Stance Scorer: Zero-shot Stance Detection on Social Media with Fine-grained Labels

Qinlong Fan
State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou, Henan, China
fan_qinlong@163.com

Jicang Lu*
State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou, Henan, China
lujicang@sina.com

Qiankun Pi
State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou, Henan, China
piqiankun2000@163.com

Yi Xia
State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou, Henan, China
summer_one520@163.com

Yilin Liu
State Key Laboratory of Mathematical Engineering and Advanced Computing
Zhengzhou, Henan, China
19293315260@163.com

***Abstract*—Social media platforms serve as crucial arenas for the expression of public opinion, where user-generated content inherently conveys latent attitudes and stances toward specific entities and topics. Recent developments in zero-shot stance detection have achieved remarkable progress. However, most studies merely focus on the three-class stance, and the coarse-grained approach restricts both the depth of semantic comprehension and the optimization of inference strategies. To address this limitation, we propose a novel Fine-grained Stance Scoring framework (Stance Scorer). Specifically, we first establish a quantifiable stance intensity metric through a meticulously designed fine-grained labeling system. Next, we develop an LLM-powered data transformation pipeline to systematically convert categorical labels into precise stance scores. Finally, we implement a model refinement strategy by fine-tuning the derived stance score metrics, with the optimized model serving as the stance scorer. Experimental results show that the proposed Stance Scorer outperforms GPT-4, achieving average F1-score improvements of 6.03%, 3.91%, and 4.12% on the SemEval 2016 Task 6 A, P-Stance, and COVID-19 datasets, respectively. These results validate the effectiveness and robustness of our approach.**

***Index Terms*—stance detection,fine-grained,social media.**

I. Introduction

With the proliferation of online opinions and discussions, stance detection has emerged as a crucial technique for extracting attitudes and identifying polarization from large-scale text data. Traditional stance detection aims to classify the stance of a text toward a specific target (e.g., entity, claim, or topic) into three categories: favor, against, neither [1]. Among its challenges, Zero-Shot Stance Detection, which identifies stances toward unseen targets, has gained significant attention. In practical applications, particularly in online debates, detecting argumentative relationships between posts is essential. This task requires not only stance polarity but also stance strength (e.g., strong, moderate, slight) [2]. Recent studies demonstrate that integrating polarity and strength information enhances stance classification and aids in detecting phenomena like opinion polarization [3] and deviation [4].

Generalization ability is a critical factor in Zero-shot Stance Detection, as it determines a model's capacity to identify stances toward unseen targets. This capability hinges on a deep understanding of stance concepts and robust reasoning strategies. Early work [5] introduced a benchmark dataset and foundational models for Zero-shot Stance Detection, spurring subsequent advancements [6]–[9]. However, current methods still face limitations in both performance and generalization, largely due to insufficient conceptual understanding of stance. Specifically, stance intensity varies based on users' perspectives and knowledge levels, often implicitly encoded in text rather than explicitly stated. While most studies focus on stance polarity and intensity, accurately detecting these nuances remains challenging. Effectively integrating polarity and intensity information is essential, as it enhances conceptual understanding and reasoning, addressing a key challenge in stance detection research.

To address these limitations, we propose a novel Stance Scorer framework that integrates fine-grained stance quantification (stance scores) and an enhanced prediction methodology. Specifically, we introduce a refined stance scoring system that extends the traditional three-class labeling scheme to enable precise characterization of textual stances toward targets. We further augment the original dataset using pre-trained language models, constructing a comprehensive stance score corpus. The Stance Scorer is fine-tuned on this dataset using the open-source LLaMA3.1-8B-Instruct model, enabling fine-grained stance perception and classification. Experiments on three benchmark datasets (SemEval 2016 Task 6 A, P-Stance, and COVID-19) demonstrate that our framework, leveraging fine-grained stance labels, achieves significant improvements in stance detection performance.

Our main contributions are as follows:

*Corresponding author.

1) We explore the role of stance scores in stance detection, showing how they reflect stance intensity at a fine-grained level and improve model understanding, leading to more accurate classification.
2) We propose a fine-grained stance scoring framework that integrates stance polarity and intensity. Our approach includes an LLM-driven data transformation pipeline to systematically convert categorical labels into continuous stance scores, and a stance scorer designed for zero-shot detection.
3) Experiments on three public stance detection benchmarks achieves significant performance improvements on social media data, underscoring the effectiveness of fine-grained stance representations in advancing detection accuracy.

II. Related Work

We review recent progress in Zero-shot Stance Detection and analyze how fine-grained label representations enhance stance detection performance.

A. Zero Shot Stance Detection

Zero-shot Stance Detection aims to identify stances toward unseen topics, which poses a significant challenge in NLP. [6] analyzed GPT-4's impact on stance detection paradigms and evaluation, highlighting paradigm shifts in the pre-trained model era. [10] introduced debate-style role-interactive LLM agents for zero-shot stance detection aligned with human reasoning. Additionally, approaches relying on external knowledge bases are limited by incomplete or irrelevant information, reducing their real-world applicability [11]. Most existing methods adopt a three-class framework (favor, against, neither), which fails to capture the nuanced and implicit variations in real-world stance expressions. This oversimplification restricts models' ability to accurately represent stance complexity, highlighting the need for more sophisticated approaches to address the subtlety and diversity of real-world stance detection.

