PiKGL: Leveraging Pruned Knowledge Graphs for Explainable Stance Detection

Bingbing Wang[1,2*]**, Jingjie Lin**[1*]**, Zhixin Bai**[1]**, Xintong Song**[1]**, Qianlong Wang**[1]
Min Yang[3]**, Xi Zeng**[4]**, Jing Li**[2]**, Ruifeng Xu**[1,5†]
[1] Harbin Institute of Technology, Shenzhen, China
[2] The Hong Kong Polytechnic University, Hong Kong, China
[3] SIAT, Chinese Academy of Sciences, Shenzhen, China
[4] The 30th Research Institute of China Electronics Technology Group Corporation, Chengdu, China
[5] Peng Cheng Laboratory, Shenzhen, China

Abstract

Stance detection on social media plays a vital role in understanding public opinion on contentious topics. While prior work leverages external knowledge sources like Wikipedia to enrich limited target information, it primarily introduces conceptual content, neglecting the interpretability potential of knowledge and often leading to the incorporation of irrelevant or redundant information that hinders stance prediction performance. To address this, we introduce **PiKGL**, a **P**runed **i**nterpretable **K**nowledge **G**raph **L**earning framework for explainable stance detection. Specifically, we first extract event triplets and topics to obtain real-world knowledge, which is then used to construct an interpretable knowledge graph. To ensure precision and minimize noise, we introduce a retrieval-guided pruning strategy that incorporates commonsense knowledge, filtering redundant information of the interpretable knowledge graph. Finally, the pruned knowledge graph is injected into a large language model to jointly model textual, target, and commonsense for improved stance comprehension. Experimental results conducted on three public datasets demonstrate our PiKGL achieves state-of-the-art performance on stance detection.

1 Introduction

Stance detection, a fundamental task in natural language processing, aims at determining whether an author conveys a positive, negative, or neutral attitude toward a specific subject (Jiang et al., 2022; Wen and Hauptmann, 2023). This task plays a vital role in capturing diverse viewpoints across various media platforms, such as social media and news outlets (Li et al., 2021; Lauscher et al., 2022; He et al., 2021). The contexts provided in stance detection tasks are often brief, posing significant challenges for data-driven models to accurately infer the stance toward a target due to the limited information (Yan et al., 2024; Wei and Mao, 2019).

Recently, Pre-trained Language Models (PLMs) have become the standard backbone for enhancing stance detection models by integrating learned commonsense knowledge, achieving significant advancements in this field (Zhang et al., 2024a; Chunling et al., 2023). A common strategy to further enrich target-specific knowledge is to incorporate background knowledge related to the target as supplementary information for the pretrained stance detection model. This approach has been shown to improve model performance substantially (Yan et al., 2024; Zhu et al., 2022; Chen et al., 2024a). However, these methods often depend heavily on the quality and relevance of knowledge retrieved from sources like Wikipedia, as well as the reasoning capabilities of the models. As illustrated in Figure 1, Wikipedia typically provides definitional knowledge. In contrast, our approach constructs and then prunes a knowledge graph, a distillation process that filters out noise to retain salient, target-relevant information. The resulting high-precision graph empowers the model to better discern complex relationships, thereby enhancing stance detection performance.

To address this, we propose the **P**runed **i**nterpretable **K**nowledge **G**raph **L**earning (**PiKGL**) framework for stance detection, combining external knowledge in the graph structure and using the pruning method to enhance both accuracy and interpretability. To be specific, we first extract event triplets and topics from

*Equal contribution. † Corresponding authors.
`https://github.com/HITSZ-HLT/PiKGL`.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 217–232, 2026. https://doi.org/10.1162/TACL.a.612
Action Editor: Nitin Madnani. Submission batch: 4/2025; Revision batch: 8/2025; Published 3/2026.

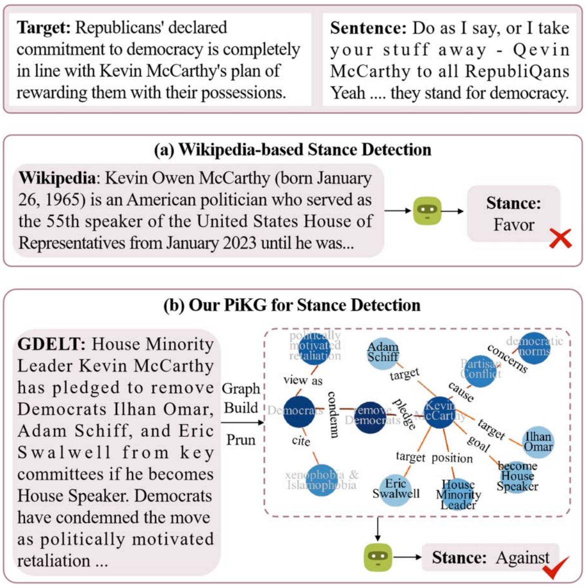

Figure 1: An example demonstrating two approaches.

text and target, leveraging the Global Database of Events, Language, and Tone (GDELT) project[1] (Leetaru and Schrodt, 2013) to obtain event-aware knowledge, which is utilized to construct an enriched interpretable knowledge graph representing real-world relationships through degree-prioritized graph connection. Subsequently, we implement a retrieval-guided pruning strategy that employs commonsense knowledge derived from the target to filter the redundant or irrelevant information, ensuring the graph remains concise and highly relevant to the target stance. Finally, the refined knowledge graph is injected into the Large Language Model (LLM), enabling joint modeling of text, target, and commonsense knowledge for improved stance detection. The main contributions of our work are summarized as follows:

- We introduce PiKGL, a novel framework that constructs an interpretable knowledge graph enriched with real-world knowledge for explainable stance detection.
- A retrieval pruning mechanism is designed to eliminate redundant knowledge and preserve highly relevant information, ensuring the focus of the knowledge graph.
- Extensive experiments on three datasets demonstrate that our PiKGL achieves superior performance compared to baselines.

[1]The project is available at `https://www.gdeltproject.org/`.

2 Related Work

2.1 Knowledge Enhancement

Knowledge enhancement (KE) aims to improve model reasoning, particularly in contexts with limited data, by incorporating external knowledge (Cohen et al., 2024; Arora et al., 2023). KE approaches are broadly classified as structured or unstructured. Structured KE leverages expert-curated Knowledge Bases (KBs) with semantic hierarchies. For instance, task-relevant commonsense knowledge from ConceptNet has been introduced for reasoning support (Liu et al., 2020), and entity linking has been utilized to align textual entities with KB counterparts, enhancing semantic understanding (Yu et al., 2022; Zhang et al., 2019). While offering strong task focus and interpretability due to stable structured knowledge, this approach is constrained by the predefined KBs, hindering adaptability to diverse, open-domain tasks and lacking real-world dynamism and specificity.

