

Article

Distantly Supervised Explainable Stance Detection via Chain-of-Thought Supervision

Daijun Ding ^{1,†}, Genan Dai ^{2,†}, Cheng Peng ³, Xiaojiang Peng ² , Bowen Zhang ^{2,*} and Hu Huang ^{4,*} ¹ College of Applied Science, Shenzhen University, Shenzhen 518052, China; 2100411011@email.szu.edu.cn² College of Big Data and Internet, Shenzhen Technology University, Shenzhen 518118, China; daigenan@sztu.edu.cn (G.D.); pengxiaojiang@sztu.edu.cn (X.P.)³ School of Computing, Zhongshan Institute, University of Electronic Science and Technology of China, Zhongshan 528402, China; pengcheng@zsc.edu.cn⁴ Shenzhen Graduate School, Peking University, Shenzhen 518055, China

* Corresponding: zhang_bo_wen@foxmail.com (B.Z.); h.huang@pku.edu.cn (H.H.)

† These authors contributed equally to this work.

Abstract: Investigating public attitudes on social media is crucial for opinion mining systems. Stance detection aims to predict the attitude towards a specific target expressed in a text. However, effective neural stance detectors require substantial training data, which are challenging to curate due to the dynamic nature of social media. Moreover, deep neural networks (DNNs) lack explainability, rendering them unsuitable for scenarios requiring explanations. We propose a distantly supervised explainable stance detection framework (DS-ESD), comprising an instruction-based chain-of-thought (CoT) method, a generative network, and a transformer-based stance predictor. The CoT method employs prompt templates to extract stance detection explanations from a very large language model (VLLM). The generative network learns the input-explanation mapping, and a transformer-based stance classifier is trained with VLLM-annotated stance labels, implementing distant supervision. We propose a label rectification strategy to mitigate the impact of erroneous labels. Experiments on three benchmark datasets showed that our model outperformed the compared methods, validating its efficacy in stance detection tasks. This research contributes to the advancement of explainable stance detection frameworks, leveraging distant supervision and label rectification strategies to enhance performance and interpretability.

Keywords: stance detection; prompt-tuning; chain-of-thought**MSC:** 68T50

check for updates

Citation: Ding, D.; Dai, G.; Peng, C.; Peng, X.; Zhang, B.; Huang, H.

Distantly Supervised Explainable Stance Detection via Chain-of-Thought Supervision.

Mathematics **2024**, *12*, 1119.<https://doi.org/10.3390/math12071119>

math12071119

Academic Editor: Hsien-Chung Wu

Received: 21 February 2024

Revised: 21 March 2024

Accepted: 27 March 2024

Published: 8 April 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Stance detection is a fundamental task in the field of natural language processing (NLP), where the aim is to categorize the attitudes expressed towards a particular target based on opinionated input texts [1]. This task has garnered significant interest in recent years due to its relevance in various domains, including political analysis, social media monitoring, and customer feedback analysis. In the initial phases of stance detection research, the focus was predominantly on online debates characterized by a uniform sentence structure, wherein the user's attitude is generally expressed in a direct fashion [2,3]. With the rapid expansion of the Internet, platforms like Twitter have witnessed remarkable growth in popularity. This surge has prompted researchers to explore the potential of social media as a rich resource for stance detection [4,5].

Stance detection methods are usually formulated as sentence-level classification tasks based on a specific target and can be broadly categorized as non-pretrained or pretrained language models (PLMs). Non-pretrained models predominantly utilize deep neural networks (DNNs), such as long short-term memory (LSTM), graph convolutional networks

(GCN), and attention-based models for the purpose of stance classification. For example, Du et al. [6] used an attention model leveraging target information, while Du et al. [6] developed separate LSTMs to filter non-neutral text and classify attitudes. Sun et al. [7] proposed hierarchical attention to learn text representations via linguistic features, and Liang et al. [8] introduced a GCN approach to distinguish target-specific and invariant features. Furthermore, Devi and Kannimuthu [9] incorporated focal-loss and context-embedding-based data augmentation to handle the data imbalance. Inspired by promising PLM results, fine-tuning strategies have been developed to enhance the accuracy of stance detection [10]. These methods entail the adaptation of pretrained models, such as BERT [11] and RoBERTa [12], using datasets specific to stance detection, thereby tailoring the models to this particular task. In summary, these approaches predominantly conceptualize stance detection as a target-oriented, sentence-level text classification task. Nonetheless, the challenge of data sparsity, exacerbated by the informal and abbreviated nature of social media content, remains a significant obstacle to the efficacy of these methods. Recently, some research has addressed the issue of data sparsity by integrating external knowledge, thus enhancing both the performance and the interpretability of stance detection processes. For example, He et al. [13] augmented text classifiers by supplementing them with relevant Wikipedia documents about the target. Diaz et al. [14] constructed a stance tree using external knowledge extracted from a knowledge base and utilized it as evidence to enhance stance prediction and detection precision.

While these works demonstrated enhancements in performance and interpretability, the practical application of these methods encounters several challenges: (i) Deep neural networks (DNNs) are often perceived as “black box” mechanisms, due to their inability to furnish explicit rationales for their decision-making processes. As a result, DNNs may not be suitable for applications where interpretability is a crucial requirement. (ii) Existing methods in stance detection largely rely on extensive datasets that require manual annotation, a process that is time-consuming and labor-intensive. Although zero-shot learning settings have been introduced, they still necessitate significant data annotation within the source domain, complicating the direct application to unseen targets. (iii) The impracticality of deploying very large language models (VLLMs) with interpretative capabilities in stance detection arises from their substantial resource consumption and local deployment complexities, alongside potential data privacy concerns associated with techniques like chain-of-thought (CoT) processing, especially in areas like business decision-making and political analysis. The issue in question has been reported and has raised concerns. Consequently, it is imperative to propose a novel technique for transferring stance detection capabilities from VLLM to smaller, locally deployable models that can effectively address these concerns.

In response, this study aimed to develop a stance detection method that can simultaneously achieve interpretability, local deployment, and high accuracy with limited annotation. (1) To satisfy the interpretability requirement, we aimed to develop an understandable stance detection method that can generate the reasoning process of the stance predictor. (2) To meet the limited annotation requirement, we aimed to develop a method that can rely on only a small number of manual labels, while achieving comparable accuracy to state-of-the-art baselines. (3) To satisfy the local deployment requirement, our objective was to develop a method that can be trained on seen data and enable direct prediction on unseen data. In particular, the method should approximate the predictive performance of large-scale models.

To achieve this goal, in this paper, we proposed a distantly supervised explainable stance detection framework (DS-ESD). The DS-ESD model consists of three modules: an instruction-based chain-of-thought (CoT) method, generative network, and transformer-based stance predictor. The CoT method involves using manually designed prompt templates to extract the stance detection analysis process from VLLM in a CoT manner. This method was inspired by Wei et al. [15], who demonstrated the ability of large-scale models to comply with prompt instructions without requiring parameter training updates. Further-

more, a generative network is utilized to learn the mapping between input and inference process, with the expectation that it can generate the inference process independently of the VLLM during the prediction process. Finally, we constructed a stance classifier that takes as input the tweet and the generated inference process, and that is trained with VLLM-annotated stance labels, thus making it a form of distant supervision. Notably, for the stance classifier, we proposed a label rectification strategy to mitigate the impact of erroneous labels by controlling the probability distribution of the labels.