B. Fine-Grained Labeling

Fine-grained labeling categorizes data into detailed classes to capture subtle stance distinctions, improving model performance. Prior work, such as [12], classifies stances into five categories (Strongly Favor, Favor, Neutral, Against, Strongly Against), while [13] refines this by incorporating both polarity and intensity. However, these methods often fail to capture nuanced stance differences effectively. For example, [2] proposes a regression-based approach, encoding stances as continuous values (-1.0 to +1.0 in 0.2 increments), where sign indicates polarity and magnitude reflects intensity. While this represents progress, it remains limited by discretization, inadequate handling of linguistic nuances, lack of contextual integration, and difficulty addressing ambiguous stances. Similarly, [14] extends classification to four categories (Favor, Against, Neutral, Irrelevant) but struggles to differentiate degrees of support or opposition within Favor and Against. Existing methods rely heavily on manual annotation, which introduces human biases, inconsistencies, and challenges in capturing subtle linguistic and contextual cues. This underscores the need for more precise, automated annotation techniques to minimize human-induced inaccuracies and enhance stance representation.

III. Stance Scores

We analyze why stance scores should be used in Zero-shot Stance Detection tasks, based on the characteristics of real-world data, and describe how stance scores can be applied to address stance detection tasks.

A. Why Use Stance Scores?

In real-world scenarios, textual stance toward a target often exhibits varying degrees of favor or opposition, typically expressed implicitly rather than explicitly. Existing research primarily focuses on stance polarity, overlooking nuanced distinctions in stance degrees. This limitation hinders models' ability to capture implicit semantic information, making them susceptible to irrelevant signals and compromising stance understanding. Thus, integrating stance polarity with fine-grained degree information to model subtle stance variations remains a critical yet understudied challenge in stance analysis.

To address the challenge of stance detection, we propose integrating stance polarity with fine-grained degree information through stance scores. This extension of traditional polarity classification enables a more nuanced representation of stances, enhancing the model's ability to handle implicit semantics and complex contexts. Our approach offers two key advantages. First, stance scores enable the identification of partial agreement, a prevalent but often overlooked phenomenon in real-world discussions where users express nuanced positions rather than absolute favor or opposition. Capturing such subtleties allows for a more comprehensive analysis of diverse perspectives. Second, fine-grained labeling improves semantic detail extraction, optimizing feature representation and leveraging implicit degree information in the text. This not only strengthens stance understanding but also refines reasoning strategies, enhancing their adaptability to diverse scenarios. These contributions provide novel insights for stance detection research.

Digitization, the process of converting information into digital form, further enhances stance detection by enabling fine-grained representation of stance degrees through numerical scores. A well-defined score range balances the reflection of stance variations with label simplicity, avoiding noise from excessive granularity. This approach aligns with human cognitive patterns, improving model training efficiency and predictive performance. By transforming stances into computationally tractable formats, digitization facilitates deeper analysis of textual attitudes and complex arguments, advancing stance detection toward greater precision. Additionally, it captures subtle stance differences, reflects their intensity and direction, and provides structured, information-rich data for model training, ultimately improving overall task performance.

B. Scoring Analysis

Key statistical methods are computed as follows:

- **Fleiss' Kappa** (κ) [15]:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \begin{cases} \bar{P} & \text{observed inter-rater agreement} \\ \bar{P}_e & \text{expected chance agreement} \end{cases} \quad (1)$$

  ▷ *Computation*: Let n=subjects, k=categories, m=raters. Define $p_j = \frac{1}{nm}\sum_{i=1}^{n} n_{ij}$ (proportion of assignments to category j), then:

$$\bar{P}_e = \sum_{j=1}^{k} p_j^2, \quad \bar{P} = \frac{1}{n}\sum_{i=1}^{n} \frac{\sum_{j=1}^{k} n_{ij}^2 - m}{m(m-1)}$$

  ▷ *Interpretation*: $\kappa > 0.60$ (substantial agreement). Weaknesses: Sensitive to category prevalence/rater bias.

- **Spearman's** ρ (nonparametric rank corr.) [16]:
  - *Data*: Rank pairs $(R_x(i), R_y(i))$, $R_x(i), R_y(i) \in \{1, \ldots, n\}$
  - *Ties adjustment*: Assign average ranks for ties, use adjusted formula:

$$\rho = \frac{\sum_{i=1}^{n}(R_x(i) - \bar{R}_x)(R_y(i) - \bar{R}_y)}{\sqrt{\sum_{i=1}^{n}(R_x(i) - \bar{R}_x)^2 \sum_{i=1}^{n}(R_y(i) - \bar{R}_y)^2}} \quad (2)$$

  - *No-ties simplification*: $d_i^2 = [R_x(i) - R_y(i)]^2$, $\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$
  - *Hypothesis testing*: $t = \rho\sqrt{\frac{n-2}{1-\rho^2}}$ (df = $n-2$); $p = 2 \cdot \min\{P(T \le t), P(T \ge t)\}, T \sim t(n-2)$. Reject H_0 (independence) if $p < 0.001$. Robust to outliers (assumes monotonicity).