In contrast, unstructured KE provides greater flexibility by drawing knowledge from broader open corpora. Methods include encoding task-relevant text like Wikipedia for stance detection (He et al., 2022), and fine-tuning PLMs on domain corpora to embed domain knowledge (Beltagy et al., 2019; Nguyen et al., 2020). Beyond fine-tuning, approaches span self-generating common sense via prompts (Shwartz et al., 2020) to retrieval-based approaches retrieving knowledge from feature pools (Karpukhin et al., 2020) or online sites (Yao et al., 2022). Recent trends integrate data synthesis with Retrieval-Augmented Generation (RAG), Supervised Fine-Tuning (SFT), and Continual Pre-Training (CPT) for improved knowledge infusion (Zhang et al., 2024b). However, challenges remain, such as performance degradation due to irrelevant knowledge and limited transparency in the injection process, underscoring the need for more robust and interpretable unstructured KE frameworks.

2.2 Stance Detection

Stance detection, the task of discerning a user's attitude (e.g., positive, negative, or neutral) towards a given topic or target, is a critical task for understanding subjective opinions expressed in text. To effectively perform stance detection, especially in

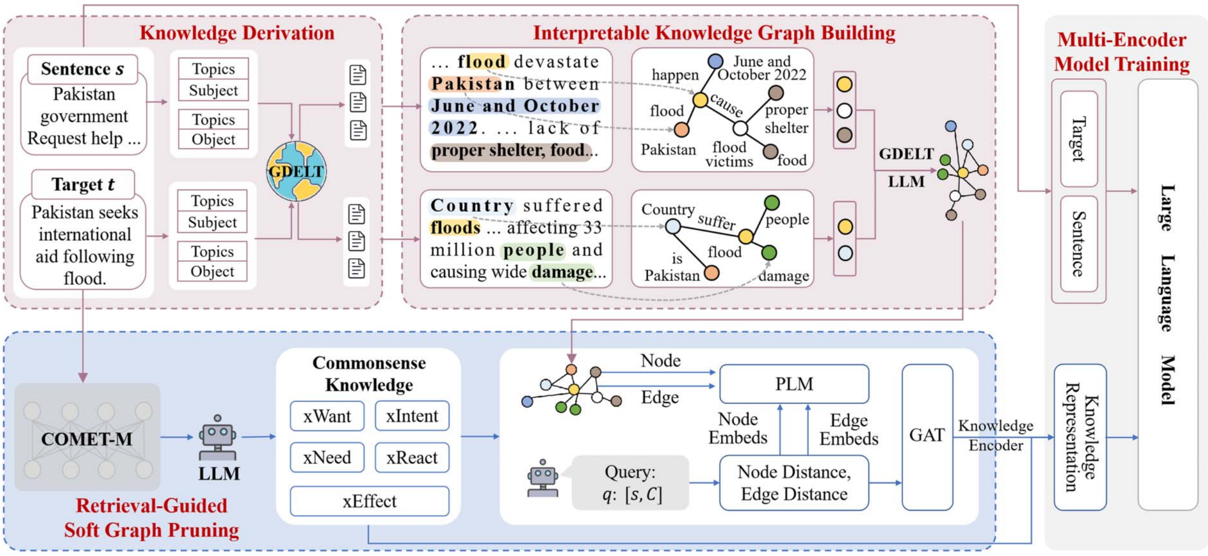

Figure 2: Workflow of our PiKGL framework, which consists of knowledge derivation, interpretable knowledge graph building, retrieval-guided soft graph pruning, and multi-encoder model training.

complex scenarios, models increasingly rely on the integration of external knowledge.

Integrating external knowledge is a key direction in stance detection. Previous studies leverage common sense from ConceptNet (Speer et al., 2017; Liu et al., 2021) and inject sentiment via graph-based modeling and cross-attention (Luo et al., 2022). Wikipedia is also widely used through content encoding (He et al., 2022), retrieval (Zhu et al., 2022), or hybrid methods combining episodic and generated discourse knowledge (Li et al., 2023). More recently, Chen et al. (2024a) proposed a unified framework for joint encoding of external knowledge and sentiment from Wikipedia and ConceptNet, and Zhang et al. (2024c) explored LLM-generated contextual knowledge with prototype contrastive loss to refine stance representation alignment. Despite the demonstrated performance improvements achieved through knowledge-enhanced methods, significant challenges persist in realizing both relevant and generalizable knowledge integration, alongside efficient knowledge filtering. Specifically, ensuring contextual pertinence, generalization to unseen targets, and mitigating redundancy to enhance reasoning remain critical areas requiring further investigation into advanced knowledge infusion strategies for stance detection.

3 Methodology

In this section, we first present the definition of the task and then introduce our novel PiKGL framework for explainable stance detection. As illustrated in Figure 2, PiKGL comprises four primary components: 1) **Knowledge Derivation**, which extracts event triplets and topic to obtain knowledge. 2) **Interpretable Knowledge Graph Building**, which constructs an enriched interpretable knowledge graph by open information extraction and degree-prioritized graph connection. 3) **Retrieval-Guided Soft Graph Pruning**, which incorporates target-specific commonsense knowledge to filter redundant nodes. 4) **Multi-Encoder Model Training**, which jointly encodes knowledge and text for stance detection.

3.1 Task Definition

Given a labeled dataset $X = \{(s_i, t_i, y_i)\}_{i=1}^{N}$, where s_i denotes a sentence instance (e.g., a post or a comment), t_i represents the corresponding target, and $y_i \in \{\texttt{FAVOR}, \texttt{AGAINST}, \texttt{NEUTRAL}\}$ is the stance label, the goal of stance detection is to infer the stance attitude y for a sentence-target pair (s, t). Formally, the objective is to learn a mapping function: $f : (s, t) \rightarrow y$, and apply it to previously unseen pairs at test time.

3.2 Knowledge Derivation

To obtain more authentic knowledge, we retrieve relevant knowledge from the GDELT database for both the input sentence and its corresponding target. GDELT offers a real-time snapshot of the world's collective events, enabling the exploration of complex patterns and relationships within global society. By analyzing this data, it becomes possible to track the evolution of contentious issues and understand the broader context of public sentiment, which is crucial for providing a deeper and more accurate background for stance detection. Given a sentence $s \in X$, we first extract event triplets using an LLM $\mathcal{M}$, denoted as $u = \{(a_j, o_j, e_j)\}_{j=1}^{n}$, where n is the number of event triplets, and a, o, e denote the event subject, event trigger, and event object. To further refine the scope of retrieval, we employ BERTopic (Grootendorst, 2022) to generate a candidate topic list $\mathcal{T}$ from s and utilize the LLM to select the most relevant topic $\hat{\mathcal{T}}$:

$$\mathcal{T} = \text{BERTopic}(s) \tag{1}$$

$$\hat{\mathcal{T}} \leftarrow \mathcal{M}(s, \mathcal{T}) \tag{2}$$

For each triplet in u, we form two topic-aware queries by concatenating $\hat{\mathcal{T}}$ first with the event subject and then with the event object. Using the queries, we retrieve the top-k relevant texts from the GDELT database, obtaining subject-based and object-based text sets. We then textually combine these two text sets to produce the triplet-level retrieved corpus for the current triplet.