We summarize our contributions as follows:

- To the best of our knowledge, we present the first study on a distantly supervised stance detection framework, which also facilitates the generation of explanations for the stance analysis process. Our approach has advanced the field of stance detection.
- We propose a DS-ESD framework, which uses an instruct-based chain-of-thought approach to construct the supervised signal, upon which a generative model is subsequently built to generate explanations.
- We propose a novel label-rectification strategy for correcting label errors that arise from the distantly supervised approach.
- In order to evaluate the effectiveness of our proposed model, we conducted extensive experiments on three benchmark datasets. Our experimental results demonstrated that our model consistently outperformed the state-of-the-art methods in terms of predictive accuracy. Moreover, we conducted a manual evaluation of the generated explanations, which revealed that they were highly effective in providing clear and intuitive justifications for the model's predictions.

The remainder of this paper is structured as follows: Section 2 provides a comprehensive review and discussion of the related literature, including some traditional and recent methods of stance detection. Section 3 presents a detailed description of the proposed model. In Section 4, we describe the experimental setup, comprising the datasets employed for evaluation, the methods used for comparison, and report the quantitative evaluation results. Section 5 presents the conclusions and discusses future work.

2. Related Work

2.1. Stance Detection

The objective of stance detection is to ascertain and scrutinize the viewpoint expressed in a given text concerning a specific subject [16,17].

(1) Within the context of an in-target setting, existing methodologies can generally be classified into two types: non-pretrained and pretrained approaches. Non-pretrained methods frequently employ deep neural networks, such as attention (Att) and graph convolutional networks (GCN), for training stance classifiers. The Att techniques prioritize target-specific information as the attention query and implement an attention mechanism to deduce the stance polarity [6,7,18,19]. The GCN methods employ a graph convolutional network to delineate the interrelation between the target and the text, thus facilitating a nuanced analysis of their connection [20–22].

(2) In the realm of cross-target stance detection, researchers have introduced a variety of methodologies, which can be fundamentally segmented into two distinct classes. The initial class encompasses word-level transfer approaches, which exploit the commonality of words across different targets to mitigate knowledge disparities [23]. On the other hand, the second class addresses cross-target challenges through the adoption of concept-level knowledge transfer, wherein concepts shared between two targets are utilized to facilitate understanding and analysis [24–26].

(3) Zero-shot stance detection represents a particularly formidable challenge, necessitating a stance detection model to deduce the stance towards a target that has not been previously encountered. Addressing this complexity, Allaway and McKeown [27] constructed a comprehensive dataset for stance detection annotated by human experts and tailored for the zero-shot framework. Furthermore, Allaway et al. [28] applied adversarial learning techniques to derive target-invariant features and utilized a target-specific stance

detection dataset to facilitate zero-shot stance detection. Liu et al. [10] introduced a graph-based model that integrates intra- and extra-semantic information, as well as common sense knowledge, leveraging BERT to enhance the semantic insights garnered. In addition, Liang et al. [8] developed a sophisticated methodology for identifying target-specific or target-invariant characteristics, aiming to secure transferable features for stance detection.

2.2. Background Knowledge Enhanced Stance Detection Methods

The incorporation of background knowledge to enhance stance detection capabilities has garnered significant interest, representing a promising strategy to amplify its efficacy [29]. He et al. [13] introduced an approach that integrates target-related background knowledge, such as encyclopedic information from Wikipedia, and devised a fine-tuning methodology to augment the model’s learning proficiency. In a similar vein, Liu et al. [10] constructed a knowledge graph representing background knowledge and employed graph neural network techniques to develop an advanced stance prediction model. Moreover, Huang et al. [30] explored the utility of *#hashtag* background knowledge to refine content learning processes. Additionally, Luo et al. [31] merged sentiment knowledge into their framework to enhance the learning of attitudinal nuances.

2.3. Explainable Stance Detection

Traditional methods for explainable stance detection have focused on revealing areas that have a significant impact on the final prediction. For example, Draws et al. [32] introduced user search terms and built an interpretable stance detection model. Gómez-Suta et al. [33] proposed an approach for explaining stance labels by identifying the most relevant terms within topics derived from corresponding tweets. Huang et al. [30] further learned the topic words and integrated them into a prompt-based model to enhance the performance of stance detection.

These techniques identify information that significantly contributes to the final predictions, but the underlying reasoning process (i.e., the rationale for the prediction) remains obscured. Jayaram and Allaway [34] manually annotated the stance reasoning process and verified that it could effectively improve the performance of stance prediction. Inspired by this work, we present, for the first time, a study on automatically generating explanations, which advances the stance detection community.

3. Our Methodology

As illustrated in Figure 1, our method mainly consists of three modules: instruct-based CoT, generative model, and the transformer-based stance network.

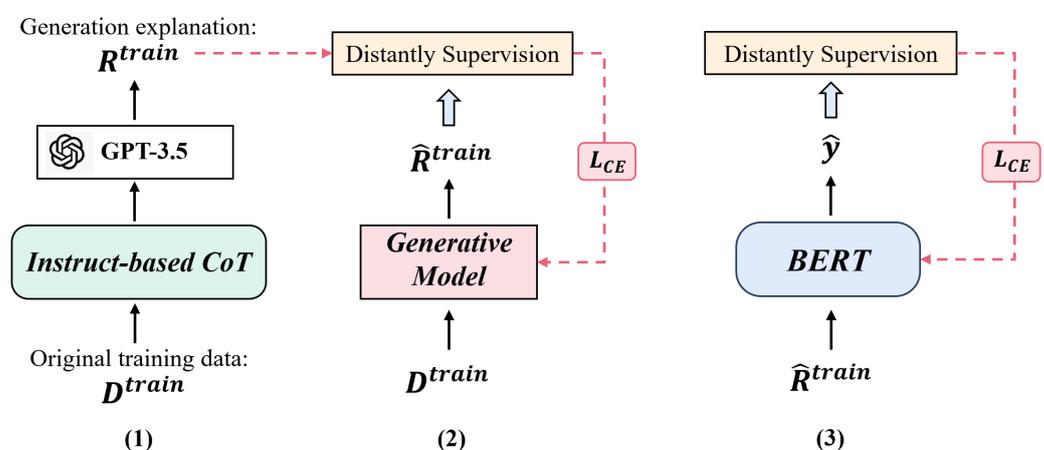


Figure 1. The overall structure of propose DS-ESD, including instruct-based CoT (1), generative model (2), and the transformer-based stance network (3).

3.1. Problem Formulation

We use $D^{train} = \{x_i, p_i\}$ to denote the collection of labeled data, where x and p denote the input text and the corresponding target, respectively. Each (x, p) pair in D^{train} is assigned a stance label y . Given an input sentence x^t and a corresponding target p^t as a test set (unseen target), this study aimed to predict the rationale of prediction r with a stance label \hat{y} for the input sentence x^t towards a given target p^t by using the proposed DS-ESD method.

3.2. Model Process

Our method is divided into two stages: training and prediction. During training, we incorporate the VLLM to aid in model training. In the prediction stage, we aim to achieve high accuracy using a generative model to produce explanations independently of large models.

