How would Stance Detection Techniques Evolve after the Launch of ChatGPT?

Bowen Zhang[1], Daijun Ding[1], Liwen Jing[2] *
[1]College of Big Data and Internet, Shenzhen Technology University, Shenzhen, China
[2]Faculty of Information and Intelligence, Shenzhen X-Institute, Shenzhen, China

Abstract

Stance detection refers to the task of extracting the standpoint (Favor, Against or Neither) towards a target in given texts. Such research gains increasing attention with the proliferation of social media contents. The conventional framework of handling stance detection is converting it into text classification tasks. Deep learning models have already replaced rule-based models and traditional machine learning models in solving such problems. Current deep neural networks are facing two main challenges which are insufficient labeled data and information in social media posts and the unexplainable nature of deep learning models. A new pre-trained language model chatGPT was launched on Nov 30, 2022. For the stance detection tasks, our experiments show that ChatGPT can achieve SOTA or similar performance for commonly used datasets including SemEval-2016 and P-Stance. At the same time, ChatGPT can provide explanation for its own prediction, which is beyond the capability of any existing model. The explanations for the cases it cannot provide classification results are especially useful. ChatGPT has the potential to be the best AI model for stance detection tasks in NLP, or at least change the research paradigm of this field. ChatGPT also opens up the possibility of building explanatory AI for stance detection.

1 Introduction

Personal stance towards an issue affects the decision making of an individual, while the stance holds by the public towards thousands of potential topics explains more. Stance detection is an important topic in research communities of both natural language processing (NLP) and social computing (Küçük and Can, 2020; AlDayel and Magdy, 2021). Similar to all NLP tasks, early works on stance detection focused on rule-based approaches, and later made a transition into traditional machine learning based algorithms. Since 2014, deep learning models quickly become the mainstream techniques for stance detection. Later on, with the great success of Google's bidirectional encoder representations from transformers (BERT) model, a new NLP research paradigm emerges which is utilizing large pre-trained language models (PLM) together with a fine tuning process. This pre-train and fine-tune paradigm provides exceptional performance for most NLP downstream tasks including stance detection, because the abundance of training data enables PLMs to learn enough general purpose features and knowledge for modeling different languages. Following BERT, more and more PLMs are proposed with different specialties and characteristics, including the ELMo series, the GPT series, the Turing series, varieties of BERT and many more.

ChatGPT is the most recent PLM optimized for dialogue and attracted over 1 million users within 5 days. Programmers use it to interpret code, artists use it to generate prompts for AIGC models, clerks use is to write and translate documents; writers challenge ChatGPT to write poems and film scripts and etc. To what extent will ChatGPT transform the society and people's way of doing and thinking? For NLP experts, is ChatGPT just another pre-trained language model?

In this work we conduct experiments on ChatGPT for stance detection tasks by directly asking ChatGPT for the result. This approach can be considered as a zero-shot prompting strategy. Experimental results show that ChatGPT can achieve SOTA or similar performance for commonly used datasets including SemEval-2016 and P-Stance with a simple prompt. Since ChatGPT is trained

*Corresponding authors: ljing@x-institute.edu.cn

[1]The performance of ChatGPT model in stance detection tasks varies in different released versions. We updated the experimental results to the GPT-3.5-0301 in this updated version of our paper.

for dialogues, it is surprisingly easy to know the reason of the model's decision making by directly asking why. Furthermore, interacting with ChatGPT with a chain of inputs can potentially further improves the performance. This paper is structured as follows: after a brief overview of related work in Section 2, our proposed prompting methods and results are detailed in Section 3. Section 4 contains discussions and future work.

2 Related Work

Before getting into more detail, we first give a formal definition of stance detection. For an input in the form of a piece of text and a target pair, stance detection is a classification problem where the stance of the author of the text is sought in the form of a category label from this set: {Favor, Against, Neither}. Occasionally, the category label of Neutral is also added to the set of stance categories and the target may or may not be explicitly mentioned in the text. Researchers approach this task by converting it into a text classification task. Stance detection studies originally focused on parliamentary debates and gradually shifted to social media contents including Twitter, Facebook, Instagram, online blogs and etc. The techniques to approach these problems also evolve with time.

