Multi-modal Stance Detection: New Datasets and Model

Bin Liang[1,3,4*], **Ang Li**[1,3*], **Jingqian Zhao**[1,3], **Lin Gui**[5], **Min Yang**[6], **Yue Yu**[2], **Kam-Fai Wong**[4], **and Ruifeng Xu**[1,2,3†]
[1] Harbin Institute of Technology, Shenzhen, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
[4] The Chinese University of Hong Kong, Hong Kong, China [5] King's College London, UK
[6] SIAT, Chinese Academy of Sciences, Shenzhen, China
bin.liang@cuhk.edu.hk, angli@stu.hit.edu.cn, xuruifeng@hit.edu.cn

Abstract

Stance detection is a challenging task that aims to identify public opinion from social media platforms with respect to specific targets. Previous work on stance detection largely focused on pure texts. In this paper, we study multi-modal stance detection for tweets consisting of texts and images, which are prevalent in today's fast-growing social media platforms where people often post multi-modal messages. To this end, we create five new multi-modal stance detection datasets of different domains based on Twitter, in which each example consists of a text and an image. In addition, we propose a simple yet effective Targeted Multi-modal Prompt Tuning framework (TMPT), where target information is leveraged to learn multi-modal stance features from textual and visual modalities. Experimental results on our five benchmark datasets show that the proposed TMPT achieves state-of-the-art performance in multi-modal stance detection.

1 Introduction

Stance detection is an important task for learning public opinion from social media platforms, which aims to determine people's opinionated standpoint or attitude (*e.g.*, *Favor*, *Against*, or *Neutral*, etc.) expressed in the content towards a specific target, topic, or proposition (Somasundaran and Wiebe, 2010; Augenstein et al., 2016). Existing conventional machine learning-based methods (Hasan and Ng, 2013; Mohammad et al., 2016; Ebrahimi et al., 2016) and deep learning-based methods (Augenstein et al., 2016; Sun et al., 2018; Zhang et al., 2020; Chen et al., 2021; Allaway et al., 2021; Liang et al., 2022a) have made promising progress in different types of stance detection tasks for pure texts.

Target: Donald Trump **Stance:** Against
Target: Joe Biden **Stance:** Favor
Text: It is exactly what I want to say!!!
Image:

Figure 1: An example of a user expressing an "Against" stance towards "Donald Trump" and a "Favor" stance towards "Joe Biden" using multi-modal information.

However, more and more present-day social media platforms like Twitter allow people to post multi-modal messages, which encourages people to express their stances and opinions through multi-modal content, posting texts with images for example. That is, detecting stance from the pure text modality may not accurately identify the user's real view of a target. For example, Figure 1 shows a post composed of a text and an image. The stance expression towards "Donald Trump" and "Joe Biden" in this example can not be accurately identified based on text information unless combined with the information of the visual modality. Therefore, how to detect users' stances on a topic from multi-modal posts might help better identify public opinion in social media.

For multi-modal stance identification, Weinzierl and Harabagiu (2023) discussed the multi-modal stance towards frames of communication. Different from conventional stance detection tasks that concentrate on the stance towards several predefined targets, their focus was centered on the frames of communication within multi-modal posts. Therefore, aiming to push forward the research of multi-

[*] The first two authors contribute equally to this work.
[†] Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2024, pages 12373–12387

modal stance detection, we create five new datasets, in which each example consists of a target, a text, and an image. These datasets contain a total of 17,544 examples across 5 domains and 12 targets, including hot topics, politicians, and debates.

To deal with multi-modal stance detection, we propose a simple yet effective Targeted Multi-modal Prompt Tuning framework (TMPT), where the targeted prompt tuning is employed to adapt pre-trained models for learning stance features from different modalities. Specifically, to leverage the target information in stance detection, we first devise targeted prompts for both textual and visual modalities. Then, the targeted prompts are fed to the pre-trained language model and pre-trained visual model to learn stance features for the target from different modalities. Further, a simple vector concatenation is used to fuse the features from different modalities for multi-modal stance detection.

The main contributions of our work are summarized as follows:

1) We manually annotate five new multi-modal stance detection datasets based on Twitter data from different domains. The release of the datasets would push forward the research in this field.

2) A simple yet effective targeted multi-modal prompting tuning framework is proposed to deal with multi-modal stance detection, where the target information is used to prompt the pre-trained models for learning multi-modal stance features.

3) A series of experiments on our datasets show that the proposed method significantly outperforms the baseline models[1].

2 Related Work

Textual Stance Detection Various methods based on conventional machine learning (Hasan and Ng, 2014; Mohammad et al., 2016) and deep learning (Sun et al., 2018; Zheng et al., 2022; Li and Caragea, 2023; Li et al., 2023) have been proposed to deal with the stance detection regarding a specific target. For performing stance detection in real-world scenarios, many existing methods focus on the task of zero-shot stance detection[2](Liu et al., 2021b; Liang et al., 2022b; Wen and Hauptmann, 2023; Zhao et al., 2023).

Prompt Tuning Prompting (Liu et al., 2021a) was originally aimed to design language instructions for pre-trained language models (PLMs) to transfer learning in downstream tasks (Shin et al., 2020; Jiang et al., 2020). Recent works begin to treat prompts as continuous vectors and optimize them during fine-tuning, called Prompt Tuning (Li and Liang, 2021; Lester et al., 2021; Liu et al., 2021c). Besides, Radford et al. (2021); Zhou et al. (2022); Ju et al. (2022); Jia et al. (2022) introduce prompt into vision-language models to leverage the ability of prompt tuning in multi-modal tasks.

3 Multi-modal Stance Detection Datasets

Based on three open-source textual stance detection datasets: Twitter Stance Election 2020 (Kawintiranon and Singh, 2021), COVID-CQ (Mutlu et al., 2020), and Will-They-Won't-They (Conforti et al., 2020), and two hot topics in recent years: Russo-Ukrainian Conflict[3] and Taiwan Question[4]. We create five multi-modal stance detection datasets of different domains to provide available data for this task: Multi-modal Twitter Stance Election 2020 (MTSE), Multi-modal COVID-CQ (MCCQ), Multi-modal Will-They-Won't-They (MWTWT), Multi-modal Russo-Ukrainian Conflict (MRUC) and Multi-modal Taiwan Question (MTWQ).

3.1 Data Collection

We use Twitter Streaming API[5] to collect tweets with corresponding keywords of five datasets (shown in Appendix A), keeping the posts containing text in English and at least one image or video/GIF. For videos/GIFs, we retain their first frame because the visual information contained in consecutive frames may be very similar. For posts with multiple images, we combine the text with each image to form multiple samples due to the fact that different images may contain completely different visual information. Finally, we obtain 130000, 90000, 60000, 100000, and 80000 candidate examples of five datasets, respectively.