For the training stage, given D^{train} , we first perform instruct-based CoT to collect retrieved explanations R^{train} . Then, we pack D^{train} and \hat{R}^{train} as the training sample for training the generation model. After training the generation model, we can feed the predicted \hat{R} , which is predicted by the generation model, into the transformer-based model to train the stance classification model.

During the inference process, we simply feed the test data x^t, p^t into the generative model to generate the corresponding inference process r^t . Subsequently, we feed both r^t and x^t into the stance classifier to automatically predict the stance.

3.3. Instruct-Based CoT

Traditional distantly supervised methods are mostly based on knowledge graphs to construct weak supervision signals. Due to the remarkable knowledge and understanding ability emerging from VLLM in recent years, this paper proposes a method based on CoT to construct weak supervision signals.

The CoT methodology has revealed the potential of VLLM multi-hop reasoning, wherein an VLLM is capable of impressive chain-style reasoning when given some input prompts or instructions. We devised a methodology that leverages a large model to extract pseudo-labels via instruction. Moreover, we aimed to utilize the analytical capabilities of the VLLM to extract the reasoning process of the model. To this end, we engineered a 1-shot instructional framework, as depicted in Figure 2. This framework is designed to facilitate the large-scale model in generating a logical sequence of the new samples along with their associated stance labels, provided that an instruction and a reference sample are supplied as inputs.

API call: *Your task is to add calls to a Question Answering API to a piece of text. The questions should help you get information required to complete the text. You can call the API by writing "[RULE: IF (A) then (B)]" where "A" is the reason why "B". Here are some examples of API calls:*

Input: what's the attitude of the sentence [input text x] to the target [given target o]?
Select an answer from (favor, against, none).

Output: IF (reason) then (attitude is [stance label]).

Input: . . .

Figure 2. 1-shot instruction.

Specifically, we send the API call (instruction) to the VLLM to acquire the explanation. Figure 3 provides a concrete example, where an instruction and a single sample are given, and the large model is fed with the input sentence "She's smarter and works harder". The large model then extracts the specific reasoning process, which is "IF (*Hillary is smarter and works harder*) THEN (the attitude towards Hillary is *favor*)".

Subsequently, we leverage the extracted inference process, "Hillary is smarter and works harder" as the reasoning process, which is denoted as r , and "favor" in the expla-

nation as the pseudo-labels. Finally, the instruct-based CoT approach can be employed to obtain the inference process \hat{R}^{train} and the pseudo-labels \tilde{Y} of the training data D^{train} .



User

Your task is to add calls to a Question Answering API to a piece of text. The questions should help you get information required to complete the text. You can call the API by writing "[RULE: IF (A) then (B)]" where "A" is the reason why "B".

Here are some examples of API calls:

Input: what's the attitude of the sentence 'So when all you brave patriots stop Hillary? LOL. #losers '. to the target 'Hillary Clinton'? select an answer from (favor, against, none).

Output: RULE: IF the target: Hillary Clinton (needs to be stopped means disparage) then (the attitude is against).

Input: She's smarter and works harder



GPT

Output: RULE: IF the sentence (suggests that Hillary is smarter and works harder) Then (the attitude towards Hillary is in favor)

Figure 3. Example of instruct-based CoT.

3.4. Fine-Tuning the Generative Model with a New Dataset

Although it is possible to directly extract r from VLLM, in applications such as business decision-making, we prefer to use smaller models that can be deployed locally, to prevent data leakage when predicting stances. Therefore, we designed a generative model to learn the mapping between x and r . During practical deployment, the generation model can directly generate r for unseen targets without the need for VLLM involvement.

The generator component can be modeled using any encoder–decoder architecture. In this study, we adopted BART-large, a pretrained sequence-to-sequence transformer comprising 400 million parameters, as our choice. We concatenate the target p with the input x . This context matrix is then given as memory to the decoder. We trained the model to minimize the negative log-likelihood of the target utterance r .

In summary, our proposed model employs a pretrained BART architecture to extract pertinent knowledge of the given input and generate an explanation. When provided with a training corpus consisting of input/output pairs, denoted as (x_j, p_i) and r_j , respectively, for the purpose of fine-tuning, we employ stochastic gradient descent with the Adam optimizer to minimize the negative marginal log-likelihood of each target $P_j \log_p(r_j | x_j, p_i)$. This approach allows us to effectively optimize the network's parameters and improve its ability to predict the correct outputs for a given set of inputs and parameters.

3.5. Text Representation Method

After obtaining the explanations of the reasoning process r , we proposed the transformer-based network as the stance predictor. The input of the transformer is the reasoning process r and text x , and the output is the stance label.

Formally, we combine r and tweet x as the input. Given the hidden states of the representation, which correspond to the BERT model's output, as H . The hidden states are subsequently fed into the multi-head self-attention mechanism (MHSA) to compute the output of the transformer layer, expressed as:

$$Q = HW_q, K = HW_k, V = HW_v, \quad (1)$$

$$\hat{A} = \frac{QK^T}{\sqrt{d_K}}, \quad (2)$$

$$Attn(H) = \tilde{A}VW_v \quad (3)$$

where the matrices Q , K , and V represent the query, key, and value, respectively, as per the standard MHSA mechanism. Next, we implement a conventional residual structure that fuses the higher-level representation H and the current $Attn(H)$ and applies layer normalization (LN) to normalize the resultant output.

$$Attn(H') = LN(Attn(H) + H) \quad (4)$$

The transformer block's ultimate output is obtained by passing $Attn(H')$ through a feed-forward layer based on attention mechanisms.

$$\alpha_t = softmax(Attn(H')_t) \quad (5)$$

Subsequently, the attentive sentence representation e is learned by aggregating the embeddings of $Attn(H')$ using the attention vector α :

$$e = \sum_{t=1}^n \alpha_t Attn(H)_t \quad (6)$$

3.6. Label Rectification Strategy

In order to prevent the model from becoming overconfident and assigning excessively high probabilities to a single-label class, we leverage a label smoothing strategy, which entails assigning a fixed small probability to alternative classes. However, in our specific scenario, the pseudo-labels themselves are not reliable and cannot be used directly. To solve the problem, we introduced a novel label rectification strategy that can dynamically adjust noisy labels. Essentially, we modify the distribution of the original labels to steer them in the correct direction in the presence of potential errors, thus improving the overall accuracy of the model predictions.

More specifically, in the rectification module, a linear transformation is applied to the representation of the sentence e by the transformer layer, which results in a distributional representation \tilde{e} that is unique to the rectifier (\tilde{e} has the same dimensions as e). Subsequently, the rectifier takes \tilde{e} as input and outputs a rectification matrix M , as follows:

$$M_{i,j} = W_{i,j}^T \tilde{e} + b, \quad 1 \leq i, j \leq K \quad (7)$$

Here, K denotes the total number of stance labels, and $W_{i,j}$ possesses the same dimensions as \tilde{e} . The rectified label distribution is subsequently computed as:

$$q = softmax(T \times l) \quad (8)$$

where $\text{softmax}(q_i) = e^{q_i} / \sum_j e^{q_j}$ denotes the normalized function. $M_{i,j}$ denotes the extent to which the i -th class is misclassified as the j -th class.