Finally, we textually aggregate all these triplet-level retrieved corpus across u to form the complete event-aware knowledge $\mathcal{K}^s$ for the sentence s. In particular, if the extracted triplet set u is empty, the resulting event-aware knowledge $\mathcal{K}^s$ is also defined as empty. Analogously, we apply the same process to the target t to obtain $\mathcal{K}^t$.

3.3 Interpretable Knowledge Graph Building

This section proposes a structured method for constructing an interpretable knowledge graph that goes beyond pure embedding similarity by explicitly modeling the relational structure among nodes. Throughout this section, knowledge denotes structured relational information extracted via open information extraction (OIE) from the retrieved corpus described in Section 3.2, which we represent as an interpretable knowledge graph. We begin by constructing two event-aware subgraphs corresponding to the sentence and the target through OIE, and then introduce a degree-prioritized connection mechanism to efficiently bridge these subgraphs, thereby capturing rich semantic and contextual relationships.

Open Information Extraction. Inspired by Zhang and Soh (2024), we employ a schema-agnostic OIE approach to semantically structure the retrieved event-aware knowledge. Using few-shot prompting, we guide the LLM to extract relational triplets of the form [subject, relation, object] from the event-aware knowledge $\mathcal{K}^s$ and $\mathcal{K}^t$. Based on the extracted relational triplets, we construct the event-aware knowledge graphs G^s and G^t, where each subject and object in a triplet corresponds to a node, while the relation represents a directed edge linking the two nodes. Notably, the process of relational triplets extraction is distinct from the extraction of u, as it focuses on identifying inter-entity relations from the external news texts rather than capturing trigger-based event structures. Here comes the prompt for OIE:

Prompt for Open Information Extraction

Your task is to transform the given text into a semantic graph in the form of a list of triples. The triples must be in the form of [Entity1, Relationship, Entity2]. In your answer, please strictly only include the triples and do not include any explanation or apologies.

Here are some examples:{few_shot_examples}

Now please extract triplets from the following text.
Text: {input_text}

Degree-Prioritized Graph Connection. To enhance computational efficiency and ensure the semantic relevance of graph integration, we propose a degree-prioritized strategy for connecting the event-aware knowledge graphs. We select the top-p nodes with the highest degrees from G^s and G^t. Formally, the selected node set for G^s is defined as:

$$V^s = \{v \in V(G^s) \mid \deg(v) \geq \deg_{(p)}\} \tag{3}$$

where $V(G^s)$ denotes the set of nodes in G^s, $\deg(v)$ is the degree of node v, and $\deg_{(p)}$ is the degree of the p-th highest-degree node in G^s. Using the same procedure, we obtain the selected node sets V^t from G^t. To establish meaningful connections between the two subgraphs, we consider all node pairs (v_j, v_k) such that $v_j \in V^s$ and $v_k \in V^t$. For each candidate pair, we retrieve news texts $l_{j,k}$ from GDELT database $\mathcal{D}$:

$$l_{j,k} = \mathcal{D}(v_j \oplus v_k) \tag{4}$$

Here, $\oplus$ denotes concatenating the texual lables of two nodes into a single query string. We then pass $l_{j,k}$ to the LLM, which assesses whether a valid semantic relationship exists between v_j and v_k, that is, an explicit relation that can be extracted from the text to link the two entities. If such a relation is identified, an edge is established; otherwise, the nodes remain unconnected. Notably, in cases where v_j and v_k refer to the same underlying entity, we still connect them to preserve the structural and semantic completeness of the graph.

Graph Structuring. Through the above process, we effectively link nodes across G^s and G^t, leading to the construction of a unified interpretable knowledge graph $G = (V, E)$: **(1) Nodes:** Each node $v \in V$ represents an event entity, which are transform T_v into the vector representation $h^v \in \mathbb{R}^d$ via PLM. **(2) Edges:** Each edge $e = (v_j, r, v_k) \in E$ captures the semantic relationship r between entities. The same PLM is used to obtain its vector representation $h^e \in \mathbb{R}^d$. Here, d denotes the PLM's embedding size.

3.4 Retrieval-Guided Soft Graph Pruning

Commonsense Knowledge Generation. To reveal the implicit meanings within the claim-based target t and enhance the model's understanding of its context and semantics, we utilize COMET-M (Ravi et al., 2023) to generate target-centered commonsense knowledge. Inspired by Sabour et al. (2022), we introduce a set $\mathcal{R}$, including five special relation tokens `[xReact]`, `[xWant]`, `[xNeed]`, `[xIntent]`, `[xEffect]`. For each token $r \in \mathcal{R}$, we input the target t and the relation token r into COMET-M to generate a commonsense inference c_r:

$$c_r = \text{COMET-M}(t, r) \tag{5}$$

We collect all such inferences into a commonsense knowledge set $\mathcal{C} = \{c_r \mid r \in \mathcal{R}\}$.

Soft Graph Pruning. To suppress irrelevant or noisy substructures, we introduce a soft retrieval-guided pruning mechanism based on semantic proximity to both the source sentence and the generated commonsense knowledge. Specifically, we encode the input text s and commonsense knowledge $\mathcal{C}$ using PLM to obtain query vectors:

$$h^s = \text{PLM}(s) \in \mathbb{R}^d, h^c \quad = \text{PLM}(\mathcal{C}) \in \mathbb{R}^{5\times d} \tag{6}$$

where h^s denotes the text query vector. h^c is the stacked commonsense query vectors corresponding to the five relation tokens. Using these query vectors, we perform a distance-driven pruning operation as follows:

$$\begin{aligned} w^v &= \text{MLP}\phi_1[\text{dist}(h^v, h^s) + \text{dist}_{\text{avg}}(h^v, h^c)], \\ w^e &= \text{MLP}\phi_2[\text{dist}(h^e, h^s) + \text{dist}_{\text{avg}}(h^e, h^c)] \end{aligned} \tag{7}$$

where $\text{dist}(\cdot)$ denotes the Euclidean distance between the node or edge representation and a specific query vector h^s, while $\text{dist}_{\text{avg}}(\cdot)$ refers to the average Euclidean distance to all commonsense query vectors h^c. The weights w^v and w^e are then computed by a Multilayer Perceptron (MLP), capturing both specific and generalized similarity signals for pruning, where ϕ_1 and ϕ_2 indicate independent sets of trainable parameters for the MLPs.

Update. We use a Graph Attention Network (GAT) to aggregate the graph's topological and semantic information, while the computed weights guide the message passing. At each layer l, the representation of node v is updated as:

$$h^{v(l)} = \text{GATConv}^{(l)}(w^v \cdot h^{v(l-1)}, w^e \cdot h^e) \tag{8}$$

where $h^{v(l-1)}$ is the node representation from the previous layer, and h^e is the corresponding edge features. The weighted node and edge features are aggregated through GAT to compute the updated node embedding $h^{v(l)}$.