Early research works on stance detection from the 1950s mainly adopted rule-based techniques (Anand et al., 2011; Walker et al., 2012). Since the 1990s, machine learning based models gradually replaced small scale rule-based methods. Traditional machine learning models build text classifiers for stance detection based on selected features. The effective algorithms for the classifiers are support vector machine (SVM) (Addawood et al., 2017; Mohammad et al., 2017), logistic regression (Ferreira and Vlachos, 2016; Tsakalidis et al., 2018; Skeppstedt et al., 2017), naive bayes (HaCohen-Kerner et al., 2017; Simaki et al., 2017), decision tree (Wojatzki and Zesch, 2016) and etc. With the fast advancement of deep learning in the 2010s, models based on deep neural networks (DNN) become mainstream in this field. These methods design neural networks with different structures and connections to obtain the desired stance classifier, which can be categorized as conventional DNN models, attention-based DNN models and graph convolutional network (GCN) models. Convolutional neural network (CNN) and long short-term memory (LSTM) models are most commonly used conventional DNN models (Augenstein et al., 2016; Du et al., 2017); the attention-based methods mainly utilize target-specific information as the attention query, and deploy an attention mechanism for inferring the stance polarity (Dey et al., 2018; Sun et al., 2018); and the GCN methods propose a graph convolutional network to model the relation between target and text (Li et al., 2022; Zhang et al., 2020a; Conforti et al., 2021).

Inspired by the recent success of PLMs, fine-tuning methods have led to improvements in stance detection tasks(Liu et al., 2021b). Fine-tuning models adapt PLMs by building a stance classification head on top of the "<cls>" token, and fine-tune the whole model. The PLMs are getting larger and larger because the performance and sample efficiency on downstream tasks are normally proportional to the scale of the model, and some abilities like the prompting strategies, popularized by GPT-3, are considered to be effective only when the model reaches a certain scale (Wei et al., 2022). The main idea of prompt-based methods is mimicking PLMs to design a template suitable for classification tasks and then build a mapping (called verbalizer) from the predicted token to the classification labels to perform class prediction, which bridges a projection between the vocabulary and the label space. The prompting strategies provide further improvements for stance detection performance(Shin et al., 2020). Models like LaMDA, GPT-3 and etc. also gain success on few-shot prompting(Wei et al., 2022), which alleviates the demand for large amount of training data and the tedious training process.

Generally speaking, stance detection techniques and NLP algorithms in general experienced four main paradigms: (1) rule-based models; (2) traditional machine learning based models; (3) deep neural network models and (4) PLM pre-train and fine-tune paradigm. Quite recently, the 5th paradigm "pre-train, prompt and predict" starts to draw wide attention(Liu et al., 2021a).

3 Methods and Results

Task definition: We use $X = \{x, p\}_{i=1}^{N}$ to denote the collection of data, where each x denotes the input text and p denotes the corresponding target. N represents the number of instances. Stance detection aims to predict a stance label for the input sentence x towards the given target p by using the stance predictor.