Let M_i denote the i -th row of M . Assuming that the label corresponding to noise is k , such that $l_k = 1$ and $l_j = 0$ for $j \neq k$, we obtain $q_i = M_{ik}$, which is equivalent to $q = T_k$ (where T_k denotes the k -th column of matrix T). As such, M_{ik} quantifies the extent to which the true label is i but the labeled noise is k . Through this approach, matrix M enables modification of the original label distribution l to a new distribution q .

3.7. Adaptive Training Mechanism

As illustrated in Figure 4, our approach consists of two distinct losses. Notably, training the rectification module ($loss_2$) without any direct human guidance or involvement poses a significant challenge. To overcome this challenge, we introduced a novel adaptive mechanism based on curriculum learning. Curriculum learning simulates the human learning process by starting with simpler tasks and gradually increasing the difficulty level. In our approach, we first concentrate on the prediction module ($loss_1$) to minimize the discrepancy between the predicted distribution and the ground truth distribution. We then gradually increase the level of complexity to enable the model to learn to cope with the noise present in the data. In the second step, we balance the two losses and obtain the final loss function $loss = \alpha \times loss_1 + (1 - \alpha) \times loss_2$, where α is the balancing coefficient. We dynamically compute the coefficient using the available information. Specifically, we use p_k as the coefficient, where k denotes the annotated label. We validate this strategy in two scenarios: (1) A value of p_k approaching 1 signifies a high level of confidence that k is the appropriate label, resulting in reduced emphasis on the rectification module. Given the initially small value of $loss_1$, the overall loss remains relatively low, and the $(1 - \alpha)$ term restricts the magnitude of the second component. (2) Conversely, a value of p_k approaching 0 implies that the annotated label may be incorrect, necessitating greater attention towards the rectification module to rectify it.

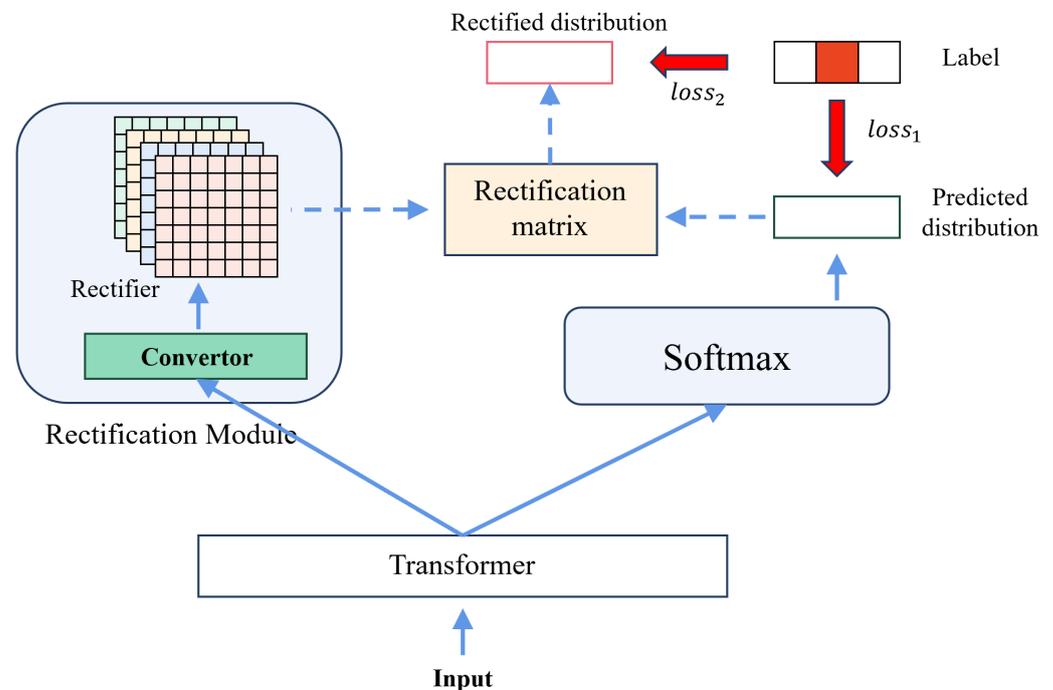


Figure 4. Framework of label rectification strategy.

In fact, q is utilized to explore the entire space and identify the true distribution of the labels. The search process commences with the labeled distribution l and adopts $(1 - p_k) \times loss_2$ as the loss function. This approach allows q to assimilate information from

the labeled distribution l , predicted distribution p , and specific context x . The matrix T is initially initialized as the identity matrix, signifying that noise is not considered.

Finally, the first loss function can be effectively implemented through the utilization of the standard cross-entropy method:

$$loss_1 = -\log p(y_i | (x, p)_i^j; \theta) \quad (9)$$

Every ground-truth label, y_i , pertaining to the i -th individual sample, is represented in one-hot format. To optimize the network, the standard gradient descent algorithm is employed. The second loss is utilized to automatically adjust the distribution of incorrect categories. Finally, we combine this as follows:

$$loss_2 = \text{sqr}t\left(\sum_{k=1}^K (p(k | (x, p)_i^j; \theta) - q(k | (x, p)_i^j; \theta))^2\right) \quad (10)$$

$$J(\theta) = p(y_i | (x, p)_i^j; \theta) \times loss_1 + (1 - p(y_i | (x, p)_i^j; \theta)) \times loss_2$$

4. Experiments

4.1. Experimental Data

This paper presents experimental results on robust benchmark datasets, including SemEval-2016 Task 6 (SEM16) [35], COVID-19 [36], and VAST [27]. Their statistics are shown in Tables 1 and 2.

Table 1. Statistics of SemEval16 and COVID-19 datasets.

Dataset	Target	Favor	Against	Neutral
SEM16	H	163	565	256
	F	268	511	170
	L	167	544	222
COVID-19	Fauci	492	610	762
	Home	615	250	325
	Mask	190	400	782
	School	693	668	346

Table 2. Statistics of VAST dataset.

	Train	Valid	Test
Examples	13,477	2062	3006
Unique Comments	1845	682	786
Zero-shot Topics	4003	383	600

- SEM16. The SEM16 dataset contains 4870 tweets, each targeting various subjects and annotated with one of three stance labels: “favor”, “against”, or “neutral”. Following the framework suggested by [24], four specific targets—Donald Trump (D), Hillary Clinton (H), Legalization of Abortion (L), and Feminist Movement (F)—were selected for the analysis of stance detection efficacy in our research. For the cross-target configuration [8,24,25], we formulated four distinct cross-target stance detection tasks ($D \rightarrow H$, $H \rightarrow D$, $F \rightarrow L$, $L \rightarrow F$), indicating the source target on the left and the destination target on the right of the arrow.
- COVID-19. The COVID-19 Stance dataset comprises 6133 tweets that pertain to users’ stance towards four targets related to COVID-19 health mandates. The tweets were manually annotated for stance based on three categories: in-favor, against, and neither.
- VAST. Introduced by Allaway and Mckeown [27], the VAST dataset incorporates a wide array of targets across different sectors, including politics, education, and public health, and features three stance labels: “pro”, “neutral”, and “con”. It consists of

4003 samples for training, with the development and test sets containing 383 and 600 samples, respectively. In alignment with Liang et al. [8], our model's performance is assessed on topics in a zero-shot learning context.