3.2 Data Annotation

To ensure consistency with previous stance detection work, we follow the guidelines of Twitter

[1]To facilitate future research, our datasets and code are publicly available at https://github.com/Leon-Francis/Multi-Modal-Stance-Detection

[2]Following (Allaway and McKeown, 2020), the term "zero-shot" here refers to the model's ability to detect stance towards targets it has not encountered during training.

[3]https://en.wikipedia.org/wiki/Russo-Ukrainian_War

[4]https://en.wikipedia.org/wiki/Cross-Strait_relations

[5]https://developer.twitter.com/en/products/twitter-api

Dataset	Target	# Samples and Proportion of Labels								
MTSE		Favor	%	Against	%	Neutral	%	-	-	Total
	Donald Trump (DT)	231	14.03	1297	78.75	119	7.23	-	-	1647
	Joe Biden (JB)	602	47.78	524	41.59	134	10.63	-	-	1260
MCCQ		Favor	%	Against	%	Neutral	%	-	-	Total
	Chloroquine (CQ)	455	33.58	503	37.12	397	29.30	-	-	1355
MWTWT		Support	%	Refute	%	Comment	%	Unrelated	%	Total
	CVS_AET	426	24.38	65	3.72	866	49.57	390	22.32	1747
	CI_ESRX	321	35.71	91	10.12	298	33.15	189	21.02	899
	ANTM_CI	59	5.01	238	20.22	306	26.00	574	48.77	1177
	AET_HUM	94	9.82	287	29.99	267	27.90	309	32.29	957
	DIS_FOXA	417	13.97	103	3.45	1504	50.37	962	32.22	2986
MRUC		Support	%	Oppose	%	Neutral	%	-	-	Total
	Russia (RUS)	16	1.4	763	68.74	331	29.82	-	-	1110
	Ukraine (UKR)	742	68.64	39	3.61	300	27.75	-	-	1081
MTWQ		Support	%	Oppose	%	Neutral	%	-	-	Total
	Mainland of China (MOC)	339	24.27	834	59.70	224	16.03	-	-	1397
	Taiwan of China (TOC)	1595	82.73	74	3.84	259	13.43	-	-	1928

Table 1: Label distribution of the five multi-modal stance detection datasets.

	Image	Person	Events	Words	Memes	Mixed
MTSE	43.7	32.7	8.2	24.0	25.0	10.1
MCCQ	44.1	12.2	22.9	26.7	16.3	21.9
MWTWT	46.2	26.4	16.8	38.4	5.9	12.5
MRUC	54.8	32.1	20.7	20.4	18.5	8.3
MTWQ	41.0	36.8	42.9	8.1	2.4	9.8

Table 2: The statistics of whether the image conveys stance information (%Image) and the type of each image (%Person, %Events, %Words, %Memes, %Mixed).

Stance Election 2020 (Kawintiranon and Singh, 2021), COVID-CQ (Mutlu et al., 2020), and Will-They-Won't-They (Conforti et al., 2020) to annotate the multi-modal stance of MTSE, MCCQ and MWTWT. For MRUC and MTWQ, the annotation guidelines are shown in Appendix B. The meaningless/noisy posts or those that do not comply with Twitter's policies or annotation guidelines are discarded during the annotation. We invite eight experienced researchers[6] to label the stance for each example. Each sample will be annotated by three different annotators, and the gold label is obtained by majority vote. For the disagreed results among the three annotators, we invited three additional annotators to annotate and then performed a majority vote to obtain the gold label[7].

3.3 Quality Assessment

We use Cohen's Kappa Statistic to evaluate the inter-annotator agreement (Cohen, 1960). The average Cohen's Kappa between our annotator pairs for MTSE is 0.703, for MCCQ is 0.689, for MWTWT is 0.729, for MRUC is 0.752, and for MTWQ is 0.691. This demonstrates that the Kappa scores of all datasets are substantial. In addition, the average Cohen's Kappa reported in the related textual stance detection dataset Will-They-Won't-They (Conforti et al., 2020) is 0.67, which also indicates the high quality of our new datasets from another angle.

3.4 Data Analysis

We finally got 17,544 well-annotated multi-modal samples across 5 domains and 12 targets. Each example consists of a text in English and an associated image. The statistics of datasets are reported in Table 1. Note that differences in label distribution between targets are common, which is also observed in other textual stance detection datasets (Mohammad et al., 2016; Mutlu et al., 2020; Conforti et al., 2020). Through analyzing the indication of whether the image conveys stance information introduced in Section 3.2, and analyzing the type of each image, we obtained statistics shown in Table 2. It can be seen that nearly half of the stance information in all five datasets comes from visual modalities. Further, the proportion of image types varies greatly among different datasets. This complicates image comprehension for stance detection, which is also the main challenge of multi-modal stance detection.

4 Methodology

In this section, we introduce our proposed targeted multi-modal prompt tuning framework (TMPT) in

[6]We recruit experienced researchers who have worked on multi-modal learning over 3 years.

[7]During the annotation process, only 2.54% of the data need to be allocated to additional annotators.

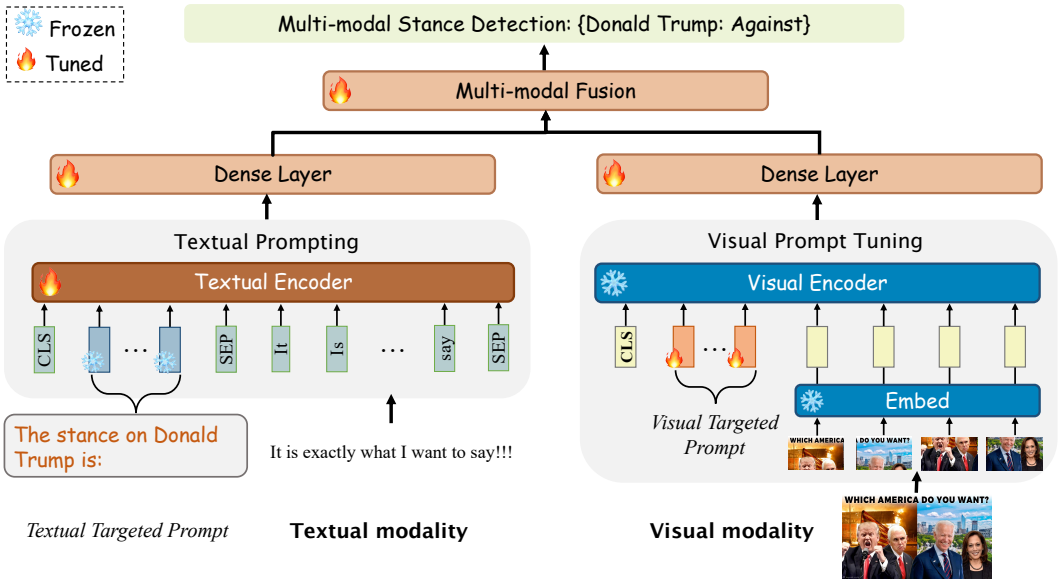


Figure 2: The overall architecture of our proposed TMPT. Textual Prompting is devised for adapting the large pre-trained language model. Visual Prompt Tuning is devised for adapting the large pre-trained vision model.