Finally, we obtain a unified knowledge representation ξ from the updated node and edge embeddings. Let L_g denote the total number of GAT layers. We define the final GAT-updated node and edge representations as:

$$H^v = h^{v(L_g)}, \quad H^e = w^e \cdot h^e \tag{9}$$

where $H^v \in \mathbb{R}^{n\times d_g}$ and $H^e \in \mathbb{R}^{m\times d_g}$, with n and m denoting the number of nodes and edges in the graph, and d_g representing the hidden dimension of GAT outputs. These features are then projected into the LLM embedding space through learnable linear transformations:

$$\tilde{H}^v = \text{Proj}_v(H^v), \quad \tilde{H}^e = \text{Proj}_e(H^e) \tag{10}$$

where $\tilde{H}^v \in \mathbb{R}^{n\times d_{\text{LLM}}}$ and $\tilde{H}^e \in \mathbb{R}^{m\times d_{\text{LLM}}}$, with d_{LLM} representing the embedding dimension of the

LLM. The projected node and edge embeddings are concatenated along the sequence dimension and passed through a Transformer-based encoder to generate the final structured knowledge representation:

$$\xi = f_{\text{enc}}(\tilde{H}^v \oplus \tilde{H}^e) \in \mathbb{R}^{L \times d_{\text{LLM}}} \tag{11}$$

Here, $\oplus$ denotes the concatenation operation, f_{enc} is a multi-layer Transformer encoder, and L is the total number of output token representations. The resulting ξ serves as a graph-aware knowledge sequence aligned with the input format of the LLM.

3.5 Multi-Encoder Model Training

To effectively integrate interpretable knowledge ξ and commonsense knowledge C, we leverage a Multi-Head Attention (MHA) mechanism as a knowledge encoder f_{ke}. Specifically, we first concatenate the token sequences of ξ and C to form a single, combined input sequence. The MHA module then operates on this sequence, enabling each token to attend to all other tokens. This facilitates a powerful form of cross-attention, where, for example, entity tokens from the graph knowledge can directly weigh the importance of and draw context from the commonsense inferences, and vice-versa. This process adaptively fuses the two knowledge types into a single, comprehensive, and nuanced representation h^k:

$$h^k = f_{ke}(\xi \oplus C) \tag{12}$$

Finally, the representation is combined with the text feature h^s and the target feature h^t, and then fed into an LLM to produce stance-aware features h. And the stance logits are then predicted by a linear classifier:

$$h = \text{LLM}([h^s; h^t; h^k]) \tag{13}$$

$$\hat{y} = \text{softmax}(Wh + b) \tag{14}$$

where $\hat{y}$ denotes the predicted probabilities for the stance label, $W \in \mathbb{R}^{d_l \times d_{LLM}}$ and $b \in \mathbb{R}^{d_l}$ are trainable parameters, and d_l denotes the number of stance label classes. The model is optimized by minimizing the cross-entropy loss function:

$$\underset{\Theta}{min} \mathcal{L} = -\sum_{i=1}^{N}\sum_{j=1}^{d_l} y_i^j \log \hat{y}_i^j + \lambda \|\Theta\|^2 \tag{15}$$

where y_i^j and $\hat{y}_i^j$ are the ground truth and predicted probability for the j-th class of the instance i, respectively. λ is a regularization hyperparameter, and Θ denotes the model parameters.

4 Experiment

4.1 Experimental Setups

We employ LLaMA-8B as the backbone LLM for multi-encoder model training, trained using the Adam optimizer with a learning rate of 8e^{-5}, and a batch size of 32 for all datasets. For subject-based knowledge acquisition and high-degree node selection, we set k and p to 3. In the case of noun phrase targets, the target itself is used as the commonsense query. Topic selection, event extraction, and OIE are performed using GPT-4o-mini. Graph elements are embedded using the ''all-MiniLM-L6-v2'' (384-dimensional), and then aligned with the LLM embedding space via MLP-based projection. The knowledge encoder utilizes 8 attention heads, and the regularization coefficient is set to $\lambda = 1 \times 10^{-5}$. All experiments are conducted on a single RTX A6000 GPU.

4.2 Dataset and Metrics

Dataset. We evaluate PiKGL on three datasets: **EZ-Stance** (Zhao and Caragea, 2024), a zero-shot English dataset comprising 40,678 targets and 47,316 tweets with both noun phrases and claims; **C-Stance** (Zhao et al., 2023), a Chinese dataset collected from Sina Weibo with 40,204 targets and 48,126 microblogs; and **P-Stance** (Li et al., 2021), which focuses on English tweets related to Joe Biden, Bernie Sanders, and Donald Trump. Following prior work, samples labeled as ''None'' are excluded due to annotation inconsistencies.

Metric. We report the Macro-averaged F1 score on EZ-Stance and P-Stance. For C-Stance, we additionally report the mean F1 score averaged across stance labels, in line with previous studies (Zhao and Caragea, 2024; Zhao et al., 2023).

4.3 Baselines

We compare our method with four baselines: 1) **Neural Network-based Models**: BiCE (Augenstein et al., 2016), CrossNet (Xu et al., 2018). 2) **Pre-trained Language Models**: BART-MNLI-e (Zhao and Caragea, 2024), BART-MNLI-e$_p$ (Zhao and Caragea, 2024), TGA-Net (Allaway and McKeown, 2020), BERT (Devlin

Model	C-Stance Claim				C-Stance Noun				EZ-C	EZ-N
	Against	Favor	Neutral	All	Against	Favor	Neutral	All		
BiCE	33.50	35.80	30.20	33.20	56.00	51.50	59.00	55.50	31.60	52.90
CrossNet	44.10	39.50	58.80	47.50	60.70	56.70	60.10	59.20	52.30	55.10
BERT	79.70	82.70	89.90	84.10	70.80	69.30	64.70	68.30	–	–
RoBERTa	79.70	81.90	89.90	83.80	71.20	70.10	66.90	69.40	85.60	65.60
XLNet	80.50	82.90	90.00	84.50	72.10	70.10	66.70	69.60	–	–
BART-MNLI-e	–	–	–	–	–	–	–	–	88.80	67.50
BART-MNLI-e_p	–	–	–	–	–	–	–	–	–	68.70
TGA-Net	48.80	62.50	69.90	60.40	69.40	67.40	67.00	67.90	59.60	60.60
Mistral-7B	42.63	44.60	39.02	42.13	48.77	59.57	42.33	50.22	41.85	53.68
GPT3.5	50.99	44.13	53.75	49.62	63.03	54.62	52.93	56.86	44.00	60.40
LLaMA-8B	58.05	52.78	44.93	51.94	69.57	66.63	46.76	60.98	49.01	52.67
Qwen-7B	69.48	48.37	63.57	60.47	72.01	59.96	64.27	65.42	44.57	60.59
Ours	**91.78***	**92.82***	**93.37**	**92.61***	**80.07***	**77.75***	**67.69**	**75.18***	**91.62***	**70.06***

Table 1: Performance (%) on C-Stance and EZ-Stance datasets in zero-shot setting. Bold indicates the best results. ''–'' indicates that the original paper does not report results for the corresponding setting. EZ-C and EZ-N denote the Claim-based and Noun-based settings of the EZ-Stance dataset.

et al., 2019), RoBERTa (Liu, 2019), XLNet (Yang et al., 2019). 3) **Large Language Models**: Qwen-7B (Yang et al., 2024), LLaMA-8B (Dubey et al., 2024), Mistral-7B (Jiang et al., 2023), GPT3.5 (Zhang et al., 2022), FACTUAL-LLaMA2, FACTUAL-ChatGPT. We prompt these models to obtain the results. 4) **Wikipedia Knowledge-based Models**: WS-BERT (He et al., 2022), KASD-ChatGPT, KASD-BERT, KASD-LLaMA2 (Li et al., 2023).