4.2. Compared Baseline Methods

To evaluate the performance of our proposed model, a comprehensive analysis and comparison with existing baseline models was conducted. These baseline models are delineated as follows:

Statistics-based methods:

- **BiLSTM** [23] employs a bidirectional long short-term memory (LSTM) network to independently encode the text and its associated target.
- **BiCond** [23] utilizes a bidirectional LSTM to encode both the text and the target concurrently.
- **CrossNet** [37] extends BiCond by incorporating a self-attention layer to identify salient words within the text.
- **MemNet** [38] introduces a multi-hop attention mechanism within a memory network to effectively encode textual data.
- **AoA** [39] deploys two LSTM networks to separately model the target and context, integrating an interactive attention mechanism to examine their interaction.
- **ASGCN** [40] employs a dependency tree and graph convolutional networks (GCN) to derive compact and expressive textual representations.
- **TAN** [6] integrates target-specific attention with a long-short term memory network for stance detection.
- **TPDG** [41] introduces a convolutional graph model adaptable to the target, enhancing stance detection accuracy through the utilization of shared features from similar targets.
- **AT-JSS-Lex** [42] presents a target-adaptive graph convolutional network for stance detection, focusing on the extraction of shared latent features from similar targets.
- **TOAD** [28] employs adversarial learning to achieve generalization across different topics.
- **GCAE** introduces a gated convolutional network based on a CNN framework, which integrates target-specific information and employs a gating mechanism to exclude irrelevant information.

Fine-tuning based methods:

- **BERT** [11] employs a pretrained BERT model for stance detection, adapting the input format to "[CLS] + text + [SEP] + target + [SEP]" to facilitate the model training and fine-tuning processes.
- **BERT-NS** [43] represents a semi-supervised approach that applies self-training and knowledge distillation to improve the efficacy of a teacher model through the use of unlabeled data.
- **BERT-DAN** [44] is designed to explicitly capture both subjective and objective elements within tweets and allows the use of labeled data from related tasks to inform the training of a model for the target task.
- **PT-HCL** [8] introduces an innovative method for cross-target and zero-shot stance detection employing contrastive learning. This model uses a BERT-based framework to create a unified representation space for various targets.

Prompt-tuning based methods:

- **MPT** [45] provides a prompt-tuning based method for stance detection, which employs a verbalizer defined by human experts.
- **KPT** [46] incorporates external lexical resources to define the verbalizer component within the prompt engineering framework.
- **PIN-POM** [47] puts forth a soft prompt approach tailored for short text categorization, an adaptation readily amenable to stance detection tasks.
- **TAPD** [48] uses a prompt setting method for position detection, using PLM to learn effective representations for stance detection tasks.

Knowledge-enhanced methods:

- **SEKT** [25] provides a graph convolutional network enhanced with semantic knowledge to detect attitudes.
- **WS-BERT-Dual** [28] introduces target-related wiki knowledge to enhance stance detection ability.
- **TarBK** [29] incorporates the targeted background knowledge for stance detection.

Variants of DS-ESD:

- **S-ESD** refers to supervised learning using backpropagation with a small set of labeled training samples. In contrast, DS-ESD does not require manual annotation and is applicable to more open real-world scenarios, while S-ESD is suited for current in-target and cross-target task settings.

4.3. Implementation Details

In the experimental configuration, we opted for the BART-large architecture with 400 million parameters for the generator component. Subsequently, for the stance classification model, we elected to utilize pretrained language models based on the BERT-base architecture with 340 parameters. To train the model, we utilized the Adam optimizer with a mini-batch size of 32 and a learning rate of 0.0002. The hardware environment for these experiments was provisioned with an A100 40G GPU. To further improve on the current state of the art, we comprehensively describe the templates used to fine-tune pretrained language models throughout this paper.

As per the recommendations of previous works [8,25], we employed the micro-average F1 score as our primary evaluation metric. Our first step in this process involved calculating the F1 scores for the categories “favor” and “against”:

$$F1_{favor} = \frac{2P_{favor}R_{favor}}{P_{favor} + R_{favor}} \quad (11)$$

$$F1_{against} = \frac{2P_{against}R_{against}}{P_{against} + R_{against}}$$

The F1-score could be computed based on precision and recall.

$$F1 = \frac{F1_{favor} + F1_{against}}{2} \quad (12)$$

4.4. Overall Performance

4.4.1. In-Target Setup

Tables 3 and 4 present the results of in-target stance detection using diverse robust benchmarks. Our DS-ESD model consistently outperformed most of the baseline models across all datasets, thereby validating the efficacy of our proposed approach for stance detection. Moreover, the significance tests conducted on DS-ESD relative to the top-performing competitor demonstrated that DS-ESD yielded a statistically significant enhancement in terms of most evaluation metrics, with p -value < 0.05 (indicated as [†]). Specifically, compared with the static-based model (GCAE) that performed poorly, our DS-ESD improved on it by 18.8% on average for the COVID-19 dataset. Compared to KPT and MT, the best competitors of the BERT-based model, our DS-ESD improved by 1.9% and 1.04% on average over SEM16 and COVID-19, respectively.

Table 3. Performance comparison of in-target setups for SEM16. The best scores are in bold.

	Methods	F	L	H
Sta.	BiLSTM †	51.6	59.1	55.8
	BiCond †	52.9	61.2	56.1
	TAN †	55.8	63.7	65.4
	AT-JSS-Lex †	61.5	68.4	68.3
	MemNet †	51.1	58.9	52.3
	AoA †	55.4	58.3	51.6
	ASGCN †	56.2	59.5	62.2
	TPDG	67.3	74.7	73.4
BERT	FT	62.3	62.4	67.0
	S-MDMT	63.8	67.2	67.2
	STANCY	61.7	63.4	64.7
	TAPD	63.9	63.9	70.1
	MPT	63.1	62.9	70.4
	PIN-POM	62.1	62.9	69.2
	KPT	63.3	63.5	71.3
	DS-ESD	72.2 †	65.6	77.5 †
	S-ESD	71.7 †	68.0	78.5 †

Table 4. Performance comparison of in-target setups for COVID-19. The best scores are in bold.

	Methods	Fauci	School	Home	Mask
Sta.	BiLSTM †	63.0	54.8	64.5	56.7
	TAN †	54.7	53.4	53.6	54.6
	ATGRU	61.2	52.7	52.1	59.9
	GCAE †	64.0	49.0	64.5	63.3
BERT	BERT	81.8	75.5	80.0	80.3
	BERT-NS	82.1	75.3	78.4	83.3
	BERT-DAN	83.2	71.7	78.7	82.5
	MT-LRM-BERT	83.7	79.3	82.7	84.7
	DS-ESD	78.9	76.5	78.9	81.5
	S-ESD	83.8	79.3	85.1 †	86.5 †

It is noteworthy that utilizing the method of distant supervision, which obviates the need for manual data annotation, achieved significant improvements in effectiveness across multiple settings compared to strong baselines. This finding indicates the effectiveness of our proposed approach, which leverages VLLM to annotate labels and conducts distant supervision.