Dataset	Target	Textual Targeted Prompt
MTSE	DT	The stance on Donald Trump is:
MCCQ	CQ	The stance on the use of Chloroquine and Hydroxychloroquine for the treatment or prevention from the coronavirus or COVID 19 is:
MWTWT	CVS_AET	The stance on merger and acquisition between CVS Health and Aetna is:
MRUC	RUS	The stance on Russia is:
MTWQ	MOC	The stance on Mainland of China is:

Table 3: The example of textual targeted prompts.

detail. Given a text S and an image I, the goal of multi-modal stance detection is to identify the stance label y for the specific target t based on S and I. Therefore, to leverage the target information for multi-modal stance detection, we design targeted multi-modal prompt tuning for both textual and visual modalities, which are utilized to prompt the pre-trained models for learning multi-modal stance features. The architecture of our TMPT is illustrated in Figure 2, containing four main components: 1) *Textual Prompt Tuning*, which encodes the input of textual modality based on the textual targeted prompt; 2) *Visual Prompt Tuning*, which encodes the input of visual modality based on the visual targeted prompt; 3) *Multi-modal Fusion*, which fuses the representations from textual and visual modalities to capture the stance features. 4) *Multi-modal Stance Detection*, which derives the stance label for an input example according to the multi-modal stance features.

4.1 Textual Prompt Tuning

Textual Prompt Construction Inspired by the textual prompt tuning (Liu et al., 2023), considering the characteristics of stance detection, we devise a textual targeted prompt for each text to adapt the pre-trained language model to the stance detection task. As the example shown in Figure 2, take the target "Donald Trump" as an example, the textual targeted prompt is defined as:

$$\mathcal{P}^T = \text{The stance on Donald Trump is :} \quad (1)$$

That is, the textual targeted prompts are designed according to the targets and the purpose of stance detection. The examples of textual targeted prompts are shown in Table 3. Other textual prompt settings are introduced in Section C.1.

Textual Encoder Based on the textual targeted prompts, we construct the input of each text on target t. Given a text S consists of a sequence of words $S = \{w_i\}_{i=1}^{n}$, n is the length of S. The input of the textual modality is represented as:

$$\mathcal{I}^T = \texttt{[CLS]}\mathcal{P}^T\texttt{[SEP]}s\texttt{[SEP]} \quad (2)$$

Then, we adopt the pre-trained uncased BERT-base model (Devlin et al., 2019) to map each textual token into a d^T-dimensional embedding:

$$[\boldsymbol{x}^T_{\texttt{[CLS]}}, \boldsymbol{x}^T_1, \cdots, \boldsymbol{x}^T_{m+n+2}] = \text{BERT}(\mathcal{I}^T) \quad (3)$$

Where $\boldsymbol{x}^T_{\texttt{[CLS]}}$ represents the embedding of the [CLS] token, m is the length of the textual targeted prompt. In this way, the pre-trained language

model can encode the input text according to the target, and obtain the feature representation that contains the target-specific stance information.

4.2 Visual Prompt Tuning

Visual Embedding Layer Following (Dosovitskiy et al., 2021), we first divide the image I into r fixed-sized patches $I = \{\boldsymbol{p}_i \in \mathbb{R}^{l^2 \times c}\}_{i=1}^{r}$, where (l, l) is the resolution of each patch, c is the number of channels. Then, following (Jia et al., 2022), we flatten the patches and map to d^V-dimensional vector with a trainable linear embedding projection $\boldsymbol{E}$. We refer to the output of this projection as the patch embeddings. Then we concatenate these patch embeddings in the sequence dimension and added the standard learnable 1D position embeddings $\boldsymbol{E}_{pos}$ to the patch embeddings to retain positional information:

$$\boldsymbol{v}_j = \boldsymbol{p}_j \boldsymbol{E} \qquad \boldsymbol{v}_j \in \mathbb{R}^d, j \in \{1, r\} \tag{4}$$

$$\boldsymbol{V}^0 = [\{\boldsymbol{v}_j\}_{j=1}^{r}] + \boldsymbol{E}_{pos} \tag{5}$$

Where $\boldsymbol{V}^0$ is the embedding for the input image.

Visual Prompt Construction Inspired by the visual prompt tuning proposed by (Jia et al., 2022), we devise visual targeted prompt, aiming to instruct the pre-trained vision model to learn the features according to the specific target. Specifically, we introduce continuous embedding (visual prompt tokens), as the visual prompt for target t. Each prompt consists of λ learnable embedding, which can be formulated as follows:

$$\mathcal{P}^V = \{\boldsymbol{e}_i \in \mathbb{R}^{d^V} | 1 \leq i \leq \lambda\} \tag{6}$$

Here, each different target corresponds to a different set of visual prompt embedding. In Section C.2, we introduced different initialization methods of visual prompt embedding.

Visual Encoder Based on the visual embedding $\boldsymbol{V}^0$ and the visual targeted prompts $\mathcal{P}^V$, we use the pre-trained Vision Transformer model ViT (Dosovitskiy et al., 2021) with N layers to encode the input image for learning visual stance features on target t. Here, the targeted prompts are inserted into the first Transformer layer L^1. Therefore, the first layer targeted prompted ViT is defined as:

$$[\boldsymbol{x}^V_{\texttt{[CLS]}_1}, \boldsymbol{Z}_1, \boldsymbol{V}_1] = L^1([\boldsymbol{x}^V_{\texttt{[CLS]}_0}, \mathcal{P}^V, \boldsymbol{V}_0]) \tag{7}$$

For layer $k \in \{2, N\}$, the targeted prompted ViT is defined as:

$$[\boldsymbol{x}^V_{\texttt{[CLS]}_k}, \boldsymbol{Z}_k, \boldsymbol{V}_k] = L^k([\boldsymbol{x}^V_{\texttt{[CLS]}_{k-1}}, \boldsymbol{Z}_{k-1}, \boldsymbol{V}_{k-1}]) \tag{8}$$

Here, we use $\boldsymbol{x}^V_{\texttt{[CLS]}}$ to represent the embedding of `[CLS]` token learned by the final Transformer layer N, *i.e.*, $\boldsymbol{x}^V_{\texttt{[CLS]}} = \boldsymbol{x}^V_{\texttt{[CLS]}_N}$. By using the visual prompts corresponding to each target, the pre-trained visual model can encode the input image according to the target, and obtain the feature representation that contains the target-specific stance information.