4.4 Main Results

To assess the effectiveness of our proposed approach, the main experimental results are reported in Table 1 and Table 2. The paired t-test ($p < 0.05$, marked with *) is conducted only against baselines for which we reproduced results under the same data splits and obtained per-instance predictions. Baselines reported with '–' indicate results not given in the original papers and are excluded from the significance test. We observe that our PiKGL consistently achieves superior performance across most datasets, confirming the effectiveness of our framework in improving stance understanding and reasoning. Furthermore, pre-trained language models such as BERT and RoBERTa generally outperform LLMs across all datasets, highlighting the advantages of fine-tuning in stance detection tasks. Notably, PiKGL not only excels on English datasets but also achieves substantial gains on the C-Stance dataset. As C-Stance is derived from Chinese social media, it complements existing work that primarily focuses on English contexts. The strong performance of our framework in this setting demonstrates its robust generalization across both languages and platforms, showcasing its adaptability in diverse real-world scenarios.

Model	Trump	Biden	Sanders	Average
RoBERTa	82.70	84.29	79.56	82.18
WS-BERT	85.80	83.50	79.00	82.77
KASD-LLaMA2	79.59	71.32	67.89	72.93
KASD-ChatGPT	85.06	84.59	79.96	83.20
KASD-BERT	85.35	85.66	80.39	83.80
Qwen-7B	72.55	79.81	76.15	76.13
Mistral-7B	76.99	79.04	73.91	76.64
LLaMA-8B	76.45	78.79	76.73	77.32
FACTUAL-GPT3.5	84.95	86.03	81.60	84.20
FACTUAL-LLaMA2	85.58	86.34	**83.06**	84.99
PiKGL(Ours)	**90.94***	**88.51***	82.66	**87.37***

Table 2: Results (%) on P-Stance in in-target setting.

Moreover, we further evaluate the performance of the PiKGL framework in the in-target setting using the P-Stance dataset. As shown in Table 2, PiKGL consistently outperforms all baseline methods, confirming its effectiveness under standard evaluation settings. This success is driven by our core methodology: by structuring information into an interpretable knowledge graph and then applying a retrieval-guided pruning mechanism, PiKGL distills raw data into precise and coherent knowledge. This refined knowledge, derived from GDELT's timely news events, provides

Number	Configuration	Description
(I)	Default	1.Apply text s and target t to fine-tune LLM
(II)	(I)+GDELT	1.Apply **Knowledge Derivation** to retrieve event-aware knowledge $\mathcal{K}^s, \mathcal{K}^t$. 2.Feed s, t, into LLM.
(III)	GDELT + RAG	1.Apply **Knowledge Derivation** to retrieve event-aware knowledge $\mathcal{K}^s, \mathcal{K}^t$. 2.Use RAG to select a target-relevant knowledge $\mathcal{K}^r$ from $\mathcal{K}^s, \mathcal{K}^t$. 3.Feed s, t, and $\mathcal{K}^r$ into LLM.
(IV)	(II) + COMET-M	1.Apply **COMET-M** to get commonsense knowledge C for the target t. 2.Feed s, t and C into LLM.
(V)	(IV) + Interpretable Knowledge Graph Building	1.Apply **Interpretable Knowledge Graph Building** to convert $\mathcal{K}^s, \mathcal{K}^t$ into an interpretable knowledge graph G. 2.Feed s, t, C and G into LLM.
(VI)	(V) + Retrieval-Guided Soft Graph Pruning	1.Apply **Retrieval-Guided Soft Graph Pruning** to get graph-aware knowledge representation ξ. 2.Feed s, t, C, and ξ into LLM.
(VII)	(VI) + Knowledge Encoder	1.Apply **Knowledge Encoder** to form knowledge representation h^k based graph-aware knowledge ξ and C. 2.Feed s, t, and h^k into Multi-Encoder Model.

Table 3: Model configurations of the ablation study. Our vital components are bold.

Model	C-Stance Claim				C-Stance Noun				P-Stance				EZ-C	EZ-N
	Ag	Fa	Ne	All	Ag	Fa	Ne	All	Trump	Biden	Sanders	Avg.		
(I)	87.69	88.35	89.04	88.36	76.45	76.16	63.89	72.17	87.68	85.97	81.13	84.93	86.71	66.87
(II)	89.10	88.56	90.04	89.23	77.86	76.51	63.83	72.73	88.43	86.01	81.19	85.21	88.04	67.69
(III)	89.39	87.51	91.70	89.53	78.39	77.02	64.47	73.29	88.91	86.74	81.47	85.71	88.92	68.09
(IV)	89.76	88.83	91.21	89.93	–	–	–	–	–	–	–	–	88.75	–
(V)	90.80	89.32	91.22	90.45	79.04	76.10	64.94	73.36	88.70	86.88	81.94	85.87	89.14	68.16
(VI)	91.59	92.31	93.26	92.39	79.61	77.73	66.20	74.63	90.79	88.50	82.57	87.28	91.33	69.64
Ours (VII)	**91.78**	**92.82**	**93.37**	**92.61**	**80.07**	**77.75**	**67.69**	**75.18**	**90.94**	**88.51**	**82.66**	**87.37**	**91.62**	**70.06**

Table 4: Results (%) of the ablation study. Ag, Fa, and Ne represent Against, Favor, and Neutral, respectively. EZ-C and EZ-N denote the Claim-based and Noun-based settings of the EZ-Stance dataset. Avg. denotes the average performance. ''–'' indicates that commonsense knowledge is not applicable to the Noun-based setting.

clearer advantages in task relevance and contextual appropriateness over the generic definitional information from Wikipedia.

Beyond overall performance, a per-class analysis is critical for understanding PiKGL's nuanced behavior, especially its effectiveness in handling difficult and often-ambiguous classes. Two points underscore our framework's superiority. Firstly, the balanced and significant gains across all classes confirm that PiKGL achieves a holistic improvement in contextual understanding, rather than being biased towards a specific class. Second, our model shows superior performance on the challenging *Neutral* class compared to all baselines, suggesting it is more effective at disambiguating nuanced or implicit opinions.