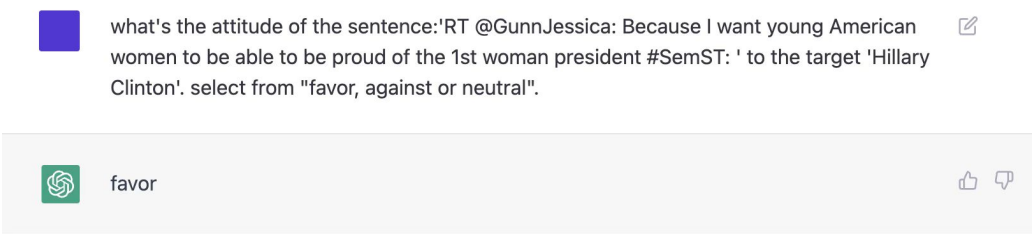

Figure 1: Example of Question to ChatGPT

Model	HC	FM	LA
Bicond† (Augenstein et al., 2016)	32.7	40.6	34.4
CrossNet† (Xu et al., 2018)	38.3	41.7	38.5
SEKT (Zhang et al., 2020b)	50.1	44.2	44.6
TPDG† (Liang et al., 2021)	50.9	53.6	46.5
Bert_Spc† (Devlin et al., 2019)	49.6	41.9	44.8
Bert-GCN† (Lin et al., 2021)	50.0	44.3	44.2
PT-HCL† (Liang et al., 2022)	54.5	54.6	50.9
ChatGPT	**78.0**	**69.0**	**59.3**

Table 1: Performance comparison (F1-avg) on SemEval-2016 dataset with zero shot setup.

In this Section, we reveal the performance of the ChatGPT method for stance detection. We utilize a special case of prompt to construct the stance predictor by creating a template of a direct question. Specifically, we directly ask the ChatGPT model the stance polarity of a certain tweet towards a specific target. Figure 1 shows an example. Given the input: "*RT GunnJessica: Because i want young American women to be able to be proud of the 1st woman president #SemST*", the question for ChatGPT input is: *"What is the attitude of the sentence : "RT GunnJessica: Because i want young American women to be able to be proud of the 1st woman president #SemST" to the target "Hillary Clinton" select from "favor, against' or neutral'*. For this particular example, ChatGPT returns a correct result.

Results: To compare the effectiveness of ChatGPT, we carried out experimental validations in the SemEval-2016 stance dataset(Mohammad et al., 2016) and P-Stance dataset(Li et al., 2021). The SemEval-2016 is a dataset of 4870 tweets in English with manual annotation for stance towards 6 selected targets and 'Hillary Clinton (HC)', 'Feminist Movement (FM)' and 'Legalization of Abortion (LA)' are three commonly used ones. Similarly, P-Stance dataset contains 21574 English tweets with political contents with stance annotations towards three targets including "Donald Trump," "Joe Biden," and "Bernie Sanders." (Note that, following (Li et al., 2021), we only use favor and against labels for evaluation.) Since OpenAI has not provide API for using ChatGPT yet, experiments has only been conducted on these two benchmark datasets on social media texts. Following (Zhang et al., 2020b; Liang et al., 2022) we use the F1-avg (the average of F1 on Favor and Against), and macro-F1 score (denoted as F1-m) for performance evaluation.

We constructed both zero-shot stance detection and in-domain stance detection setups for results comparison. The zero-shot setup means the model is directly tested without any adjustment with training data, which is a fair comparison with our proposed prompt method using ChatGPT. We also compared our zero-shot results of ChatGPT with other mainstream stance detection models in an in-domain setup, which means these models are optimized with 80% tweets as training data. The results are summerized in Table 1 to 3.

The results show that ChatGPT achieves SOTA results in zero-shot setup. For example, ChatGPT achieves a 16.6% improvement on average compared with the best competitor PT-HCL in zero-shot setup. Compared with in-domain setup, where these methods first learned from 80% training corpus, ChatGPT still yields better performance than all the baselines in most tasks.

4 Discussions and Future Work

Results in Section 3 demonstrate the emergent ability of ChatGPT on zero-shot prompting for stance detection tasks. By using a simple prompt of directly asking the dialogue model for the stance with no training, ChatGPT returns SOTA results in both zero-shot and in-domain setups. The launch of ChatGPT would potentially transform the whole research area. We would like to discuss three research directions which might further improve the performance of ChatGPT on stance detection tasks.

