5-32B)	66.8	82.2	77.4	70.4	69.7	68.7	72.5	82.3	82.1	78.3	80.9
DeepSeek-V3	69.3	**85.8**	72.5	66.5	47.9	81.5	70.6	86.3	86.3	**82.6**	85.1
LLMTriStance (DeepSeek-V3)	69.4	84.9	**77.6**	**71.9**	66.0	**84.7**	**75.8**	**87.2**	**87.2**	**82.6**	**85.7**

Table 2: Overall results on SemEval-2016 and P-Stance datasets. The best scores are marked in bold.

not require any labeled data and leverage the reasoning capabilities of LLMs for direct inference.

5.3 Experiment Settings

We use the DeepSeek-V3 model as our LLM backbone and set the temperature to zero for ensuring replicable. Additionally, to validate the adaptability and effectiveness of our approach, we also employ two smaller open-source LLMs, Qwen2.5-14B and Qwen2.5- 32B, since relying on APIs of large models is not always feasible in real-world scenarios, especially when data privacy, latency, or cost constraints are critical concerns. Following previous work (Allaway et al., 2021; Lan et al., 2024), we calculate the average F1 score of the favor and against stances (F_{avg}) as the metric. We report both the initial results generated by prompting the three LLMs for the first time and the final results achieved after applying our proposed method.

For other baselines, we directly adopt the results from previous papers (Lan et al., 2024; Li et al., 2023). Since the results of KASD-ChatGPT on DT, AT and CC are not included, we reproduced this model using the codes provided by the authors[3].

We use spaCy[4] to implement POS tagging. For the embedding-based object similarity calculation, we choose SentenceTransformer[5] as the sentence encoder. As to the support thresholds in the two-step pattern mining, we set $\epsilon_1 = \epsilon_2 = 0.02$, while also requiring the occurrence number of each mined pair to be greater than 1 to avoid issues with too small datasets. Besides, we set the iterative number for the main model as 2. The experiments were conducted using Python 3.10.15 in a CentOS-7 server with 6 NVIDIA A40 GPUs.

5.4 Overall Results

The overall results are shown in Table 2. We can see that LLMTriStance (DeepSeek-V3) significantly outperforms the best baseline on both datasets, improving 5.6% and 2.3% over COLA respectively. For individual targets, the advantage is consistent across 8 out of 9 targets except AT, as for this target stances are expressed with scattered objects lacking strong commonality and similarity.

Notably, our model is entirely unsupervised yet still surpasses supervised and zero-shot methods, especially in handling the diversity and ambiguity of stance expressions, which can give the credit to the intrinsic knowledge embedded in LLMs. Furthermore, compared to the poor performance of GPT-3.5, GPT-3.5+COT and KASD-ChatGPT on controversial topics such as AT and CC, our approach demonstrates prominent enhancement, enabling more accurate differentiation of favor/against stances from the neutral stance. This certifies that while LLMs have the potential to reduce reliance on labeled data, for challenging and nuanced scenarios, the common stance expressions related to each document need to be actively provided as additional knowledge via prompt design to optimize LLMs' predictions.

Additionally, comparing initial results with those after our model's refinement, the average accuracy of Qwen2.5-14B, Qwen2.5-32B and DeepSeek-V3 increases on both datasets. This underscores the adaptability and robustness of our rule-based iterative approach, regardless of the choice of LLM backbones, which ranges from the latest DeepSeek-V3 to smaller open-source LLMs. How-

[3]https://github.com/HITSZ-HLT/KA-Stance-Detection
[4]https://spacy.io/
[5]https://www.sbert.net/

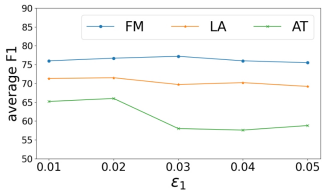

Figure 3: Varied threshold ϵ_1.

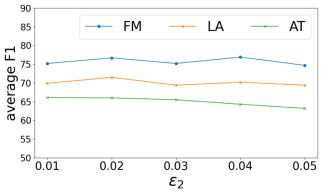

Figure 4: Varied threshold ϵ_2.

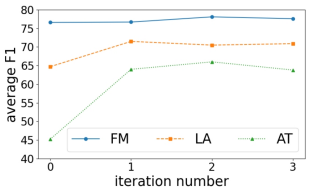

Figure 5: Varied iteration number.