4.5 Ablation Study

We conducted an ablation study to analyze the contribution of each component in our PiKGL framework. The different model configurations are detailed in Table 3, and the corresponding experimental results are presented in Table 4. An analysis of these results is provided below.

Effect of GDELT Text (I vs. II). The integration of GDELT knowledge (Configuration II) yields a significant performance improvement over the baseline (Configuration I) across all datasets. This result underscores the value of external knowledge, demonstrating that event-aware context significantly enhances the model's understanding and its ability to detect stance accurately.

Effect of RAG (II vs. III). To benchmark our approach against standard retrieval methods, we established a simple RAG baseline (Configuration III). Following Chirkova et al. (2024), Configuration (III) retrieves the top-3 GDELT texts using BGE-M3 (Chen et al., 2024b). While Table 4 shows that this RAG approach provides a marginal improvement over using unfiltered GDELT knowledge (II), its performance is substantially lower than our full framework (VI). This demonstrates that while basic retrieval-based

Model	C-Stance Claim				C-Stance Noun				P-Stance				EZ-C	EZ-N
	Ag	Fa	Ne	All	Ag	Fa	Ne	All	Trump	Biden	Sanders	Avg.		
PiKGL$_L$	91.78	92.82	93.37	92.61	80.07	77.75	67.69	75.18	90.94	88.51	82.66	87.37	91.62	70.06
PiKGL$_Q$	92.21	92.69	92.88	92.59	79.53	76.70	68.24	74.82	89.30	86.62	79.16	85.03	90.98	69.94
PiKGL$_M$	91.62	92.44	93.47	92.50	79.85	77.62	67.41	74.93	90.19	88.66	83.61	87.48	91.59	70.27

Table 5: Performance (%) of different backbones. Ag, Fa, and Ne represent Against, Favor, and Neutral, respectively. EZ-C and EZ-N denote Claim-based and Noun-based setting of EZ-Stance dataset.

filtering is beneficial, our proposed graph-based pruning and integration method is far more effective than a standard RAG for knowledge selection and utilization.

Effect of Commonsense Knowledge (II vs. IV). Configuration (IV) enhances the model by adding commonsense knowledge generated by COMET-M for the given target. The performance improvement over (II) confirms that for claim-based targets, this external commonsense knowledge is useful for the model to capture the implicit meanings and context of the target, leading to better stance comprehension.

Effect of Interpretable Knowledge Graph Building (IV vs. V). In configuration (V), we convert the textual knowledge into an interpretable knowledge graph. The subsequent performance gain over (IV) shows that providing knowledge to the model in a structured format helps it better understand the complex relationships between entities and events. This structured representation also implicitly filters some less important textual information compared to feeding raw text.

Effect of Soft Graph Pruning (V vs. VI). The addition of our retrieval-guided soft graph pruning mechanism in (VI) results in a notable performance increase. This component serves two purposes: first, by assigning weights based on semantic proximity to the source text and the target, it effectively suppresses irrelevant and noisy information within the knowledge graph. Second, it acts as a crucial bridge that aligns the external knowledge with the specific context of the stance detection task, ensuring the information is highly relevant to both the text and the target.

Effect of Knowledge Encoder (VI vs. VII). Our full model, configuration (VII), includes a knowledge encoder that uses a Multi-Head Attention mechanism to fuse the pruned graph-aware knowledge representation ξ and the commonsense knowledge C. The final performance boost demonstrates the effectiveness of this module. It adaptively integrates different forms of knowledge, creating a comprehensive and nuanced representation that empowers the model to make more accurate stance predictions.

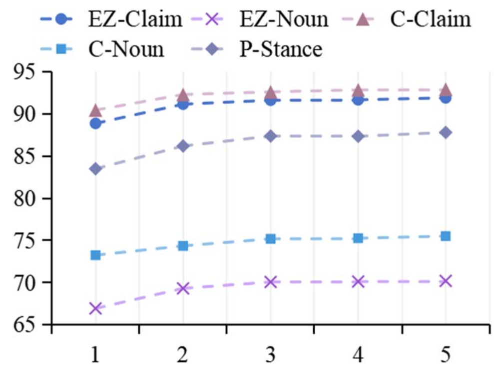

Figure 3: Average performance (%) for different p values, where graphs are constructed using the top-p high-degree nodes ($p = 1$–5).

4.6 Effect of Different Backbones

We evaluate the performance of our PiKGL framework with alternative backbones for multi-encoder model training phrase, including Qwen2.5-7B and Mistral-7B. Experimental results shown in Table 5 demonstrate that PiKGL consistently surpasses all baseline methods across the evaluated datasets, highlighting its robustness and strong generalization ability across different model architectures. Notably, Mistral-7B exhibits slightly better performance on certain datasets, which we attribute to its superior capacity for modeling long-context inputs, particularly when leveraging external knowledge from GDELT news content.

4.7 Effect of Different Degree

To investigate the impact of the number of top high-degree nodes p used for graph connections on the performance of our PiKGL framework, we conduct experiments by varying p from 1 to 5, as shown in Figure 3. When $p = 1$, performance remains stable, likely due to the relevance of the initial subject- and object-based event-aware

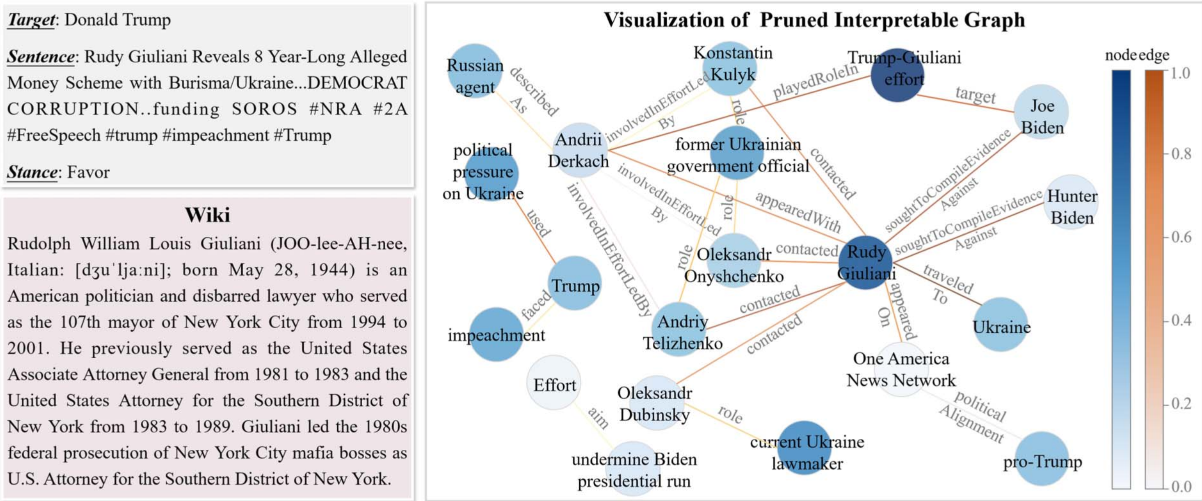

Figure 4: Visualization of wiki data and our pruned interpretable knowledge graph, where darker node and edge colors indicate higher weights, reflecting greater node importance and stronger associations between nodes.

knowledge. As p increases, the model benefits from richer contextual information, leading to improved performance. However, beyond a certain threshold, the gains begin to saturate, indicating that excessive node connections may introduce redundant information and increase computational complexity. To balance effectiveness and efficiency, we set $p = 3$ in our final experiments.