4.3 Multi-modal Fusion

Based on the embeddings of `[CLS]` tokens, we first use two dense layers with Leaky ReLU (Maas et al., 2013) to derive the hidden representations of textual and visual modalities:

$$\boldsymbol{h}^T = f(\boldsymbol{W}^T \boldsymbol{x}^T_{\texttt{[CLS]}} + \boldsymbol{b}^T) \tag{9}$$

$$\boldsymbol{h}^V = f(\boldsymbol{W}^V \boldsymbol{x}^V_{\texttt{[CLS]}} + \boldsymbol{b}^V) \tag{10}$$

Where $\boldsymbol{h}^T \in \mathbb{R}^{d^h}$ and $\boldsymbol{h}^V \in \mathbb{R}^{d^h}$ are the hidden representation of textual and visual modalities respectively. d^h is the dimensionality of the hidden representation. $f(\cdot)$ represents the Leaky ReLU activation function. $\boldsymbol{W}^T \in \mathbb{R}^{d^h \times d^T}$ and $\boldsymbol{W}^V \in \mathbb{R}^{d^h \times d^V}$ are weight matrices. $\boldsymbol{b}^T \in \mathbb{R}^{d^h}$ and $\boldsymbol{b}^V \in \mathbb{R}^{d^h}$ are biases.

In our `TMPT`, vector concatenation is used to fuse the feature vectors from different modalities:

$$\boldsymbol{h} = \boldsymbol{h}^T \oplus \boldsymbol{h}^V \tag{11}$$

Where $\boldsymbol{h} \in \mathbb{R}^{2d^h}$ is the final multi-modal stance representation. $\oplus$ represents the vector concatenation operation.

4.4 Multi-modal Stance Detection

Then, the final multi-modal stance representation is fed into a fully-connected layer with a softmax function to capture a probability distribution $\hat{\boldsymbol{y}} \in \mathbb{R}^{d^p}$ in the stance decision space:

$$\hat{\boldsymbol{y}} = \text{softmax}(\boldsymbol{W}^o \boldsymbol{h} + \boldsymbol{b}^o) \tag{12}$$

Where d^p is the dimensionality of stance labels. $\boldsymbol{W}^o \in \mathbb{R}^{d^p \times 2d^h}$ and $\boldsymbol{b}^o \in \mathbb{R}^{d^p}$ are weight matrix and bias respectively.

5 Experimental Setup

To advance and facilitate research in the field of multi-modal stance detection, based on our new datasets, we conduct experiments on **In-target Multi-modal Stance Detection**, training and testing on the same target, and **Zero-shot Multi-modal Stance Detection**, performing stance detection on unseen targets based on the known targets.

Task	Dataset	Target	# Train	# Valid	# Test
In-target	MTSE	DT	1150	170	327
		JB	882	128	250
	MCCQ	CQ	934	141	280
	MWTWT	CSV_AET	1216	179	352
		CI_ESRX	628	91	180
		ANTM_CI	825	114	238
		AET_HUM	674	97	186
		DIS_FOXA	2081	306	599
	MRUC	RUS	777	111	222
		UKR	756	108	217
	MTWQ	MOC	977	140	280
		TOC	1349	193	386
Zero-shot	MTSE	DT	1114	146	1647
		JB	1434	212	1260
	MWTWT	CVS_AET	5253	737	1747
		CI_ESRX	5994	841	899
		ANTM_CI	5694	804	1177
		AET_HUM	5884	840	957
	MRUC	RUS	945	136	1110
		UKR	971	139	1081
	MTWQ	MOC	1686	242	1397
		TOC	1222	175	1928

Table 4: Statistics of the experimental data.

5.1 Data Partition

For in-target multi-modal stance detection, each dataset is divided into training, development, and testing sets with a ratio of 7:1:2[8]. For zero-shot multi-modal stance detection, the training and development set is built on known target(s), and the testing set is built on unknown target(s). Since there is only one target in the MCCQ dataset, we only use it for in-target multi-modal stance detection. In the MWTWT dataset, following (Conforti et al., 2020), we select four targets (CVS_AET, CI_ESRX, ANTM_CI, and AET_HUM) to perform zero-shot scenario since DIS_FOXA is not in the same domain as them. The statistics of each dataset are shown in Table 4.

5.2 Comparison Models

Pure textual modality baselines: 1) BERT (Devlin et al., 2019), the uncased BERT-base; 2) RoBERTa (Liu et al., 2019a), the RoBERTa-base; 3) KEBERT (Kawintiranon and Singh, 2022), a BERTweet-base model with specific knowledge of Twitter political posts. 4) LLaMA2 (Touvron et al., 2023), the LLaMA2-70b-chat; 5) GPT4 [9]. **Pure visual modality baselines**: 1) ResNet (He et al., 2016), the ResNet-50 v1.5; 2) ViT (Dosovitskiy et al., 2021), the vit-base-patch16-224; 3) Swin Transformer (SwinT) (Liu et al., 2021d), the swinv2-base-patch4-window12-192-22k. **Multi-modal baselines**: 1) ViLT (Kim et al., 2021), the vilt-b32-mlm; 2) CLIP (Radford et al., 2021), the clip-vit-base-patch32; 3) BERT+ViT, utilizing BERT as the textual encoder and ViT as the visual encoder, and concatenating the `[CLS]` vectors of textual and visual modalities for stance detection. 4) Qwen-VL (Bai et al., 2023), the Qwen-VL-Chat-7b. 5) GPT4-Vision [10].

5.3 Experimental Settings

To leverage the powerful capabilities of large language models, following Gatto et al. (2023), we propose a variant of our TMPT model, named `TMPT+CoT`. By utilizing GPT4-Vision to generate a chain of thought from the text and image of the sample, we concatenate this chain of thought with the text to serve as the textual modality input for `TMPT+CoT`. The images are used as the visual modality input. This approach is employed for both the training and testing phases. We utilize the pre-trained uncased BERT-base (Devlin et al., 2019) to embed each word as a 768-dimensional embedding and employ the pre-trained ViT-base (Dosovitskiy et al., 2021) to embed each image patch as a 768-dimensional embedding, *i.e.*, $d^T = d^V = 768$. The resolution of the visual patch is set to $(16, 16)$. The length of visual prompt tokens is set to $c = 7$. The dimensionality of hidden vectors is set to $d_h = 768$. We use *Macro F1-score* to measure the model performance. The experimental results of our models are averaged over 5 runs to ensure the final reported results are statistically stable. For detailed settings of the experiments, please refer to Appendix C.

6 Experimental Results

6.1 In-target Multi-modal Stance Detection

The results of in-target multi-modal stance detection are shown in Table 5. It can be seen that our proposed `TMPT` outperforms fine-tuning based baselines on most of datasets, denoting that exploiting targeted prompt tuning can preferably leverage the target-specific multi-modal stance information, thus improving stance detection performance. Further, `TMPT+CoT` performs overall better than `TMPT`, which demonstrates that leveraging the knowledge

[8] To address the possibility of one text corresponding to multiple images, we use the Twitter ID to split datasets, thus avoiding the issue of data leakage.