Method	FM	LA	AT
LLMTriStance-opp	76.2	71.3	49.7
LLMTriStance-ind	75.4	69.1	65.1
LLMTriStance-sim	73.9	67.0	45.6
LLMTriStance (DeepSeek-V3)	**78.1**	**71.5**	**66.0**

Table 3: Results of ablation study.

ever, LLMTriStance (DeepSeek-V3) experiences a slight performance drop on the HC target. Through observations, besides a few labeling errors, this can be attributed to the context-dependent correlation between popular opposite objects and the target.

5.5 Ablation Study

In order to analyze the role of key components in our approach, we design three variants in terms of the model structure. The first two aim at the mined core sets of objects for matching, neglecting opposite objects (LLMTriStance-opp) and stance-indicative objects (LLMTriStance-ind) respectively. The third variant removes the process of identifying similar objects when precise matching does not exist (LLMTriStance-sim). The ablation experiments were conducted on three targets with relatively poor performance to highlight the impact of these components, with the results shown in Table 3.

At first, the performance decline of the first two variants confirms the importance of the two object sets in understanding implicit expressions of stances. Many objects exhibit a contrasting nature with respect to the target , such as god is contrastive to atheism and life is contrastive to legalization of abortion, while some others like freethinker, inherently imply a deterministic attitude toward atheism.

Moreover, eliminating the soft matching of objects through similarity computation severely degrades accuracy, particularly for the AT target. This suggests that the irregularity of social media texts makes perfect word matching challenging, underlining the necessity of leveraging semantic matching with the help of word embeddings.

5.6 Hyper-parameter Analysis

Support threshold ϵ_1 and ϵ_2 These two hyper-parameters in Equations 3 and 6 determine how many objects and object-evaluation pairs are retained during the two-step mining respectively. We vary the values in the range of [0.01,0.05], and the results are shown in Figures 3 and 4. We observe that the model is not sensitive to these parameters, and nearly optimal accuracy can be achieved when both are set to 0.02. striking a balance between the representativeness and coverage of the mined rules.

Iteration number We change this critical number from 0 to 3, and the results shown in Figure 5 exhibit a trend of first rising and then stabilizing after about two iterations. This leads to a consistent choice, which not only embodies the effect of mutual enhancement of LLMs and mined rules, but also maintains low costs to get good performance.

6 Conclusion

Stance detection is a difficult NLP task, as expressions toward a specific target on social media is highly diverse. Inspired by the expanded stance triangle framework from linguistics, which features the concepts of indirect references and implicit mentions through explicit objects, this paper investigates the conservativeness of LLMs in tackling the stance detection task, and explores a novel unsupervised paradigm to achieve mutual enhancement of LLMs' predictions and actively revealed reasoning rationales. By mining objective-evaluation pairs as target-specifc stance expression rules and identifying conflicts to obtain three types of representative objects, document-specific guidance is adaptively generated and provided to LLMs for building necessary correlations and facilitating reflection. The proposed approach LLMTriStance demonstrates superior accuracy over SOTA methods of various types and offers strong interpretability for understanding targets. For future work, we aim to extend this paradigm to other intricate classification tasks in NLP, such as rumor detection.

7 Limitation

Our model relies on extracting object-evaluation pairs under each stance of a specific target, which necessitates a sufficient amount of target-specific data for each target with diverse expressions of stances. However, this requirement poses a limitation: when the data for a target lacks diversity in terms of indirect references and implicit mentions, it becomes challenging to extract meaningful pairs as reasoning rules, potentially leading to suboptimal model performance.

In addition, our model currently focuses on identifying individual object-evaluation pairs but has not explored higher-order patterns such as the combinations of different pairs co-occurring in the same document. This extension is worth studying for enhancing the model's ability to capture more complicated stance expressions toward targets involving multiple factors.

References

Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2023. Can we trust the evaluation on chatgpt? *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, page 47–546.

Emily Allaway and Kathleen R. McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 8913–8931.

Emily Allaway, Malavika Srikanth, and Kathleen R. McKeown. 2021. Adversarial learning for zero-shot stance detection on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4756–4767.

Douglas Biber and Edward Finegan. 1988. Adverbial stance types in english. *Discourse processes*, 11(1):1–34.

Henrik Bøhler, Petter Asla, Erwin Marsi, and Rune Sætre. 2016. Idi$@$ntnu at semeval-2016 task 6: Detecting stance in tweets using shallow features and glove vectors for word representation. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016*, pages 445–450.