4.4.2. Cross-Target Setup

The procurement of a comprehensively annotated large dataset necessitates substantial time and resources. Consequently, we proposed to evaluate the efficacy of our method within a cross-target framework. The objective of this framework was to predict the stance towards the target destination using labeled data from the source target. The F1 scores are detailed in Table 5. According to these findings, our proposed methodologies (DS-ESD and S-ESD) surpassed the competing baselines by a notable margin. Specifically, DS-ESD exhibited an average enhancement of 12.85% in F1 score over the top-performing statistical method (TPDG), affirming the efficiency of employing a distantly supervised approach in a cross-target context. Furthermore, when compared to the leading fine-tuning-based method (PT-HCL), DS-ESD registered an average improvement of 12.05% in F1 score.

Table 5. Performance comparison of cross-target stance detection. The best scores are in bold.

	Methods	F→L	L→F	H→D	D→H
Sta.	BiLSTM †	44.8	41.2	29.8	35.8
	BiCond	45.0	41.6	29.7	35.8
	CrossNet	45.4	43.3	43.1	36.2
	VTN	47.3	47.8	47.9	36.4
	SEKT	53.6	51.3	47.7	42.0
	TPDG	58.3	54.1	50.4	52.9
BERT	BERT-FT	47.9	33.9	43.6	36.5
	MPT	42.1	47.6	47.1	58.7
	KPT	49.1	54.2	54.6	60.9
	JointCL	58.8	54.5	52.8	54.3
	PT-HCL	59.3	54.6	53.7	55.3
	TarBK	59.1	54.6	53.1	54.2
	DS-ESD	66.0 †	66.3 †	69.3 †	69.5 †
	S-ESD	66.4 †	69.3 †	70.8 †	71.7 †

4.4.3. Zero-Shot Stance Detection

In instances where the text’s target was not present in the training dataset, we undertook a comparative analysis against the foremost competitors in the domain. The outcomes of these analyses are documented in Table 6. It is crucial to note that, given the intrinsic challenges and constraints associated with zero-shot stance detection, all techniques manifested a diminished performance relative to the in-target configuration. In particular, methods predicated exclusively on statistical analysis exhibited inferior results. In contrast, fine-tuning-based strategies, such as PT-HCL, TarBK, and TTS, consistently surpassed those reliant on statistical analyses. This phenomenon underscored the substantial advantages of harnessing knowledge derived from extensive corpora. Despite the inherent difficulties of zero-shot stance detection, our DS-ESD model showcased notable efficacy, rivaling the performance of the leading benchmark methods. When equipped with labeled data from the source domain, our approach (S-ESD) recorded the highest F1 score. Consequently, our findings suggest that DS-ESD constitutes an effective approach for navigating the complex task of zero-shot stance detection through distantly supervised methods.

Table 6. Performance comparison of zero-shot stance detection. The best scores are in bold.

	Methods	All	Pro	Con	FLD→H	LDH→F	FDH→L	FLH→D
Sta.	BiCond	42.8	44.6	47.4	32.7	40.6	34.4	30.5
	CrossNet	43.4	46.2	43.4	38.3	41.7	38.5	35.6
	SEKT	41.8	50.4	44.2	50.1	44.2	44.6	46.8
	TPDG	51.9	53.7	49.6	50.9	53.6	46.5	47.3
BERT	BERT-FT	66.1	54.6	58.4	49.6	41.9	44.8	40.1
	CKE-Net	70.2	61.2	61.2	-	-	-	-
	MPT	66.6	54.9	62.0	52.0	47.6	50.2	48.7
	KPT	70.2	61.1	61.4	47.3	50.9	50.6	52.2
	JointCL	72.3	64.9	63.2	54.8	53.8	49.5	50.5
	PT-HCL	-	-	-	54.5	54.6	50.9	50.1
	TarBK	-	-	-	55.1	53.8	48.7	-
	TTS	-	-	-	71.6	64.4	62.6	65.8
	DS-ESD	71.2	66.4 †	63.8 †	72.2 †	68.4 †	65.1 †	69.5 †
	S-ESD	72.9 †	62.0	68.1	73.8 †	69.5 †	68.6 †	70.2 †

4.5. Ablation Study

To investigate the impact of each part on our model, we performed an ablation test by discarding the label rectification strategy (denoted as w/o LRS) and the adaptive training

mechanism (denoted as w/o ATM), respectively. Specifically, for the w/o LRS, the model was trained using standard cross-entropy. Additionally, following [49], we constructed a method that relied solely on ChatGPT to verify performance without LRS, denoted as “ChatGPT”.

The ablation study results are summarized in Table 7. The findings reveal that both the LRS and the ATM contributed significantly to enhancing the performance of the proposed approach. In particular, the performance significantly dropped when LRS was removed. This is because using ChatGPT’s results as pseudo-labels directly introduced a considerable amount of noise, thereby adversely affecting performance. The empirical outcomes additionally corroborated the efficacy of the proposed LRS. Not unexpectedly, integrating all components yielded the optimal outcomes across all experimental setups.

Table 7. Experimental results of ablation study.

Methods	F	L	H
DS-ESD	72.2	65.6	77.5
-w/o LRS	71.9	64.0	76.7
-w/o ATM	72.0	65.1	77.1
ChatGPT	68.4	58.2	79.5

Experimental Results on Varying Amounts of Labeled Data. The size of the labeled samples is crucial for the proposed method, as it significantly impacts both the model’s performance and running time. In this experiment, we conducted tests on the F and L task in the SEM16 dataset, with varying amounts of labeled data, ranging from 0% to 100%. Figure 5 shows the results. Notably, when using 0% labeled data, we refer to our method as DS-ESD, while using 100% labeled samples represents S-ESD. The results are shown in Figure 5. The empirical findings illustrate that the performance, as expected, progressively improved with the volume of annotated data, an outcome that aligned with our expectations. Specifically, the results when using 0% labeled data exceeded the baseline method in most settings, indicating the effectiveness of our distant supervision method.

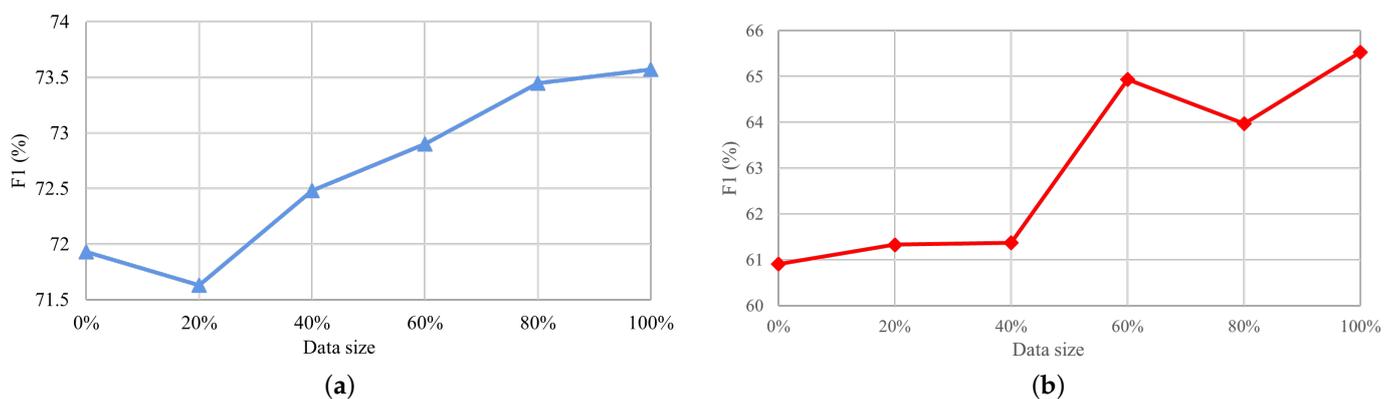


Figure 5. Experimental results on varying amounts of labeled data; (a) F1 score on “L”; (b) F1 score on “F”.