[9] https://openai.com/research/gpt-4

[10] https://openai.com/research/gpt-4v-system-card

Modality	Method	MTSE		MCCQ	MWTWT					MRUC		MTWQ	
		DT	JB	CQ	CA	CE	AC	AH	DF	RUS	UKR	MOC	TOC
Textual	BERT	48.25	52.04	66.57	75.62	60.85	63.05	59.24	81.53	41.25	46.80	57.77	45.91
	RoBERTa	58.39	60.79	66.57	69.56	65.03	69.74	67.99	79.21	39.52	57.66	55.22	48.88
	KEBERT	64.50	69.81	66.84	71.67	67.56	69.29	69.74	80.57	41.55	59.01	58.15	47.75
	LLaMA2	53.23	52.67	47.40	34.89	41.95	49.09	44.32	30.21	38.84	38.54	55.31	46.51
	GPT4	68.74	66.39	65.84	63.14	65.12	**69.93**	**71.62**	52.69	41.64	53.76	58.05	49.81
Visual	ResNet	37.89	38.59	47.16	39.89	42.20	43.52	37.05	50.34	35.10	40.00	42.02	33.94
	ViT	40.48	40.42	46.64	46.63	50.00	40.16	46.32	50.86	33.31	39.87	38.63	35.53
	SwinT	39.89	40.43	48.80	46.30	46.99	41.02	47.39	51.32	35.01	40.89	35.03	35.47
Multi-modal	BERT+ViT	41.86	45.82	61.32	63.20	44.71	56.45	46.85	73.71	39.28	48.41	47.47	40.86
	ViLT	35.32	48.24	47.85	62.70	56.44	58.06	60.22	73.66	34.62	42.41	44.43	59.51
	CLIP	53.22	65.83	63.65	70.93	67.17	67.43	70.86	79.06	44.99	59.86	55.29	40.98
	Qwen-VL	43.31	45.13	50.51	43.06	45.49	49.79	46.04	27.73	36.50	40.78	42.14	39.34
	GPT4-Vision	**70.46**	**72.82**	61.63	44.59	47.07	57.47	57.90	37.61	44.83	56.40	66.72	56.90
	TMPT	55.41	61.61	67.67	**76.60**	63.19	67.25	62.92	81.19	43.56	59.24	55.68	46.82
	TMPT+CoT	66.61	68.75	**71.79***	74.40	**69.96***	68.43	63.00	**82.71***	**45.04***	**60.52**	**68.95***	**59.87***

Table 5: Experimental results (%) of in-target multi-modal stance detection. The dark background results are for our TMPT. Best scores of each group are in bold. Results with $\star$ denote the significance tests of our TMPT over the baseline models at $p-\text{value} < 0.05$. The dashed line represents a separation between fine-tuned methods, non-fine-tuned LLM-type methods, and our TMPT.

Modality	Method	MTSE		MWTWT				MRUC		MTWQ	
		DT	JB	CA	CE	AC	AH	RUS	UKR	MOC	TOC
Textual	BERT	32.52	29.97	63.55	61.30	59.18	52.89	22.01	15.45	28.04	9.57
	RoBERTa	26.60	32.21	59.22	59.22	64.86	57.46	27.10	19.98	30.62	15.84
	KEBERT	26.17	31.81	59.70	62.56	63.92	55.53	24.68	28.18	29.17	19.80
	LLaMA2	53.57	53.92	32.47	38.37	48.08	46.13	31.86	36.34	51.46	44.10
	GPT4	70.78	68.83	57.19	60.56	65.63	**69.01**	40.22	49.18	62.10	52.12
Visual	ResNet	25.52	29.70	23.01	24.11	25.21	25.27	23.88	25.57	27.59	24.88
	ViT	28.63	29.70	24.59	28.18	34.06	33.40	27.26	28.51	29.37	23.69
	SwinT	28.54	30.85	28.53	28.50	35.87	34.33	25.44	24.54	27.90	19.69
Multi-modal	BERT+ViT	26.70	31.57	59.21	59.30	65.04	59.28	23.33	15.21	24.76	11.70
	ViLT	28.08	29.74	38.33	46.00	55.01	48.55	21.56	23.96	23.54	19.18
	CLIP	28.21	28.99	61.08	55.67	63.80	60.06	25.62	27.40	27.21	15.69
	Qwen-VL	47.62	46.14	38.57	43.36	47.82	41.01	36.95	41.39	44.32	44.08
	GPT4-Vision	**72.68**	**71.28**	42.23	45.92	54.59	53.19	42.09	47.00	**65.00**	**52.36**
	TMPT	31.69	32.65	66.36	**66.39***	**66.32**	61.56	23.87	24.71	32.18	26.48
	TMPT+CoT	54.30	58.46	**67.28***	63.73	64.87	54.26	**48.99***	**51.75***	45.32	43.70

Table 6: Experimental results of zero-shot multi-modal stance detection. The dark background results are for our TMPT. Best scores of each group are in bold. Results with $\star$ denote the significance tests of our TMPT over the baseline models at $p-\text{value} < 0.05$. The dashed line represents a separation between fine-tuned methods, non-fine-tuned LLM-type methods, and our TMPT.

from large language models can improve the comprehension of textual modality, thereby achieving better performance. For uni-modal baselines, the performance of visual modality methods is unsatisfactory, while the performance of textual modality methods is much better. This indicates that the stance expression primarily resides in the textual modality. Further, CLIP, which considers both textual and visual modalities performs overall better than models which only consider textual modality, which proves the importance of visual modality in multi-modal stance detection. While, in datasets like MWTWT and MCCQ, GPT4-Vision underperforms GPT-4. Our analysis of images within these datasets revealed that a noteworthy proportion of images had comparatively high complexity levels. This also implies effective use of visual information is key to improving multi-modal stance detection performance. In addition, KEBERT, which integrates specific knowledge of Twitter political posts into BERTweet, achieves promising performance compared to other textual models, which indicates that exploring external target-related knowledge might improve the performance of multi-modal stance detection.

METHOD	MTSE	MCCQ	MWTWT	MRUC	MTWQ
TMPT	**60.84**	**67.67**	**77.59**	**75.76**	**67.59**
w/o $\mathcal{P}^T$	54.93	62.76	71.42	70.53	61.77
w/o $\mathcal{P}^V$	58.14	65.71	73.93	70.85	63.69

Table 7: Experimental results of ablation study. The reported results are the Macro F1-score across all targets in a dataset on in-target multi-modal stance detection.