- **KL Divergence** (information-theoretic measure) [17]:

$$D_{\mathrm{KL}}(P \parallel Q) = \begin{cases} \sum_i P(i) \log \frac{P(i)}{Q(i)} & \text{(discrete)} \\ \int_{-\infty}^{\infty} P(x) \log \frac{P(x)}{Q(x)} dx & \text{(continuous)} \end{cases} \quad (3)$$

  ◇ *Asymmetry*: $D_{\mathrm{KL}}(P \parallel Q) \neq D_{\mathrm{KL}}(Q \parallel P)$ (forward/reverse divergence matters in variational inference).

  ◇ *Non-negativity*: $D_{\mathrm{KL}} \ge 0$ (equality iff $P = Q$ a.e.).

  ◇ *Entropy relation*: $D_{\mathrm{KL}}(P \parallel Q) = H(P, Q) - H(P)$ (cross-entropy $H(P, Q)$).

To assess the reliability of the GPT-4-generated scores, we conducted a rigorous human evaluation study on 300 randomly sampled instances from the Webis_Score dataset, using the same information and prompts provided to the LLM. Three domain experts independently annotated each instance using the same scoring rubric; in cases of discrepancy, the three experts reached a consensus to determine the final score. The agreement between model-generated and human expert annotations was measured, yielding a Fleiss' kappa of $\kappa = 0.70$ (indicating substantial agreement). GPT-4 scores demonstrated a strong correlation with human judgments (Spearman's $\rho = 0.86$, $p < 0.001$).

Figure 1 compares the distribution of GPT-4-generated scores against human annotations. Although both distributions peak around the median values (GPT4 $\mu = 0.1$ vs. Human $\mu = 0.0$),

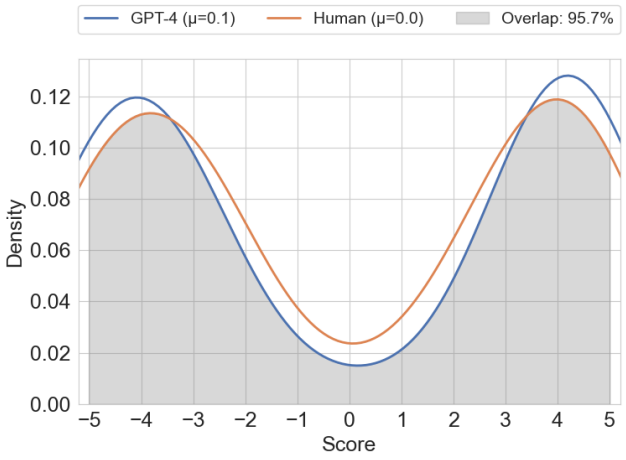

Fig. 1. Comparative Analysis of Score Distributions: GPT-4 vs. Human Annotations — Characteristics of Mean and Kurtosis

the LLM scores exhibit marginally heavier tails (kurtosis = -1.8 vs. -1.7). Further analysis using the Kullback-Leibler divergence reveals limited distributional divergence ($D_{\mathrm{KL}}(P \parallel Q) = 0.05$), with over 95.7% probability mass overlap between the two distributions.

Through multi-faceted validation (expert-human alignment: κ=0.70, ρ=0.86; distributional similarity: $D_{\mathrm{KL}}(P \parallel Q)$ = 0.05, 95.7% overlap; qualitative granularity analysis), we demonstrate GPT-4's capacity to generate robust stance scores annotations. The convergence of agreement metrics, distributional parallels, and categorical coverage establishes the reliability of automated scoring systems within defined operational boundaries.

Our findings support that our comprehensive evaluation of scoring quality, distribution characteristics and potential scoring biases establishes the practical feasibility of employing large language models for reliable stance scores annotation at scale, while maintaining rigorous psychometric properties comparable to expert human assessments.

C. Task Description

Data optimization enhances data quality and usability through various techniques. We define a stance detection dataset $D_o = \{(x_i^o, t_i^o, y_i^o)\}_{i=1}^n$, where x_i^s is the text, t_i^s is the target, and y_i^s is the stance label. Using an external model M, we infer the stance score s_i^s for each text-target pair, resulting in an optimized corpus $D_s = \{(x_i^s, t_i^s, y_i^s, s_i^s)\}_{i=1}^n$, where s_i^s represents the stance score.

Stance detection aims to identify the attitude of a text toward a specific target. Given a dataset $D = \{(x_i, t_i, y_i)\}_{i=1}^n$, where x_i denotes the text, t_i represents the target, and y_i is the stance label. The modified stance detection task, based on a stance scorer, involves inferring a stance score s_i for each x_i-t_i pair and aligning it with y_i.