4.6. Case Study

Table 8 presents three explanation examples generated by DS-ESD. The selected samples were accurately predicted by DS-ESD, whereas the baseline failed to predict the correct category.

Table 8. Examples of the explanation generated by DS-ESD.

Input	Target	Generated Explanation
@HillaryClinton now can add #ropegate to her long list of “accomplishments”.	Hillary Clinton	the sentence suggests that ropegate has a negative impact on Hillary
She’s smarter and works harder. Empty accusations, inferences, gossip, and rumors are rubbish.	Hillary Clinton	the sentence suggests that Hillary is smarter and works harder
Based on the long lines, I thought it was free burrito day at Pancheros but it was actually Hillary!	Hillary Clinton	the sentence implicitly expresses support for Hillary
A fictional character been raped and abused oppress me. Censor pls!! #thisoppresseswomen	Feminist Movement	the sentence includes #thisoppresseswomen means negative for women’s rights

In the first and fourth examples, it is evident that the DS-ESD model possessed a profound understanding of stance-aware symbols. For instance, the hashtag “#ropegate” was associated with a negative news event during Hillary Clinton’s presidential campaign, and “#thisoppresseswomen” signified opposition to women’s rights. The explanations generated by the DS-ESD model illustrate its capacity to grasp the contextual significance of these tags. In the second example, when conflicting stance-bearing words appeared in the text, DS-ESD could effectively identify the correct words that describe the target. The third example shows that DS-ESD could effectively understand semantic content that requires a deep understanding. For example, the comparison between the long queues to vote for Hillary and the Free Burrito Day queue was accurately categorized by DS-ESD as “implicitly expresses support for Hillary”.

At present, due to constraints on model parameters, the generated explanations are relatively concise and lack detailed elaboration. Future research endeavors may consider expanding the model’s parameter space to generate more comprehensive explanations.

4.7. Manual Evaluation

As the evaluation metrics for the generative model score (e.g., BLEU score) only considered the word-level similarity between ground truth and predicted outputs, we conducted a manual evaluation of the quality of the explanations produced.

Specifically, we asked three annotators to rank these explanations using three different criteria: (1) The generated explanation contains important, salient information and does not omit any essential points that contributed to the stance prediction. (2) The generated explanation does not contain any redundant, repeated, or irrelevant information to the input and the stance detection. (3) The generated explanation does not contain any contradictory pieces of information to the input and the stance detection.

We randomly selected a small set of 100 instances from the test set, and the evaluators scored them according to the above evaluation criteria with a range of ± 1 . The average scores of the three evaluators all exceeded 80, demonstrating that the generated explanations could effectively explain the rationale of the prediction.

5. Conclusions

We proposed a distantly supervised explainable stance detection framework (DS-ESD) comprising three modules: an instruction-based chain-of-thought (CoT) method, a generative network, and a transformer-based stance predictor. The CoT method leverages prompt templates to extract stance explanations from a very large language model (VLLM) like GPT-3.5. A generative network then learns the mapping between input and explanation.

The transformer-based classifier takes the tweet and generates an explanation as input, trained with VLLM-labeled stances as distant supervision. Comprehensive experiments on three benchmarks demonstrated consistently better performance over the comparisons. Furthermore, future research endeavors may explore expanding the parameter space of the model to enhance the generation of more comprehensive explanations.

Author Contributions: Conceptualization, C.P. and B.Z.; Methodology, D.D., X.P. and H.H.; Investigation, B.Z.; Writing—original draft, G.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by National Nature science Foundation of china (No. 62306184), Natural Science Foundation of Top Talent of SZTU (grant no. GDRC202320) and the Research Promotion Project of Key Construction Discipline in Guangdong Province (2022ZDJS112).

Data Availability Statement: No new data were created during this study. The study brought together existing data obtained upon request and subject to licence restrictions from a number of different sources.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Küçük, D.; Can, F. Stance detection: A survey. *ACM Comput. Surv.* **2020**, *53*, 1–37. [[CrossRef](#)]
- Walker, M.A.; Anand, P.; Abbott, R.; Grant, R. Stance classification using dialogic properties of persuasion. In Proceedings of the 2012 the North American Chapter of the Association for Computational Linguistics, Montréal, QC, Canada, 3–8 June 2012; pp. 592–596.
- Somasundaran, S.; Wiebe, J. Recognizing stances in online debates. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, 2–7 August 2009; pp. 226–234.
- Yang, M.; Zhao, W.; Chen, L.; Qu, Q.; Zhao, Z.; Shen, Y. Investigating the transferring capability of capsule networks for text classification. *Neural Netw.* **2019**, *118*, 247–261. [[CrossRef](#)] [[PubMed](#)]
- Zhang, Y.; Tiwari, P.; Song, D.; Mao, X.; Wang, P.; Li, X.; Pandey, H.M. Learning interaction dynamics with an interactive LSTM for conversational sentiment analysis. *Neural Netw.* **2021**, *133*, 40–56. [[PubMed](#)]
- Du, J.; Xu, R.; He, Y.; Gui, L. Stance classification with target-specific neural attention networks. In Proceedings of the 26th International Joint Conference on Artificial Intelligence, Melbourne, VIC, Australia, 19–25 August 2017.
- Sun, Q.; Wang, Z.; Zhu, Q.; Zhou, G. Stance detection with hierarchical attention network. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NW, USA, 20–26 August 2018; pp. 2399–2409.
- Liang, B.; Chen, Z.; Gui, L.; He, Y.; Yang, M.; Xu, R. Zero-Shot Stance Detection via Contrastive Learning. In Proceedings of the ACM Web Conference, Lyon, France, 25–29 April 2022; pp. 2738–2747.
- Devi, V.S.; Kannimuthu, S. Author profiling in code-mixed WhatsApp messages using stacked convolution networks and contextualized embedding based text augmentation. *Neural Process. Lett.* **2023**, *55*, 589–614. [[CrossRef](#)]
- Liu, R.; Lin, Z.; Tan, Y.; Wang, W. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Online, 1–6 August 2021; pp. 3152–3157.
- Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. 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, NAACL-HLT, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A Robustly Optimized Bert Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
- He, Z.; Makhberian, N.; Lerman, K. Infusing Wikipedia Knowledge to Enhance Stance Detection. *arXiv* **2022**, arXiv:2204.03839.
- Diaz, G.A.; Chesñevar, C.I.; Estevez, E.; Maguitman, A. Stance Trees: A Novel Approach for Assessing Politically Polarized Issues in Twitter. In Proceedings of the 15th International Conference on Theory and Practice of Electronic Governance, Guimarães, Portugal, 4–7 October 2022; pp. 19–24.
- Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D. Chain-of-thought prompting elicits reasoning in large language models. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 24824–24837.
- Jain, R.; Jain, D.K.; Dharana; Sharma, N. Fake News Classification: A Quantitative Research Description. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* **2022**, *21*, 1–17. [[CrossRef](#)]
- Rani, S.; Kumar, P. Aspect-based Sentiment Analysis using Dependency Parsing. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* **2022**, *21*, 1–19. [[CrossRef](#)]
- Dey, K.; Shrivastava, R.; Kaushik, S. Topical Stance Detection for Twitter: A Two-Phase LSTM Model Using Attention. In Proceedings of the European Conference on Information Retrieval, Grenoble, France, 26–29 March 2018; pp. 529–536.