Iain J. Cruickshank and Lynnette Hui Xian Ng. 2023. Use of Large Language Models for Stance Classification. *arXiv preprints arXiv:2309.13734*.

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*, pages 4171–4186. Association for Computational Linguistics.

Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2018. Topical stance detection for twitter: A two-phase LSTM model using attention. In *Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018*, volume 10772, pages 529–536.

John W Du Bois. 2008. The stance triangle. In *Stancetaking in discourse: Subjectivity, evaluation, interaction*, pages 139–182. John Benjamins Publishing Company.

John W Du Bois and Elise Kärkkäinen. 2012. Taking a stance on emotion: Affect, sequence, and intersubjectivity in dialogic interaction. *Text & Talk*, 32(4):433–451.

Javid Ebrahimi, Dejing Dou, and Daniel Lowd. 2016. Weakly supervised tweet stance classification by relational bootstrapping. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, pages 1012–1017.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023*, pages 1171–1182.

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.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media, ICWSM 2024*, pages 891–903. AAAI Press.

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023. Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 15703–15717. Association for Computational Linguistics.

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 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Yingjie Li and Yue Zhang. 2024. Pro-woman, anti-man? identifying gender bias in stance detection. In *Findings of the Association for Computational Linguistics, ACL 2024*, pages 3229–3236. Association for Computational Linguistics.

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022a. Zero-shot stance detection via contrastive learning. In *WWW '22: The ACM Web Conference 2022*, pages 2738–2747.

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*, pages 3453–3464.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022b. JointCL: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 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*, Findings of ACL, pages 3152–3157.

Zhengyuan Liu, Yong Keong Yap, Hai Leong Chieu, and Nancy F Chen. 2023. Guiding computational stance detection with expanded stance triangle framework. *arXiv preprint arXiv:2305.19845*.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan 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 2016*, pages 31–41.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Trans. Internet Techn.*, 17(3):26:1–26:23.

Maksym Taranukhin, Vered Shwartz, and Evangelos E. Milios. 2024. Stance reasoner: Zero-shot stance detection on social media with explicit reasoning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024*, pages 15257–15272. ELRA and ICCL.

Martin Tutek, Ivan Sekulic, Paula Gombar, Ivan Paljak, Filip Culinovic, Filip Boltuzic, Mladen Karan, Domagoj Alagic, and Jan Snajder. 2016. Takelab at semeval-2016 task 6: Stance classification in tweets using a genetic algorithm based ensemble. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016*, pages 464–468.

Penghui Wei, Wenji Mao, and Guandan Chen. 2019. A topic-aware reinforced model for weakly supervised stance detection. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 7249–7256.

Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. 2016. pkudblab at semeval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016*, pages 384–388.

Haoyang Wen and Alexander G. Hauptmann. 2023a. Zero-shot and few-shot stance detection on varied topics via conditional generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023*, pages 1491–1499.

Haoyang Wen and Alexander G. Hauptmann. 2023b. Zero-shot and few-shot stance detection on varied topics via conditional generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023*, pages 1491–1499.

Guido Zarrella and Amy Marsh. 2016. MITRE at semeval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016*, pages 458–463.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would Stance Detection Techniques Evolve after the Launch of ChatGPT? *arXiv preprints arXiv:2212.14548*.

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Yangyang Li, and Liwen Jing. 2023. Investigating Chain-of-thought with ChatGPT for Stance Detection on Social Media. *arXiv preprints arXiv:2304.03087*.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4568–4578.

Chenye Zhao, Yingjie Li, and Cornelia Caragea. 2023. C-STANCE: A large dataset for chinese zero-shot stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023*, pages 13369–13385.

Jiaying Zou, Xuechen Zhao, Feng Xie, Bin Zhou, Zhong Zhang, and Lei Tian. 2022. Zero-shot stance detection via sentiment-stance contrastive learning. In *34th IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2022*, pages 251–258.

A Pseudo-code of LLMTriStance

The pseudo-code of the whole model is shown in Algorithm 1, and the sub-procedure of reflection prompt generation is presented in Algorithm 2.