IV. Method

We present Stance Scorer, a Zero-shot Stance Detection framework that enhances performance by utilizing fine-grained stance scores. As shown in Figure 2, the method reformulates

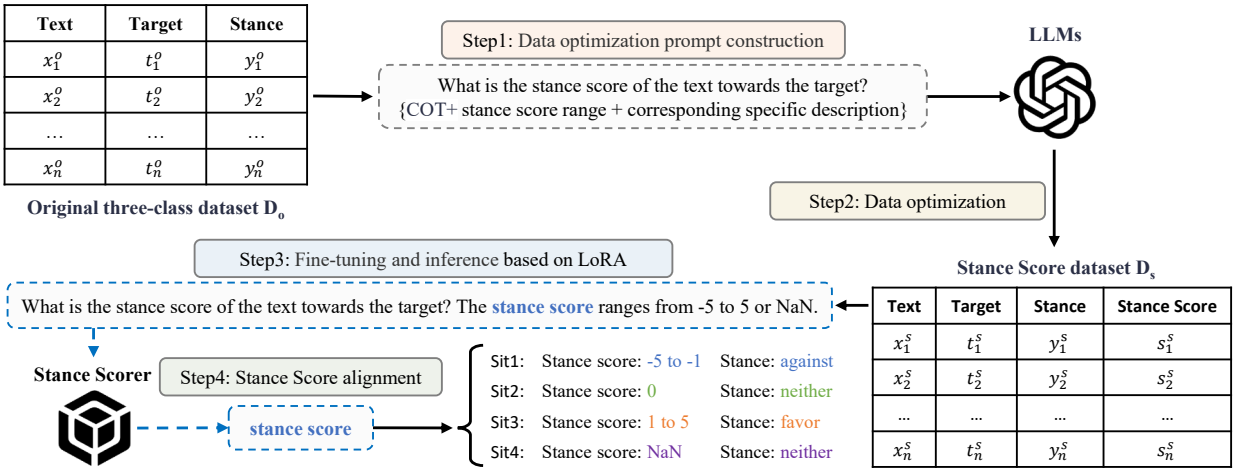

Fig. 2. The Model Architecture of the Stance Scorer

the traditional three-class stance labels into a continuous stance score representation. We propose an LLM-driven pipeline to transform categorical labels into continuous stance scores, producing a high-quality stance score corpus for downstream applications. Stance Scorer is fine-tuned on this corpus using an open-source model, and the predicted stance scores are subsequently mapped back to the original three-class labels to ensure consistency.

A. Labeling System Construction

We developed a stance labeling system based on stance scores, extending the original three-class labeling system (Favor, Against, Neither). From a fine-grained perspective, the subdivided labels effectively reflect varying degrees of attitude, capture subtle changes, and reveal the complexity of stances. From a quantitative perspective, stances can be represented by numerical scores, where the polarity and magnitude of the numbers correspond to the polarity and intensity of the stance. In classifying the degrees of "favor" and "oppose," the design of hierarchical levels is particularly crucial. Too few levels may fail to capture the full complexity of attitudes, whereas too many levels may result in over-fine-grained differentiation, introducing uncertainty and noise into the classification. Therefore, we divided attitudes into five levels: "Completely ," "Strongly," "Moderately," "Somewhat," and "Slightly." This design not only reflects the different degrees of favor and opposition in detail, avoiding oversimplification or overcomplication, but also facilitates the model in precisely capturing subtle shifts in attitude. The distinctions between these five levels are clear and intuitive, reducing potential confusion caused by inappropriate category settings and thereby improving the model's generalization ability and accuracy when handling unseen data. Additionally, the introduction of Neutral and Irrelevant labels enriches the dimensionality of stance analysis, aligning the study more closely with real-world issues and further enhancing the reliability and validity of the results. Based on this analysis, we propose a fine-grained stance annotation scheme, ranging from -5 to 5, as illustrated in Figure 3. Negative values indicate opposition, with -5 representing complete opposition and -1 indicating slight opposition. Positive values denote support, where 1 represents slight support and 5 indicates complete support. A neutral stance is labeled as 0, while NaN marks unrelated content. This scale effectively captures varying degrees of stance intensity and distinguishes relevant from irrelevant texts.

B. Data Optimization

Manual annotation has several limitations: it is prone to errors and biases, difficult to capture nuanced stances, costly, and often inconsistent with practical requirements. To mitigate these issues, we propose a data optimization method leveraging large models, centered on prompt design to generate fine-grained stance labels (stance scores). Effective prompt design is critical, requiring clarity, feasibility, and iterative testing. Our approach constructs prompts based on a predefined stance score label system, ensuring goal clarity and leveraging the model's robust natural language understanding. Iterative testing and feedback further enhance prompt quality, improving the reliability of stance score inference.

> **Data Optimization Prompts:** Please evaluate the stance score of the text towards the target based on the following criteria and generate the corresponding thought chain. The stance of the text towards the target is label. What is the stance score of the text towards the target? [Stance score range and specific description of corresponding scores, as shown in Figure 3].

We propose a method where text, target, and a three-class stance label (favor, against, neither) are input into a large model to infer fine-grained stance scores. The stance label determines the score range: -5 to -1 for "against," 1 to 5 for "favor," and 0 or NaN for "neither." Guided by the label, score range, and descriptions, the model outputs a stance score. The inferred