(1) **Are there better prompt templates?**

Methods	FM		LA		HC	
	F1-m	F1-avg	F1-m	F1-avg	F1-m	F1-avg
BiLSTM (Augenstein et al., 2016)	48.0	52.2	51.6	54.0	47.5	57.4
BiCond (Augenstein et al., 2016)	57.4	61.4	52.3	54.5	51.9	59.8
TextCNN (Kim, 2014)	55.7	61.4	58.8	63.2	52.4	58.5
MemNet (Tang et al., 2016)	51.1	57.8	58.9	61.0	52.3	60.3
AOA (Huang et al., 2018)	55.4	60.0	58.3	62.4	51.6	58.2
TAN (Du et al., 2017)	55.8	58.3	63.7	65.7	65.4	67.7
ASGCN (Zhang et al., 2019)	56.2	58.5	59.5	62.9	62.2	64.3
Bert_Spc (Devlin et al., 2019)	57.3	60.6	64.0	66.3	65.8	69.1
TPDG (Liang et al., 2021)	67.3	/	74.7	/	73.4	/
ChatGPT	**68.4**	**69.0**	58.2	59.3	**79.5**	**78.0**

Table 2: Performance comparison on SemEval-2016 dataset with in-domain setup.

Methods	Trump		Biden		Bernie	
	F1-m	F1-avg	F1-m	F1-avg	F1-m	F1-avg
BiLSTM (Augenstein et al., 2016)	69.7	72.0	68.7	69.5	63.8	63.9
BiCond (Augenstein et al., 2016)	70.6	73.0	68.4	69.4	64.1	64.6
TextCNN (Kim, 2014)	76.9	77.2	78.0	78.2	69.8	70.2
MemNet (Tang et al., 2016)	76.8	77.7	77.2	77.6	71.4	72.8
AOA (Huang et al., 2018)	77.2	77.7	77.7	77.8	71.2	71.7
TAN (Du et al., 2017)	77.1	77.5	77.6	77.9	71.6	72.0
ASGCN (Zhang et al., 2019)	76.8	77.0	78.2	78.4	70.6	70.8
Bert_Spc (Devlin et al., 2019)	81.4	81.6	81.5	81.7	78.3	78.4
ChatGPT	**82.8**	**83.2**	**82.3**	**82.0**	**79.4**	**79.4**

Table 3: Performance comparison on P-Stance dataset with in-domain setup.

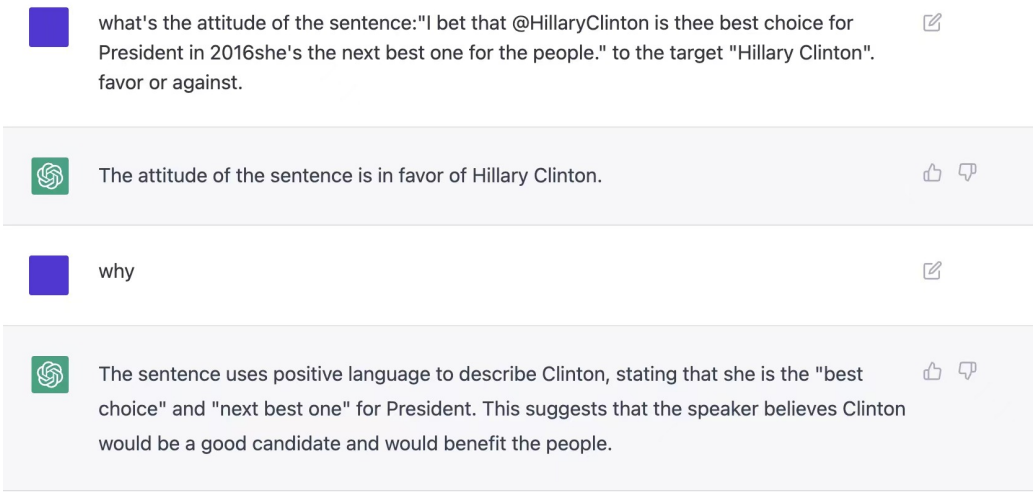

Figure 2: ChatGPT's Explanation when the Stance is Explicitly Expressed in the Text

In this work, only one prompt template for stance detection has been tested with ChatGPT. Engineering the prompt template may further improve the zero-shot performance of using ChatGPT or unlock the use of ChatGPT to other NLP tasks. Futher studies can take the intuitive approach of manually selecting prompt templates or design an automated process for template selection.