Algorithm 1 LLMTriStance

Require: An unlabeled document corpus D; a specified target t;
Ensure: The stance label $\hat{y}_d$ of each document $d \in D$.

for all document $d \in D$ **do**
  Obtain initial pseudo label $\hat{y}_d^{(0)}$ and stance holder description h_d with Equations 1 and 2;
end for
for $i = 1$ to $Iter$ **do**
  for all document $d \in D$ **do**
    Extract objects and mine frequent ones O_d with Equation 3;
    Obtain evaluation $e_{d,i}$ for each object $o_{d,i} \in O_d$ with Equation 4;
  end for
  Construct stance transaction database T^y $(y \in \{\mathtt{F}, \mathtt{A}\})$ with Equation 5;
  Mine candidate stance expression rule of frequent object-evaluation pairs $\tilde{\mathcal{R}}^y$ $(y \in \{\mathtt{F}, \mathtt{A}\})$ with Equation 6;
  Identify intra-stance conflicts and obtain stance-indicative objects S^y with Equation 7;
  Identify cross-stance conflicts and obtain purified rule $\mathcal{R}^y$ $(y \in \{\mathtt{F}, \mathtt{A}\})$ with Equation 8;
  Obtain align objects O^{ali} and opposite objects O^{opp} with Equation 9;
  for all document $d \in D$ **do**
    Obtain the reflection prompt $\mathcal{P}_d$ by invoking Algorithm 2;
    if $\mathcal{P}_d \neq \emptyset$ **then**
      Obtain new pseudo label $\hat{y}_d$ $(\hat{y}'_d)$ with Equation 10;
    end if
  end for
end for
return $\hat{y}_d$;

Algorithm 2 Reflection Prompt Generation

Require: The object-evaluation set of a document T_d (containing the object set O_d); the pseudo label of the document $\hat{y}_d$; the sets of align objects O^{ali}, opposite objects O^{opp}, and stance-indicative objects S^y $(y \in \{\mathtt{F}, \mathtt{A}\})$.
Ensure: The reflection prompt $\mathcal{P}_d$.

$\mathcal{P}_d \leftarrow \emptyset$;
for all object $o \in O_d \cap O^{\text{ali}}$ **do**
  ▷ Aligned Object Mismatch
  if $\exists e$ such that $(o, e) \in T_d$ and $\hat{y}_d = e$ **then**
    return $\emptyset$;
  end if
  $\mathcal{P}_d \leftarrow \mathcal{P}_d \circ$ "If the document supports o, the stance is F; If the statement opposes o, the stance is A";
end for
for all object $o \in O_d \cap O^{\text{opp}}$ **do**
  ▷ Opposite Object Mismatch
  if $\exists e$ such that $(o, e) \in T_d$ and $\hat{y}_d \neq e$ **then**
    return $\emptyset$;
  end if
  $\mathcal{P}_d \leftarrow \mathcal{P}_d \circ$ "If the document supports o, the stance is A; If the statement opposes o, the stance is F";
end for
for all object $o \in O_d \cap S^y$ $(y \in \{\mathtt{F}, \mathtt{A}\})$ **do**
  ▷ Stance-indicative Object Mismatch
  if $\hat{y}_d = y$ **then**
    return $\emptyset$;
  end if
  $\mathcal{P}_d \leftarrow \mathcal{P}_d \circ$ "If the document talks about o, the stance tends to be y";
end for
if $\mathcal{P}_d = \emptyset$ **then**
  $o_1 \leftarrow \text{argmax}_{o_1 \in O_d} \text{sim}(o_1, o_2)$ $(o_2 \in O^{\text{ali}})$;
  $\mathcal{P}_d \leftarrow \mathcal{P}_d \circ$ "If the document supports o_1, the stance is F; If the statement opposes o_1, the stance is A";
  $o_1 \leftarrow \text{argmax}_{o_1 \in O_d} \text{sim}(o_1, o_2)$ $(o_2 \in O^{\text{opp}})$;
  $\mathcal{P}_d \leftarrow \mathcal{P}_d \circ$ "If the document supports o_1, the stance is A; If the statement opposes o_1, the stance is F";
end if
return $\mathcal{P}_d$;

B Design of Prompt Texts for LLMs

Prompt Name	Prompt Text
Label-prompt	Given the document : {document}. What is the author's stance towards "{target}"? Select answer from "favor, against, or none". Output Format: Label: [Your chosen label]
Holder-prompt	What is the document comes from? Only return the answer with one sentence. Output Format: Source: [one sentence]
Evaluation-prompt	Given the document : {document}. What is the author's stance towards "{object}"? Select answer from "favor, against, or none". Output Format: Label: [Your chosen label]
Reflect-prompt	Given the document: {document} What is the author's stance towards "{target}"? Instructions: {rule_desc} Select answer from "favor, against, or none". Output Format: Label: [Your chosen label]