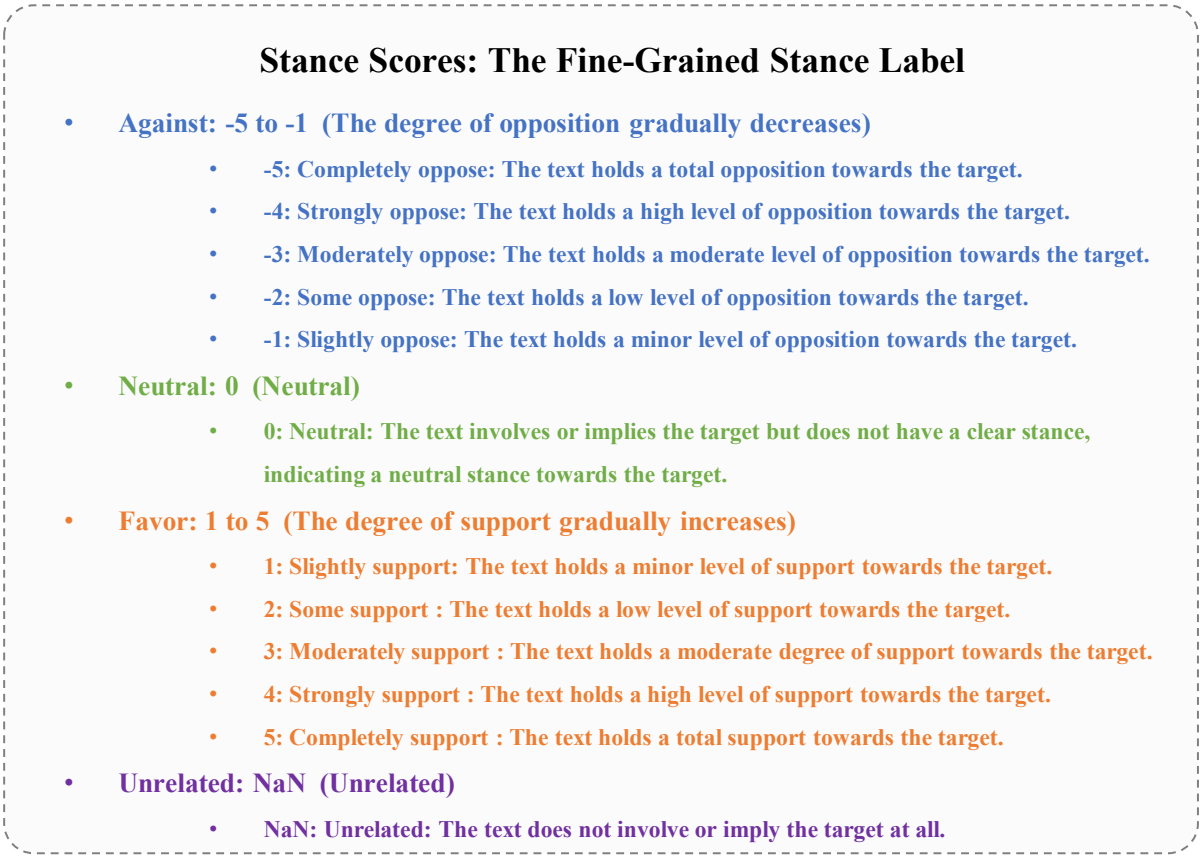

Fig. 3. A Fine-grained Labeling System Using Stance Scores

scores are integrated with the original dataset to construct a fine-grained stance score corpus, including text, target, stance label, and corresponding stance scores.

C. Fine-Tuning, Reasoning, and Alignment

We adopt a supervised fine-tuning approach based on Low-Rank Adaptation (LoRA) to fine-tune an open-source large model on the stance score corpus and perform inference. Leveraging the stance score labeling system, we design a task-specific prompt for fine-tuning and inference.

> **Fine-tuning and Inference Prompt:** Please evaluate the stance score of the text towards the target based on the following criteria. What is the stance score of the text towards the target? The stance score ranges from -5 to 5 or NaN. [Descriptions of all scores]

In the fine-tuning phase, we select the LLaMA3.1-8B-Instruct for its strong performance in NLP tasks, efficient 8-billion-parameter architecture, and low hardware and operational costs, making it ideal for resource-constrained settings. Its compatibility with mainstream frameworks (e.g., TensorFlow, PyTorch) simplifies deployment and adaptation, while extensive documentation and community support ease usability. Given its balance of performance, cost-effectiveness, and accessibility, we employ LLaMA3.1-8B-Instruct for supervised fine-tuning on fine-grained stance score data to build a stance scorer based on fine-grained labels. In the stance inference phase, we employ a fine-tuned stance scorer to evaluate stance scores of texts toward targets in the test data. The inferred results are aligned with the original three-class stance labels to assess the model's effectiveness and reliability.

TABLE I
STATISTICS OF SEMEVAL 2016 TASK 6 A, P-STANCE AND COVID-19 DATASETS.

Dataset	Target	Favor	Against	Neither
SemEval 2016 Task 6 A [1]	A	124	464	145
	CC	335	26	203
	FM	268	511	170
	HC	163	565	256
	LA	167	544	222
P-Stance [18]	JB	3217	4079	-
	BS	3551	2774	-
	DT	3663	4290	-
COVID-19 [19]	WA	515	220	172
	SC	430	102	85
	AF	384	266	307
	SH	151	201	396

V. EXPERIMENTS

We evaluate the proposed method on three benchmark stance detection datasets. We detail the datasets, evaluation metrics,

and experimental setup, and compare against several baseline models. Results, including ablation studies and a comprehensive analysis of the model, are presented and discussed.