6.2 Zero-shot Multi-modal Stance Detection

The results of zero-shot multi-modal stance detection are reported in Table 6. It can be seen that the large language models achieve superior performance due to the need for detecting stances on unseen targets. This may be attributed to the powerful zero-shot learning ability of large language models. For our TMPT, which does not use chain-of-thoughts from large language models, still achieves better performance than large language models in some targets, while also outperforming most of non-large language model baselines. This shows the promise of our TMPT in zero-shot stance detection. Further, TMPT+CoT performs overall better than TMPT, which indicates that obtaining powerful text and visual comprehension abilities from large models may be key to improving the performance of detecting stance for unseen targets.

6.3 Ablation Study

To analyze the impact of the targeted prompt tuning in our proposed TMPT, we conduct an ablation study and report the results in Table 7. Note that the removal of textual prompting (w/o $\mathcal{P}^T$) sharply degrades the performance, which verifies the significance of textual prompting in learning textual targeted stance features for multi-modal stance detection. In addition, the removal of visual prompt tuning (w/o $\mathcal{P}^V$) leads to considerable performance degradation, which indicates that utilizing visual prompt tuning can make better learning of visual targeted stance information and thus improves the performance of multi-modal stance detection.

6.4 Generalization of Targeted Multi-modal Prompt Tuning

Pre-trained Models Previous experiments have demonstrated that the stance expression primarily resides in the textual modality. Therefore, to investigate the generalization of our TMPT when used with different pre-trained language models, we conduct experiments with two variants of our TMPT by using two other promising PLMs: RoBERTa (Liu et al., 2019b) and KEBERT (Kawintiranon and Singh, 2021). The results are shown in Figure 3. Note that our Targeted Multi-modal Prompting can directly work with the two PLMs and achieve better performance, showing the compatibility of our method with various pre-trained models.

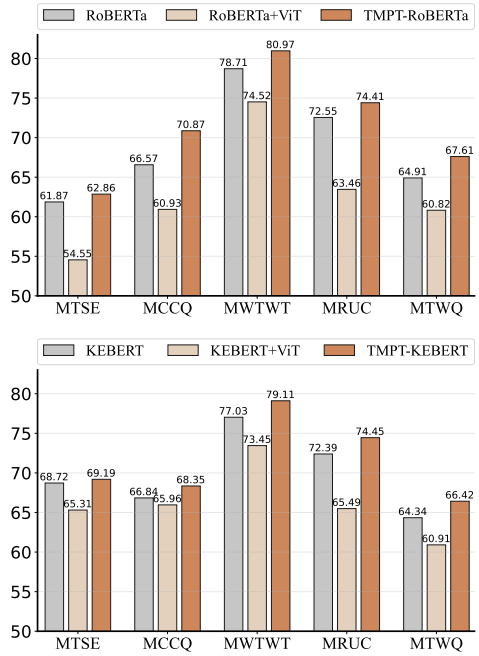


Figure 3: Performance of using different pre-trained language models: RoBERTa (Nguyen et al., 2020) (top) and KEBERT (Kawintiranon and Singh, 2021) (bottom). The reported results are the Macro F1-score across all targets in a dataset on in-target multi-modal stance detection.

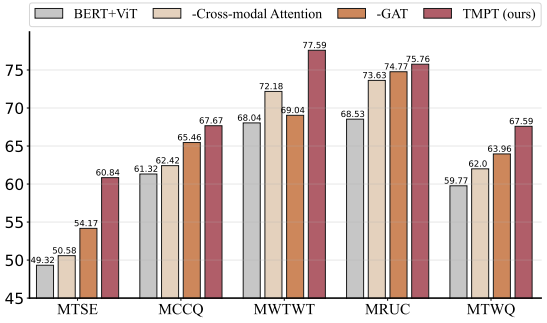


Figure 4: Performance of using different multi-modal fusion methods. The reported results are the Macro F1-score across all targets in a dataset on in-target multi-modal stance detection.

Multi-modal Fusion We choose cross-modal attention (Wei et al., 2020) and GAT (Velickovic et al., 2018) to analyze the performance of our TMPT with different multi-modal fusion methods. The results are shown in Figure 4. It can be seen

	Person	Events	Words	Memes	Mixed
ERR-Img(%)	7.8	13.4	37.7	26.6	14.5
ERR-Img-prop	0.28	0.60	1.60	1.95	1.16

Table 8: The results of error analysis on image types. ERR-Img means the proportion of different image types among the incorrect samples with stance-related content. ERR-Img-prop means ERR-Img divided by the proportion of its image types in the original dataset.

that no matter which fusion method we use, the experimental performance is better than BERT+ViT. This indicates that our TMPT can directly work with various multi-modal fusion methods and lead to improved multi-modal stance detection performance.

6.5 Error Analysis

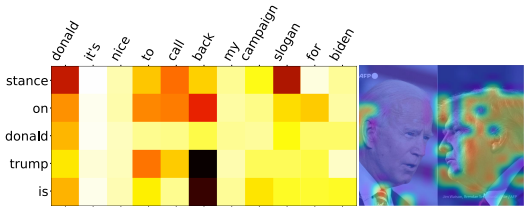
From the results in Table 6, we can see that our TMPT performs well in the MWTWT dataset, but still insufficient in other datasets on zero-shot stance detection. One possible reason is that targets in MWTWT are all about expressing views on corporate mergers, so unknown targets can easily find commonalities in the dataset. However, for other datasets, the topics are diverse, which poses a challenge to mining targeted information. Therefore, how to better learn the correlation information between targets from data on diverse topics is a potential direction to improve the performance of the zero-shot scenario.

Further, after analyzing examples of misclassification, we found that among the incorrect samples, approximately 70% (based on calculations on randomly sampled 300 incorrect samples from five datasets) of the images contained stance-related content. This indicates that for our multi-modal stance detection task, it is important for further exploration in extracting and utilizing features from the visual modality.

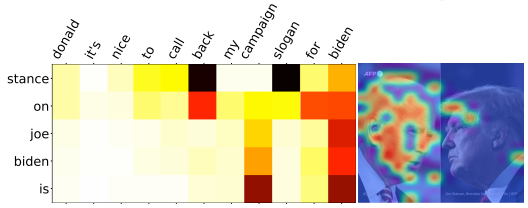
Building on the previous step, we conduct error analysis on different image types. The results are illustrated in Table 8. A larger ERR-Img-prop indicates a weaker ability of the model to handle samples with images of that category, indicating that for images containing words, memes, and mixed features, more effective methods of feature extraction and understanding remain to be proposed.

6.6 Visualization

To qualitatively investigate how our TMPT improves the performance of multi-modal stance detection, we visualize the attention values calculated by the targeted prompts and the vectors of the final layer of two encoders. The results are shown in Figure 5. It can be seen that the crucial textual tokens and the key visual patches regarding different targets are highly attended to and discriminated by our TMPT. This illustrates that the proposed TMPT can learn the important stance features for the specific target with the help of targeted prompt tuning, thus improving the learning ability of multi-modal stance detection.