4.8 Case Study

Figure 4 presents a case study from the P-Stance dataset, contrasting our pruned knowledge graph with generic Wikipedia data. As shown in Figure 4, we construct an event graph centered around Rudy Giuliani, capturing his relationships with key entities and organizations involved in the surrounding events. Centered on entities such as *Konstantin Kulyk*, *Oleksandr Onyshchenko*, *Andriy Telizhenko*, and *One America News Network*, the graph captures relevant entities and relationships. A retrieval-guided pruning mechanism assigns semantic similarity-based weights to nodes and edges, filtering out less relevant information.

Furthermore, our PiKGL method exhibits strong interpretability. Each node and edge in the pruned interpretable knowledge graph is derived from real-world news events and assigned explicit weights based on semantic similarity. This design ensures that the model's reasoning process is traceable. Users can intuitively understand how external knowledge influences the model's decisions, thereby enhancing the transparency and credibility of the model in stance detection tasks.

4.9 Error Analysis

We performed a qualitative error analysis with representative cases shown in Figure 5. Our analysis reveals two primary issues. The first issue, shown in the top case, is a failure to establish a common topic. The model treats ''Arkansas'' and ''Kansas'' as separate subjects, overlooking their shared context of ''NCAA postseason''. This led to the retrieval of disconnected evidence, causing the model to default to a ''None'' prediction and miss the implicit ''Against'' relationship. The second issue, shown in the bottom case, is incomplete event extraction. The model fails to extract the key term ''Warren Plan'' when generating event triplets. Consequently, the model works with an incomplete information profile, causing it to hallucinate a ''Favor'' prediction.

4.10 Computational Efficiency Analysis

The computational overhead of the PiKGL framework is primarily concentrated in two stages. The initial **interpretable knowledge graph building** stage is highly efficient, comprising Open Information Extraction with a complexity of $O(n_t)$, where n_t is the number of retrieved texts) and a degree-prioritized graph connection mechanism at $O(p^2)$. The **retrieval-guided graph pruning** stage involves multiple steps: structuring the graph (nodes n and edges m) into embeddings with a

Input	Topic	Entity	GDELT Content	Label
Sentence: Entering regionals, Arkansas has played 25 games against 10 teams that made the NCAA postseason. The Razorbacks are 13-12 in those games.	Arkansas	[Arkansas, played, games] [The Razorbacks, are, 13-12 in those games]	The article previews the upcoming college basketball matchup between the Arkansas Razorbacks and the LSU Tigers, highlighting Arkansas's strong late-game performance under coach Eric Musselman despite frequent slow starts. With two consecutive Elite Eight appearances...	**Ground Truth:** Against
Target: Kansas has played 10 games against 5 NCAA postseason teams, 87-71 in those games.	Kansas	[Kansas, played, postseason teams]	The article recaps the season-ending rivalry clash between the Kansas Jayhawks and Kansas State Wildcats, where Kansas once again fell short of breaking its long losing streak in the series. Kansas State capitalized on Kansas's...	**Predicted Label:** None
Sentence: Any time, a Democrat starts getting horrible publicity, Democrats should implement a Warren Plan to shut them up. A big one. Like a Wealth Tax exec order or some such.	Democrat	[Democrats, starts getting, publicity] [Democrats, should implement, Wealth Tax exec order]	The article explores the Republican Party's long-running practice of referring to the Democratic Party as the "Democrat Party," a rhetorical tactic dating back to the 1940s. Popularized by GOP leader B. Carroll Reece, the term was used to suggest the party had abandoned its...	**Ground Truth:** None
Target: Implementing a Warren Plan during times of negative publicity will be seen as a bold move by Democrats to shift the narrative and regain control of public opinion.	Democrat	[Democrats, shift, narrative]	The article covers Representative Jim Himes's appointment as the top Democrat on the House Intelligence Committee, replacing Rep. Adam Schiff after Speaker Kevin McCarthy barred Schiff from serving. House Minority Leader Hakeem Jeffries announced Himes's selection alongside other key...	**Predicted Label:** Favor

Figure 5: Error analysis with two examples.

dimension of d has a cost of $O((n + m) \cdot d)$, assigning weights based on r commonsense cues requires $O((n + m) \cdot r \cdot d)$, and a GAT-based update aggregates information with a complexity of $O(L_g \cdot (n+m) \cdot d_g)$. The most computationally intensive step is the final Transformer encoder, which applies self-attention with a quadratic complexity of $O((n + m)^2 \cdot d_{LLM})$.

In practice, our design choices lead to significant efficiency gains. The degree-prioritized graph connection mechanism reduces the average graph construction and connection time to approximately 18.1 seconds per input, a dramatic improvement over the 103 minutes required by a direct construction approach. Furthermore, the entire inference process is highly efficient, requiring only about 0.96 seconds.

5 Discussion

Definition of Explainability and Transparency. To build a more trustworthy model, PiKGL operationalizes the critical concepts of explainability and transparency. Explainability, the ability to understand why a prediction is made, is achieved through its interpretable knowledge graph, where human-readable nodes and edges trace back to real-world GDELT events. The model's reasoning is made explicit and visualizable by a weighted pruning mechanism that quantifies the importance of each piece of evidence. Transparency, the clarity of how the model functions, is ensured by its decomposable modular architecture, its use of a verifiable public knowledge source (GDELT), and an ablation study empirically validating each component's contribution. By integrating these principles, PiKGL not only achieves excellent performance but also offers a more scrutable and reliable framework for stance detection.

Model	Trump	Biden	Sanders	Average
PiKGL(Ours)	90.94	88.51	82.66	87.37
LLaMA-implement	90.44	88.30	81.98	86.90

Table 6: Performance (%) on the Open-Source Implementation.

Reproducibility and Model Independence. To test the framework's reproducibility and independence from proprietary models, we substituted GPT-4o-mini with the open-source LLaMA3-8B on the P-Stance dataset. As shown in Table 6, the open-source implementation achieved nearly identical performance. This demonstrates that the effectiveness of PiKGL originates from its architecture, such as the knowledge graph construction and pruning, rather than the specific LLM used for extraction.

Limitations. Our PiKGL framework achieves excellent stance detection by constructing and pruning an interpretable knowledge graph from the real-time GDELT news database. Although this approach is highly effective for capturing dynamic, event-related context, there is an inherent limitation in relying solely on GDELT, whose news-media focus and source demographics introduce potential coverage and linguistic biases.