A. Evaluation Dataset

We evaluate on three Twitter stance detection datasets: SemEval 2016 Task 6 A [1], P-Stance [18], and COVID-19 [19]. The statistics is shown is Table I. SemEval 2016 Task 6 A focuses on three-class stance classification (favor, against, neither) toward entities like Hillary Clinton and climate change. P-Stance provides binary stance labels for Joe Biden, Bernie Sanders, and Donald Trump; we exclude 'neither' samples due to annotation inconsistencies. COVID-19 includes 3,229 tweets on COVID-19 policies (e.g., mask-wearing, school closures), labeled as favor, against, or neither.

B. Evaluation Criteria

To ensure fair evaluation, we adopt metrics consistent with prior work for three datasets. For SemEval 2016 Task 6 A and COVID-19, we use the macro-average F1 score (F_{avg}) across the three classes (favor, against, neither) as in [1], computed as $F_{\text{avg}} = \frac{F_{\text{favor}} + F_{\text{against}} + F_{\text{neither}}}{3}$. For P-Stance, a binary classification task (favor, against), F_{avg} is computed as $F_{\text{avg}} = \frac{F_{\text{favor}} + F_{\text{against}}}{2}$.

C. Experimental Setup

All experiments were conducted on NVIDIA TESLA V100S 32GB GPUs. We fine-tuned the LLaMA3.1-8B-Instruct model on the Webis_Stance dataset [22] and the Webis_Score dataset (optimized from Webis_Stance) using AdamW optimization. The training configuration included a learning rate of 1e-5, a batch size of 16, and a maximum sequence length of 1024, with 3 epochs and a warm-up step of 0.1. Data optimization was performed using GPT4.

D. Baseline

To evaluate the proposed model, we conducted a comprehensive comparison with established baselines, including: KASD [20], MB-Cal [21], COLA [10], WS-BERT-Dual [7],BERT [23], SSCL [24], GPT4, and LLAMA.

VI. Results

We assess the performance of Stance Scorer and baseline models on Zero-shot Stance Detection using the SemEval 2016 Task 6 A, P-Stance, and COVID-19 datasets. Detailed results are presented in Table II, Table III, and Table IV.

Our experiments demonstrate that the proposed Stance Scorer significantly enhances Zero-shot Stance Detection, achieving strong performance in terms of overall mean F1 score on public benchmarks and competitive results on target-specific metrics compared to state-of-the-art models. We argue that fine-grained stance labeling with stance scores offers two key advantages over traditional three-class labeling. First, stance scores capture both polarity and intensity, providing richer information for nuanced stance representation. Second, this approach better models implicit textual information and aligns with real-world fine-grained stances, enhancing semantic understanding and reasoning capabilities. The effectiveness arises from detailed stance degree segmentation, enabling the model to effectively apply fine-grained distinctions in real-world tasks.

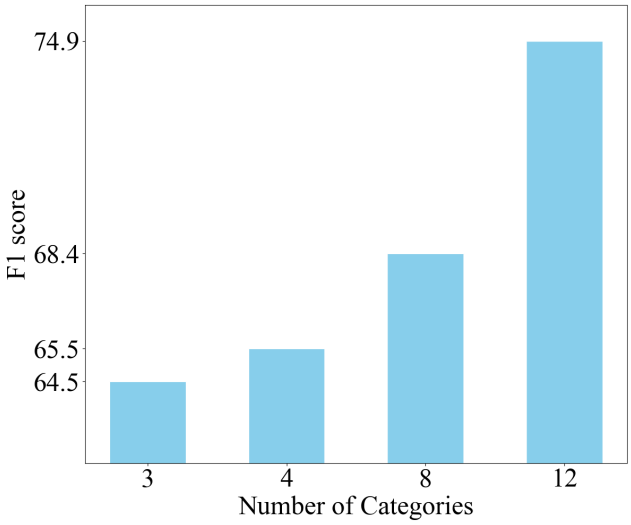

Fig. 4. The Impact of Fine-grained Stance Score Granularity on F1 Score

A. Fine-grained Analysis

We investigate the effect of fine-grained stance scoring on model performance, as illustrated in Figure 4. Our analysis compares the Stance Scorer's performance on the SemEval 2016 Task 6 A test set against multiple baselines: a three-class model (favor, against, neither), a four-class model (splitting "neither" into "neutral" and "irrelevant"), an eight-class model (refining "favor" and "against" into three levels each), and the Stance Scorer, which further divides "favor" and "against" into five levels. Results highlight the impact of granularity on stance detection accuracy.

As the granularity of stance scoring increases, the model's stance understanding and task performance improve substantially, provided that the quality of the prompts is maintained. This highlights the positive impact of fine-grained stance labeling on Zero-shot Stance Detection, underscoring the critical role of detailed data annotation in enhancing model effectiveness. However, subdividing "favor" and "against" into seven levels results in suboptimal model performance. We attribute this to two key factors: limitations in prompt design and data optimization hinder further fine-graining, and excessive granularity introduces data noise, which constrains model optimization. While fine-graining generally enhances performance, it is crucial to balance granularity with cost and efficiency to maintain model effectiveness and accuracy.