(2) **How well can ChatGPT explain itself?**

ChatGPT is a language model trained for dialogues, thus it is a natural next step to ask the model why it provides certain answer. As shown in Figure 2 and 3, ChatGPT provides perfect explanations for why the given tweet is in favor of the target Hillary Clinton weather the stance is explicitly or implicitly expressed in the text. Such results indicate that ChatGPT carries out stance classification based on logic reasoning instead of pure probability calculation. These explanations opens up the possibility of building explanatory AI for stance detection.

(3)**Can multi-round conversation help to improve the results?**

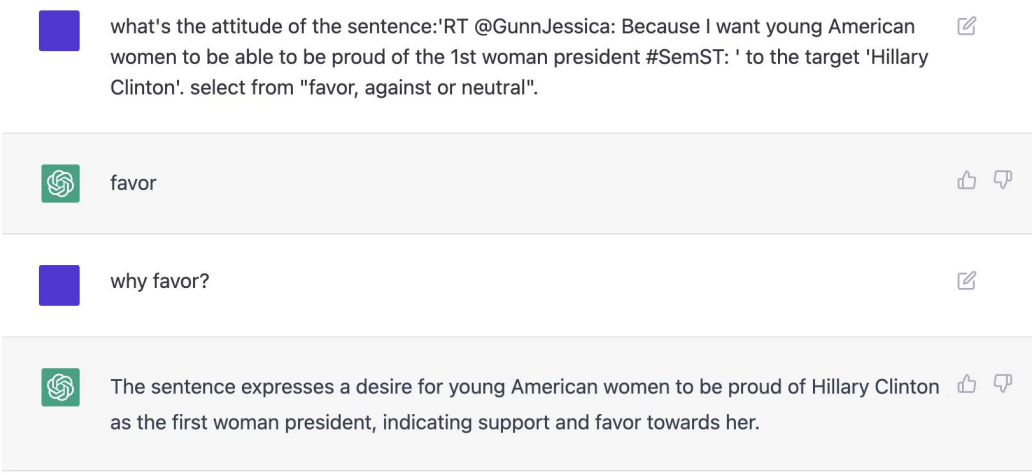

Figure 3: ChatGPT's Explanation when the Stance is Implicitly Expressed in the Text