6 Conclusion

This paper introduces PiKGL, a novel stance detection framework that leverages real-world knowledge from the GDELT project to construct an interpretable knowledge graph using degree-based graph connections. To ensure the relevance and precision of the knowledge, a pruning mechanism is employed to filter out irrelevant information while retaining target-specific and contextual commonsense knowledge. This enhances the model's ability to make informed and interpretable stance predictions. Experimental results on three benchmark datasets show that PiKGL achieves state-of-the-art performance in both zero-shot and in-target stance detection tasks, demonstrating its effectiveness and generalization.

Acknowledgments

This work was supported by the National Natural Science Foundation of China 62176076 and 62576120, the Major Key Project of PCL2023A09, CIPSC-SMP-ZHIPU Large Model Cross-Disciplinary Fund ZPCG20241119405 and Key Laboratory of Computing Power Network and Information Security, Ministry of Education under grant No. 2024ZD020, the Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (2022B1212010005).

References

Emily Allaway and Kathleen 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)*, pages 8913–8931, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.717

Simran Arora, Patrick Lewis, Angela Fan, Jacob Kahn, and Christopher Ré. 2023. Reasoning over public and private data in retrieval-based systems. *Transactions of the Association for Computational Linguistics*, 11:902–921. https://doi.org/10.1162/tacl_a_00580

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885. https://doi.org/10.18653/v1/D16-1084

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. 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 3615–3620. https://doi.org/10.18653/v1/D19-1371

Hongzhou Chen, Ke Yan, Mustafa Raad Kadhim, Kui Wu, and Ling Tian. 2024a. Sentkb-bert: Sentiment-filtered knowledge-based stance detection. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE. https://doi.org/10.1109/IJCNN60899.2024.10651178

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024b. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Nadezhda Chirkova, David Rau, Hervé Déjean, Thibault Formal, Stéphane Clinchant, and Vassilina Nikoulina. 2024. Retrieval-augmented generation in multilingual settings. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 177–188. https://doi.org/10.18653/v1/2024.knowllm-1.15

Wang Chunling, Zhang Yijia, Yu Xingyu, Liu Guantong, Chen Fei, and Lin Hongfei. 2023. Adversarial network with external knowledge for zero-shot stance detection. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 824–835. https://doi.org/10.1007/978-981-99-6207-5_26

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. 2024. Evaluating the ripple effects of knowledge editing in language models. *Transactions of the Association for Computational Linguistics*, 12:283–298. https://doi.org/10.1162/tacl_a_00644

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. https://doi.org/10.18653/v1/N19-1423

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Zhenyu He, Ce Li, Fan Zhou, and Yi Yang. 2021. Rumor detection on social media with event augmentations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2020–2024. https://doi.org/10.1145/3404835.3463001

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from wikipedia to enhance stance detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 71–77. https://doi.org/10.18653/v1/2022.wassa-1.7

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Yan Jiang, Jinhua Gao, Huawei Shen, and Xueqi Cheng. 2022. Few-shot stance detection via target-aware prompt distillation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 837–847. https://doi.org/10.1145/3477495.3531979

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen Tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 6769–6781. Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2020.emnlp-main.550

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia potentia est–on the role of knowledge in computational argumentation. *Transactions of* *the Association for Computational Linguistics*, 10:1392–1422. https://doi.org/10.1162/tacl_a_00525

Kalev Leetaru and Philip A. Schrodt. 2013. Gdelt: Global data on events, location, and tone, 1979–2012. In *ISA Annual Convention*, volume 2, pages 1–49. Citeseer.

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*, pages 15703–15717, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.972

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. https://doi.org/10.18653/v1/2021.findings-acl.208

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. https://doi.org/10.18653/v1/2021.findings-acl.278

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908. https://doi.org/10.1609/aaai.v34i03.5681

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Yun Luo, Zihan Liu, Yuefeng Shi, Stan Z. Li, and Yue Zhang. 2022. Exploiting sentiment and common sense for zero-shot stance detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7112–7123.

Dat Quoc Nguyen, Thanh Vu, and Anh-Tuan Nguyen. 2020. Bertweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14. https://doi.org/10.18653/v1/2020.emnlp-demos.2

Sahithya Ravi, Raymond Ng, and Vered Shwartz. 2023. COMET-M: Reasoning about multiple events in complex sentences. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12921–12937, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.861

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11229–11237. https://doi.org/10.1609/aaai.v36i10.21373

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629. https://doi.org/10.18653/v1/2020.emnlp-main.373

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31. https://doi.org/10.1609/aaai.v31i1.11164

Penghui Wei and Wenji Mao. 2019. Modeling transferable topics for cross-target stance detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1173–1176. https://doi.org/10.1145/3331184.3331367

Haoyang Wen and Alexander G. Hauptmann. 2023. 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)*, pages 1491–1499. https://doi.org/10.18653/v1/2023.acl-short.127

Chang Xu, Cécile 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. https://doi.org/10.18653/v1/P18-2123

Ming Yan, Tianyi Zhou Joey, and W. Tsang Ivor. 2024. Collaborative knowledge infusion for low-resource stance detection. *Big Data Mining and Analytics*, 7(3):682–698. https://doi.org/10.26599/BDMA.2024.9020021

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32.

Yunzhi Yao, Shaohan Huang, Li Dong, Furu Wei, Huajun Chen, and Ningyu Zhang. 2022. Kformer: Knowledge injection in transformer feed-forward layers. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 131–143. Springer. https://doi.org/10.1007/978-3-031-17120-8_11

Donghan Yu, Chenguang Zhu, Yiming Yang, and Michael Zeng. 2022. Jaket: Joint pre-training of knowledge graph and language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11630–11638. https://doi.org/10.1609/aaai.v36i10.21417

Bowen Zhang, Daijun Ding, Liwen Jing, Genan Dai, and Nan Yin. 2022. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An LLM-based framework for knowledge graph construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9820–9836. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.548

Hao Zhang, Yizhou Li, Tuanfei Zhu, and Chuang Li. 2024a. Commonsense-based adversarial learning framework for zero-shot stance detection. *Neurocomputing*, 563:126943. https://doi.org/10.1016/j.neucom.2023.126943

Jiaxin Zhang, Wendi Cui, Yiran Huang, Kamalika Das, and Sricharan Kumar. 2024b. Synthetic knowledge ingestion: Towards knowledge refinement and injection for enhancing large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21456–21473. https://doi.org/10.18653/v1/2024.emnlp-main.1196

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451. https://doi.org/10.18653/v1/P19-1139

Zhao Zhang, Yiming Li, Jin Zhang, and Hui Xu. 2024c. Llm-driven knowledge injection advances zero-shot and cross-target stance detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 371–378. https://doi.org/10.18653/v1/2024.naacl-short.32

Chenye Zhao and Cornelia Caragea. 2024. EZ-STANCE: A large dataset for English zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15697–15714, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.838

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)*, pages 13369–13385, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.747

Qinglin Zhu, Bin Liang, Jingyi Sun, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. Enhancing zero-shot stance detection via targeted background knowledge. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2070–2075. https://doi.org/10.1145/3477495.3531807