Our analysis of the Webis fine-tuning dataset and inference results on SemEval 2016 Task 6 A reveals that scores such as "-1," "1," "-2," and "2" are underrepresented in both datasets. Due to manual annotation constraints and limited coverage of subtle stance variations (e.g., weakly favor or against), the fine-tuned model fails to accurately capture these nuanced distinctions. This underscores a key challenge in stance detection for less pronounced expressions. While models ignoring weak stances

TABLE II
ZERO-SHOT STANCE DETECTION EXPERIMENTAL RESULTS ON SEMEVAL 2016 TASK 6 A.

Model	SemEval 2016 Task 6 A(%)					
	HC	FM	LA	AT	CC	Avg
KASD-GPT4 [20]	80.3	70.4	62.7	64.0	55.8	66.6
GPT3.5-MB-Cal [21]	78.5	**75.0**	66.0	66.9	67.2	70.7
COLA [10]	**81.7**	63.4	**71.0**	70.8	65.5	70.5
GPT4	74.4	72.0	64.2	48.9	68.3	68.9
LLaMA3.1-8B-Instruct	42.1	49.2	42.0	44.3	42.1	50.3
Stance Scorer	80.0	70.2	66.3	**72.6**	**71.9**	**74.9**

TABLE III
ZERO-SHOT STANCE DETECTION EXPERIMENTAL RESULTS ON P-STANCE.

Model	P-Stance(%)			
	JB	BS	DT	Avg
WS-BERT-Dual [7]	77.9	71.6	69.2	72.9
KASD-GPT4 [20]	83.6	79.7	84.3	82.5
GPT4	78.9	78.1	79.6	79.3
LLaMA3.1-8B-Instruct	81.1	74.6	68.5	75.2
Stance Scorer	**83.8**	**80.2**	**85.1**	**83.4**

TABLE IV
ZERO-SHOT STANCE DETECTION EXPERIMENTAL RESULTS ON COVID-19.

Model	COVID-19(%)				
	AF	SC	SH	WA	Avg
BERT [23]	47.5	45.1	39.7	44.3	44.2
SSCL [24]	51.8	53.7	54.5	59.5	54.9
GPT4	74.6	52.4	68.9	74.2	68.6
LLaMA3.1-8B-Instruct	47.6	32.5	41.7	52.1	43.3
Stance Scorer	**75.0**	**66.1**	**69.1**	**75.5**	**72.7**

may suffice for coarse-grained tasks, distinguishing weak stances is critical for identifying near-neutral attitudes, enabling deeper stance understanding. Although weak stances are less frequent, they often reflect implicit, socially significant attitudes prevalent in real-world scenarios. To enhance practical utility and alignment with human perception, stance detection models must accurately reflect these nuanced stances, improving real-world applicability and precision.

B. Ablation Study

We conduct an ablation study on the SemEval 2016 Task 6 A dataset using LLaMA3.1-8B-Instruct as the base model to evaluate the impact of different modules and fine-tuning datasets. Specifically:

- "w/o stance" excludes the Webis_Stance dataset.
- "w/o score" excludes the Webis_Score dataset and adopts a three-class stance prompt.

Table V shows that removing any dataset during fine-tuning significantly lowers the F1 score, underscoring the importance of each dataset. The Webis_Score dataset, constructed based on score prompts, is particularly crucial. Fine-grained stance labels (stance scores) outperform the three-class scheme ("favor," "against," "neither") in stance detection, as they enable the model to better analyze complex texts by capturing nuanced stance concepts. By incorporating both stance polarity and degree information, the model improves its ability to infer implicit stances and refine its reasoning strategy.

C. Case Study

Table VI presents a case study comparison between our proposed fine-grained stance scoring model and a three-class stance classification baseline. Specifically, the "Stance" column reports results from the baseline, which does not adopt the scoring stratedoes not adopt the scoring strategy and is only fine-tuned with stance prompts on the same dataset. In contrast, the "Score" column displays outputs from our stance scorer, which incorporates the stance scoring mechanism. Results indicate that our stance scorer (reflected in the "Score" column) delivers higher accuracy and more robust reasoning effectiveness than the non-scoring baseline (results in the "Stance" column). This highlights our model's improved capability for precise stance prediction and in-depth analysis.

The stance scorer enhances stance detection by integrating both stance polarity (e.g., favor, against, neither) and degree, enabling effective identification of implicit stances. For instance, in Case 5, it uncovers subtle biases in seemingly neutral text, improving prediction accuracy. This nuanced understanding boosts the model's adaptability and generalization in complex scenarios, while its ability to detect subtle stance shifts and semantic differences ensures robust performance across diverse texts. These advancements not only elevate model performance but also offer valuable insights for future research, particularly in applications requiring fine-grained stance analysis.

VII. CONCLUSION

Existing zero-shot stance detection methods predominantly rely on coarse-grained three-class labels, which fail to capture stance intensity and implicit semantic information. To address this, we propose Stance Scorer, a novel framework integrating fine-grained stance scoring and LoRA-fine-tuned inference. Extensive experiments on three benchmark datasets demonstrate that Stance Scorer significantly improves stance detection accuracy by effectively capturing subtle attitude variations and enhancing robustness. Case studies further verify its capability to identify implicit stances and optimize reasoning strategies.

TABLE V
PERFORMANCE COMPARISON OF STANCE SCORER WITH LESS FINE-TUNED DATASET.