(a) Attention Visualization of target "Donald Trump".



(b) Attention Visualization of target "Joe Biden".

Figure 5: Visualization of a typical example.

7 Conclusion

In this paper, we present MTSE, MCCQ, MWTWT, MRUC and MTWQ, five new datasets for multi-modal stance detection. Based on the created datasets, we present in-target multi-modal stance detection and zero-shot multi-modal stance detection, aiming to advance and facilitate research in the field of multi-modal stance detection. In addition, we propose a simple yet effective targeted multi-modal prompting framework (TMPT) to deal with multi-modal stance detection, where the target information is explored to prompt the pre-trained models in learning multi-modal stance features. Extensive experiments on the new datasets show that our proposed TMPT achieves overall better performance than state-of-the-art baseline methods.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (62176076), Natural Science Foundation of GuangDong 2023A1515012922, the Shenzhen Foundational Research Funding (JCYJ20220818102415032, JCYJ20210324115614039), the Major Key Project of PCL2021A06, Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies 2022B1212010005.

Limitations

Our method needs to devise a specific prompt for the given target, which needs to take time and artificial effort to analyze the target information in the real-world scenario for designing and selecting appropriate prompts. Furthermore, our proposed method does not integrate external target-specific knowledge to improve the learning of multi-modal stance information, such as the background knowledge of targets. Integrating external knowledge related to the target can improve the performance of stance detection. In addition, the current version of the data does not consider audio modality and video information, which is also an issue we need to explore in the future.

Ethics Statement

This work presents MTSE, MCCQ, MWTWT, MRUC and MTWQ, five new open-source datasets for the research community to study multi-modal stance detection. The MTSE, MCCQ, and MWTWT are the extension of Twitter Stance Election 2020 (Kawintiranon and Singh, 2021), COVID-CQ (Mutlu et al., 2020), and Will-They-Won't-They (Conforti et al., 2020), which are three open-source textual stance detection datasets for academic research. We only collect the tweet text and image content needed for our research from Twitter following the privacy agreement of Twitter for academic usage, so there is no privacy issue. To annotate extended data, we recruited 8 experienced researchers who work on natural language processing or multi-modal learning. The detailed collect and annotate process has been illustrated in Section 3. Each researcher is paid $6.5 per hour (above the average local payment of similar jobs). The entire annotation process lasted 5 months, and the average annotation time of the eight researchers was 430 hours. During the annotation process, samples that contain personally identifiable information will be discarded, and only the tweet IDs and human-annotated stance labels will be shared. Thus, our data set complies with Twitter's information privacy policy. The annotators have no affiliation with any of the companies that are used as targets in the dataset, so there is no potential bias due to conflict of interest. We used the ChatGPT service from OpenAI for our writing. We followed their term and policies. Some examples in our paper may include a stance or tendency. It should be clarified that they are randomly sampled from the dataset for better studying the dataset and task, and do not represent any personal viewpoints.

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.

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

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, Austin, Texas. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966.

Pengyuan Chen, Kai Ye, and Xiaohui Cui. 2021. Integrating n-gram features into pre-trained model: a novel ensemble model for multi-target stance detection. In *International Conference on Artificial Neural Networks*, pages 269–279. Springer.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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

Joseph Gatto, Omar Sharif, and Sarah Preum. 2023. Chain-of-thought embeddings for stance detection on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4154–4161. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan. Asian Federation of Natural Language Processing.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *European Conference on Computer Vision (ECCV)*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models for efficient video understanding. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXV*, volume 13695 of *Lecture Notes in Computer Science*, pages 105–124. Springer.

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

Kornraphop Kawintiranon and Lisa Singh. 2022. Polibertweet: A pre-trained language model for analyzing political content on twitter. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022, Marseille, France, 20-25 June 2022*, pages 7360–7367. European Language Resources Association.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 5583–5594. PMLR.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023. Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 15703–15717. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Yingjie Li and Cornelia Caragea. 2023. Distilling calibrated knowledge for stance detection. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6316–6329. Association for Computational Linguistics.

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

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022b. JointCL: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ArXiv preprint*, abs/2107.13586.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9):195:1–195:35.

Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021b. 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, Online. Association for Computational Linguistics.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021c. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *ArXiv preprint*, abs/2110.07602.

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

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

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021d. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Atlanta, Georgia, USA.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Ece C Mutlu, Toktam Oghaz, Jasser Jasser, Ege Tutunculer, Amirarsalan Rajabi, Aida Tayebi, Ozlem Ozmen, and Ivan Garibay. 2020. A stance data set on polarized conversations on twitter about the efficacy of hydroxychloroquine as a treatment for covid-19. *Data in brief*, 33:106401.

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, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 9–14. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124, Los Angeles, CA. Association for Computational Linguistics.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio.

2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. 2020. Multi-modality cross attention network for image and sentence matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10938–10947. IEEE.

Maxwell Weinzierl and Sanda Harabagiu. 2023. Identification of multimodal stance towards frames of communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12597–12609, Singapore. Association for Computational Linguistics.

Haoyang Wen and Alexander 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, Toronto, Canada. Association for Computational Linguistics.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197, Online. Association for Computational Linguistics.

Chenye Zhao, Yingjie Li, and Cornelia Caragea. 2023. C-STANCE: A large dataset for chinese zero-shot stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13369–13385. Association for Computational Linguistics.

Kai Zheng, Qingfeng Sun, Yaming Yang, and Fei Xu. 2022. Knowledge stimulated contrastive prompting for low-resource stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1168–1178. Association for Computational Linguistics.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348.

A Keywords

In this section, we introduce the keywords to retrieve tweets.

A.1 Multi-modal Twitter Stance Election 2020

- Since Twitter Stance Election 2020 (Kawintiranon and Singh, 2021) didn't explicitly give the Keywords for collecting tweets, We used the keywords related to the election while with no clear preference: one of *#vote, #Debates2020, #USElection2020, #PresidentialDebate2020, #2020Election, #votersuppression, #GetOuttheVote, #2020elections* + mention of Trump or Biden.

- Filter for posting time:

DT	01/01/2020 → 09/30/2020
JB	01/01/2020 → 09/30/2020

A.2 Multi-modal COVID-CQ

- We followed the keywords for collecting tweets from COVID-CQ (Mutlu et al., 2020): one of *hydroxychloroquine, chloroquine, HCQ*.

- Filter for posting time:

CQ	04/01/2020 → 04/30/2020

A.3 Multi-modal Will-They-Won't-They

- We followed the keywords for collecting tweets from Will-They-Won't-They (Conforti et al., 2020): one of *merge, acquisition, agreement, acquire, takeover, buyout, integration* + mention of a given company/acronym.