19. Wei, P.; Lin, J.; Mao, W. Multi-target stance detection via a dynamic memory-augmented network. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 1229–1232.
20. Li, C.; Peng, H.; Li, J.; Sun, L.; Lyu, L.; Wang, L.; Yu, P.S.; He, L. Joint Stance and Rumor Detection in Hierarchical Heterogeneous Graph. *IEEE Trans. Neural Networks Learn. Syst.* **2022**, *33*, 2530–2542. [[CrossRef](#)] [[PubMed](#)]
21. Cignarella, A.T.; Bosco, C.; Rosso, P. Do Dependency Relations Help in the Task of Stance Detection? In Proceedings of the Third Workshop on Insights from Negative Results in NLP, Dublin, Ireland, 26 May 2022; pp. 10–17.
22. Conforti, C.; Berndt, J.; Pilehvar, M.T.; Giannitsarou, C.; Toxvaerd, F.; Collier, N. Synthetic Examples Improve Cross-Target Generalization: A Study on Stance Detection on a Twitter corpus. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Online, 19 April 2021; pp. 181–187.
23. Augenstein, I.; Rocktaeschel, T.; Vlachos, A.; Bontcheva, K. Stance Detection with Bidirectional Conditional Encoding. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016.
24. Wei, P.; Mao, W. Modeling Transferable Topics for Cross-Target Stance Detection. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 1173–1176.
25. Zhang, B.; Yang, M.; Li, X.; Ye, Y.; Xu, X.; Dai, K. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 3188–3197.
26. Cambria, E.; Poria, S.; Hazarika, D.; Kwok, K. SenticNet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.
27. Allaway, E.; McKeown, K.R. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, 16–20 November 2020; pp. 8913–8931.
28. Allaway, E.; Srikanth, M.; McKeown, K.R. Adversarial Learning for Zero-Shot Stance Detection on Social Media. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, Online, 6–11 June 2021; pp. 4756–4767.
29. Zhu, Q.; Liang, B.; Sun, J.; Du, J.; Zhou, L.; Xu, R. 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, Madrid, Spain, 11–15 July 2022; pp. 2070–2075.
30. Huang, H.; Zhang, B.; Li, Y.; Zhang, B.; Sun, Y.; Luo, C.; Peng, C. Knowledge-enhanced Prompt-tuning for Stance Detection. *ACM Trans. Asian-Low Lang. Inf. Process.* **2023**, *22*, 1–20. [[CrossRef](#)]
31. Luo, Y.; Liu, Z.; Shi, Y.; Li, S.Z.; Zhang, Y. Exploiting Sentiment and Common Sense for Zero-shot Stance Detection. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022; pp. 7112–7123.
32. Draws, T.; Natesan Ramamurthy, K.; Baldini, I.; Dhurandhar, A.; Padhi, I.; Timmermans, B.; Tintarev, N. Explainable cross-topic stance detection for search results. In Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, Austin, TX, USA, 19–23 March 2023; pp. 221–235.
33. Gómez-Suta, M.; Echeverry-Correa, J.; Soto-Mejía, J.A. Stance detection in tweets: A topic modeling approach supporting explainability. *Expert Syst. Appl.* **2023**, *214*, 119046. [[CrossRef](#)]
34. Jayaram, S.; Allaway, E. Human Rationales as Attribution Priors for Explainable Stance Detection. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 5540–5554.
35. Mohammad, S.; Kiritchenko, S.; Sobhani, P.; Zhu, X.; Cherry, C. Semeval-2016 task 6: Detecting stance in tweets. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), San Diego, CA, USA, 16–17 June 2016; pp. 31–41.
36. Glandt, K.; Khanal, S.; Li, Y.; Caragea, D.; Caragea, C. 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, Virtual, 1–6 August 2021; pp. 1596–1611.
37. Xu, C.; Paris, C.; Nepal, S.; Sparks, R. Cross-Target Stance Classification with Self-Attention Networks. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, VIC, Australia, 15–20 July 2018; pp. 778–783.
38. Tang, D.; Qin, B.; Liu, T. Aspect Level Sentiment Classification with Deep Memory Network. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016.
39. Huang, B.; Ou, Y.; Carley, K.M. Aspect level sentiment classification with attention-over-attention neural networks. In Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation, Washington, DC, USA, 10–13 July 2018; pp. 197–206.
40. Zhang, C.; Li, Q.; Song, D. Aspect-based Sentiment Classification with Aspect-specific Graph Convolutional Networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 4560–4570.
41. Liang, B.; Fu, Y.; Gui, L.; Yang, M.; Du, J.; He, Y.; Xu, R. Target-adaptive Graph for Cross-target Stance Detection. In Proceedings of the WWW '21: The Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 3453–3464.

42. Li, Y.; Caragea, C. Multi-task stance detection with sentiment and stance lexicons. 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), Hong Kong, China, 3–7 November 2019; pp. 6299–6305.
43. Xie, Q.; Luong, M.T.; Hovy, E.; Le, Q.V. Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 10687–10698.
44. Xu, C.; Paris, C.; Nepal, S.; Sparks, R.; Long, C.; Wang, Y. DAN: Dual-View Representation Learning for Adapting Stance Classifiers to New Domains. In *ECAI 2020*; IOS Press: Amsterdam, The Netherlands, 2020; pp. 2260–2267.
45. Hu, S.; Ding, N.; Wang, H.; Liu, Z.; Li, J.; Sun, M. Knowledgeable Prompt-Tuning: Incorporating Knowledge into Prompt Verbalizer for Text Classification. *arXiv* **2021**, arXiv:2108.02035.
46. Shin, T.; Razeghi, Y.; IV, R.L.L.; Wallace, E.; Singh, S. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Online, 16–20 November 2020; pp. 4222–4235.
47. Dan, Y.; Zhou, J.; Chen, Q.; Bai, Q.; He, L. Enhancing Class Understanding Via Prompt-Tuning For Zero-Shot Text Classification. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 23–27 May 2022; pp. 4303–4307.
48. Jiang, Y.; Gao, J.; Shen, H.; Cheng, X. Few-Shot Stance Detection via Target-Aware Prompt Distillation. In Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, 11–15 July 2022; pp. 837–847.
49. Zhang, B.; Fu, X.; Ding, D.; Huang, H.; Li, Y.; Jing, L. Investigating Chain-of-Thought with ChatGPT for Stance Detection on Social Media. *arXiv* **2023**, arXiv:2304.03087.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.