Model	SemEval 2016 Task 6 A					
	HC	FM	LA	AT	CC	Avg
Stance Scorer	**80.0**	**70.2**	**66.3**	**72.6**	**71.9**	**74.9**
w/o stance	77.7	61.4	66.0	69.0	63.4	71.2
w/o score	62.5	64.8	62.5	43.4	67.4	65.0
w/o stance & score	42.1	49.2	42.0	44.3	42.1	50.3

TABLE VI
CASE STUDY OF SEMEVAL 2016 TASK 6 A

Target	Text	Label	Stance	Score
Hillary Clinton	I think marriage is as it has always been, between a man and a woman. -HillaryClinton #MarriageEquaility #justsaying	against	favor	-3.0(against)
Legalization of Abortion	in turkey at ze momento, just got internet and started reading @barrettweed's tweets, literally just killing it!!	neither	favor	NaN(neither)
Feminist Movement	i'm a co-omnipotent-faux-fallacious-poly-concupiscent-inter- equanimity feminist now. you may bow to me.	against	favor	-5.0(against)
Climate Change is a Real Concern	if complaining about irrelevant problems were illegal maybe we could solve some of the important issues #itsjustaflag	neither	favor	0.0(neither)
Atheism	#Religions can't all be right, but they can all be wrong.	favor	neither	2.0(favor)

REFERENCES

[1] S. Mohammad, S. Kiritchenko, P. Sobhani *et al.*, "Semeval-2016 task 6: Detecting stance in tweets," in *In Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, 2016, pp. 31–41.

[2] J. Sirrianni, X. Liu, and D. Adams, "Agreement prediction of arguments in cyber argumentation for detecting stance polarity and intensity," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 5746–5758.

[3] J. Sirrianni, X. Frank, and D. Adams, "Quantitative modeling of polarization in online intelligent argumentation and deliberation for capturing collective intelligence," in *2018 IEEE International Conference on Cognitive Computing (ICCC)*. IEEE, 2018, pp. 57–64.

[4] R. S. Arvapally, X. F. Liu, F. F. H. Nah *et al.*, "Identifying outlier opinions in an online intelligent argumentation system," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 8, p. e4107, 2021.

[5] E. Allaway and K. McKeown, "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)*, pp. 8913–8931, 2020.

[6] Y. Zhang, M. Li, T. Wang *et al.*, "How would stance detection techniques evolve after the launch of ChatGPT?" *Journal of Natural Language Processing*, vol. 30, no. 3, pp. 123–135, 2022.

[7] Z. He, N. Mokhberian, and K. Lerman, "Infusing knowledge from wikipedia to enhance stance detection," *arXiv preprint arXiv:2204.03839*, 2022.

[8] H. Zhang, Y. Li, T. Zhu, and C. Li, "Commonsense-based adversarial learning framework for zero-shot stance detection," *Neurocomputing*, vol. 563, p. 126943, 2024.

[9] M. Taranukhin, V. Shwartz, and E. Milios, "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)*, pp. 15 257–15 272, 2024.

[10] X. Lan, C. Gao, D. Jin, and Y. Li, "Stance detection with collaborative role-infused llm-based agents," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 18, 2024, pp. 891–903.

[11] K. Ma, J. Francis, Q. Lu *et al.*, "Towards generalizable neuro-symbolic systems for commonsense question answering," *In Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, p. 22–32, 2019.

[12] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, 1932.

[13] J. W. Sirrianni, X. Liu, and D. Adams, "Predicting stance polarity and intensity in cyber argumentation with deep bidirectional transformers," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 3, pp. 655–667, 2021.

[14] C. Conforti, J. Berndt, M. T. Pilehvar *et al.*, "Will-they-won't-they: A very large dataset for stance detection on twitter," *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 1715–1724, 2020.

[15] J. L. Fleiss, "Measuring nominal scale agreement among many raters." *Psychological bulletin*, pp. 378–382, 1971.

[16] C. Spearman, "The proof and measurement of association between two things." *The American Journal of Psychology*, pp. 72–101, 1904.

[17] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, pp. 79–86, 1951.

[18] Y. Li, T. Sosea, A. Sawant *et al.*, "P-stance: A large dataset for stance detection in political domain," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP*, 2021, pp. 2355–2365.

[19] K. Glandt, S. Khanal, Y. Li *et al.*, "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)*, vol. 1, 2021.

[20] A. Li, B. Liang, J. Zhao *et al.*, "Stance detection on social media with background knowledge," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 15 703–15 717.

[21] A. Li, J. Zhao, B. Liang, L. Gui, H. Wang, X. Zeng, K.-F. Wong, and R. Xu, "Mitigating biases of large language models in stance detection with calibration," *arXiv preprint arXiv:2402.14296*, 2024.

[22] Y. Ajjour, M. Alshomary, H. Wachsmuth, and B. Stein, "Modeling frames in argumentation," 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)*, 2019, pp. 2922–2932.

[23] J. Devlin, "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*, p. 4171–4186, 2019.

[24] J. Zou, X. Zhao, F. Xie *et al.*, "Zero-shot stance detection via sentiment-stance contrastive learning," in *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2022, pp. 251–258.