- Filter for posting time:

CVS_AET	02/15/2017 → 12/17/2018
CI_ESRX	05/27/2017 → 09/17/2018
ANTM_CI	04/01/2014 → 04/28/2017
AET_HUM	09/01/2014 → 01/23/2017
DIS_FOXA	07/09/2017 → 04/18/2018

A.4 Multi-modal Russo-Ukrainian Conflict

- We used the keywords: *Ukraine, Russia, Putin, Zelensky, Ukrainian, Russian*

- Filter for posting time:

RUS	01/01/2022 → 06/30/2023
UKR	01/01/2022 → 06/30/2023

A.5 Multi-modal Taiwan Question

- We used the keywords: *Taiwan, Taiwan Crisis, Nancy Pelosi, Taiwan Strait*

- Filter for posting time:

MOC	01/01/2022 → 06/30/2023
TOC	01/01/2022 → 06/30/2023

B Annotation Guidelines

To ensure consistency with previous stance detection work, we follow the guidelines of Twitter Stance Election 2020 (Kawintiranon and Singh, 2021), COVID-CQ (Mutlu et al., 2020), and Will-They-Won't-They (Conforti et al., 2020) to annotate the multi-modal stance of MTSE, MCCQ and MWTWT. For MRUC and MTWQ, the annotation guidelines are shown below:

B.1 Annotation Guidelines of MRUC and MTWQ

The annotation process consists of choosing one of three possible labels, given a tweet and an image. The three labels to choose from are Support, Oppose and Neutral.

Label 1: Support - If the tweet and image use direct or indirect expressions to support the target, or support objects which can represent the target (such as leaders, events).

Label 1: Oppose - If the tweet and image use direct or indirect expressions to oppose the target, or oppose objects which can represent the target (such as leaders, events).

Label 1: Neutral - If the tweet and image only mention the target without expressing a stance.

B.2 Details of Datasets Partition

As mentioned in Section 5.1, for each task/target, the dataset is divided into a training set, a development set, and a testing set with a ratio of 70%:10%:20%. In order to ensure that the partition is not affected by the data distribution bias. We performed 20 different random divisions for each task using the ratios above. We use the two baselines: BERT and BERT+ViT, to test every 20 groups of divisions. For each task, we take the division that can make the results of two baseline close to the median as the final partition.

C Prompt Tuning Analysis

C.1 Analysis of Textual Prompt Tuning

Frozen vs Tuned For textual prompt tuning, we utilize a frozen paradigm to make full use of the semantic information learned by the pre-trained language model. To analyze the effectiveness of the frozen paradigm, we also tried a tuned soft prompt, the results are shown in Table 9. The results of the tuned paradigm are extremely poorer than the frozen one and fluctuate greatly. One possible reason is that, unlike the continuous picture patch in the visual transformer, the token in the textual modality is discrete single words, so it is hard to use a gradient-based method to fine-tune the soft prompts like a visual modality. Thus, in this paper, we choose a manually designed fixed prompt to obtain better performance.

METHOD	MTSE	MCCQ	MWTWT	MRUC	MTWQ
Frozen	**60.84**	**67.67**	**77.59**	**75.76**	**67.59**
Tuned	53.85	58.41	62.58	59.64	52.26

Table 9: Experimental results of fixed prompt and tuned soft prompt of textual modality. The reported results are the Macro F1-score across all targets in a dataset on in-target multi-modal stance detection.

METHOD	Textual Prompts	MTSE
BERT+ViT	-	53.29
①	Trump	58.54
②	Donald Trump	59.53
③	stance on Donald Trump	60.24
④	What is the stance on Donald Trump?	58.85
⑤ (Ours)	The stance on Donald Trump is:	**60.84**

Table 10: Experimental results of using different textual prompts in MTSE dataset on target "Donald Trump".

Hand-design textual prompts To analyze the impact of the type of the hand-design textual prompts, we design several types of textual prompts and report the experiments in Table 10. Take the MTSE dataset as an example, it can be seen that the experimental results of all types of prompts are superior to those without prompts (BERT+ViT). This demonstrates the significance of the targeted prompts in this task. Further, we can also see that there are considerable differences in performance between different types of prompts. Specifically, when using both target and stance to design prompts (③, ④, and ⑤), the model performs significantly better. Therefore, in our method, we choose the type of ⑤ to devise the textual prompts for multi-modal stance detection.

C.2 Analysis of Visual Prompt Tuning

Depth of prompt tuning Following (Jia et al., 2022), we conduct experiments with different depths of visual prompt tuning to analyze the impact of the depth of prompt tuning in our model. Here, shallow prompt tuning refers to only fine-tuning the prompt tokens in the Embedding layer of ViT, while deep prompt tuning refers to fine-tuning the prompt tokens in the Embedding layer of ViT and each layer in the transformer. The ex-

METHOD	MTSE	MCCQ	MWTWT	MRUC	MTWQ
Shallow (Ours)	60.84	**67.67**	**77.59**	**75.76**	**67.59**
Deep	**63.48**	62.10	72.68	69.66	61.08

Table 11: Comparison results of using the shallow prompt tuning and deep prompt tuning. Shallow prompt tuning refers to only fine-tuning the prompt tokens in the Embedding layer of ViT, while deep prompt tuning refers to fine-tuning the prompt tokens in the Embedding layer of ViT and each layer in the Transformer. The reported results are the Macro F1-score across all targets in a dataset on in-target multi-modal stance detection.

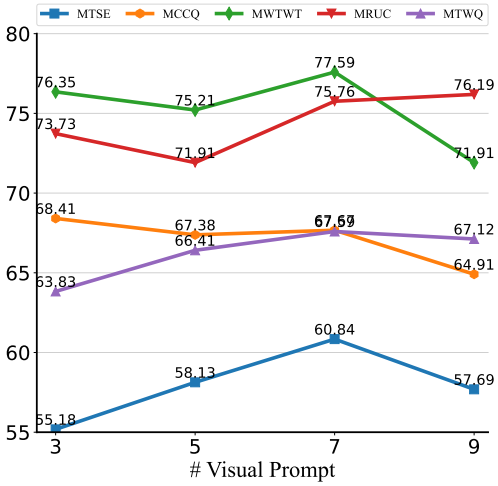


Figure 6: Experimental results of using different numbers of visual prompting tuning tokens. The reported results are the Macro F1-score across all targets in a dataset on in-target multi-modal stance detection.

perimental results are shown in Table 11. We can see that shallow prompt tuning has a better overall effect.

Number of visual prompt tokens To analyze the impact of the number of visual prompt tokens, we set the value range of tokens as {3, 5, 7, 9} for comparative experiments. The results are shown in Figure 6. Note that different values of tokens can have a certain impact on performance. When the value is 7, the overall performance of the model on all datasets is the best. Therefore, we set the number of tokens to 7 in our method.