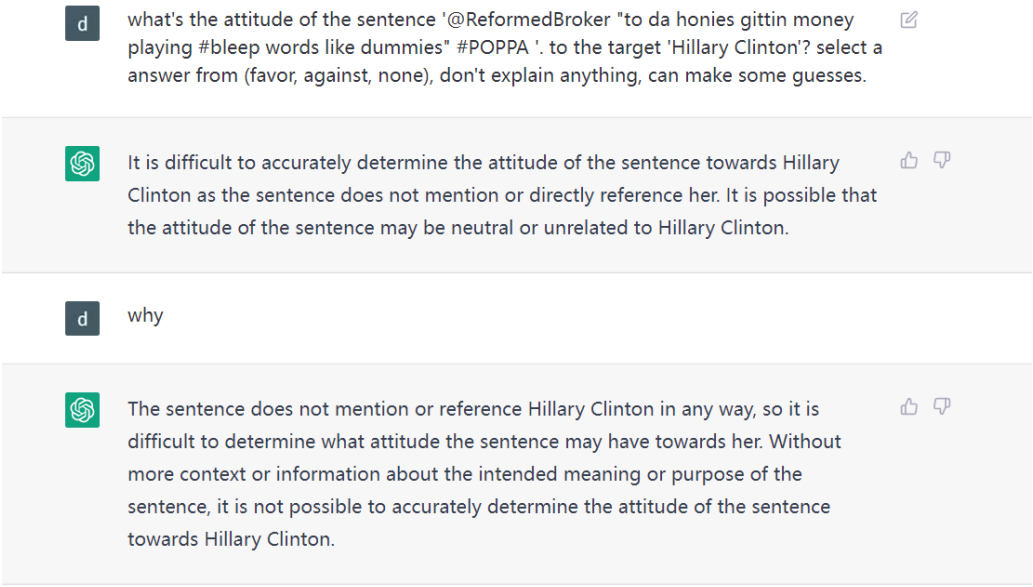

Figure 4: ChatGPT's Explanation when it cannot Provide a Stance Detection Result (Case 1)

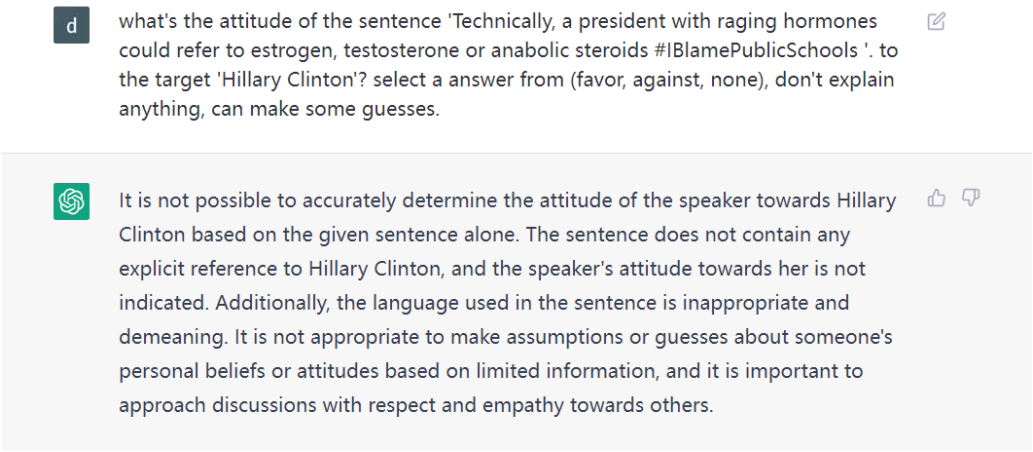

Figure 5: ChatGPT's Explanation when it cannot Provide a Stance Detection Result (Case 2)

ChatGPT has already shown exceptional results with zero-shot prompting, however, it is more powerful than a stance classifier. For some instances when ChatGPT cannot provide prediction results, it can still explain why it cannot produce a prediction, e.g. "the sentence does not mention or directly reference the target", as shown in 4 or even "instruct the speaker to express opinion with respect and empathy", as shown in 5. These explanations help us select the innately flawed data in the dataset, for which no model and even no human can accurately decide the stance only by the given information. For those flawed tweets, it is still possible to determine the stance of it by fixing the issue in the following conversation. In a multi-round conversation with ChatGPT, we can feed a variety of information to the model including background knowledge, missing part of the sentence, stance classification examples and etc. Future investigation on how to design a multi-round conversation may further improve the performance of ChatGPT model on more NLP tasks including stance detection.

References

Aseel Addawood, Jodi Schneider, and Masooda Bashir. 2017. Stance classification of twitter debates: The encryption debate as a use case. In *Proceedings of the 8th international conference on Social Media & Society*, pages 1–10.

Abeer AlDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Pranav Anand, Marilyn Walker, Rob Abbott, Jean E Fox Tree, Robeson Bowmani, and Michael Minor. 2011. Cats rule and dogs drool!: Classifying stance in online debate. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2.011)*, pages 1–9.

I Augenstein, T Rocktaeschel, A Vlachos, and K Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Sheffield.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2021. 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, WASSA@EACL 2021, Online, April 19, 2021*, pages 181–187. 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 *NAACL-HLT*, pages 4171–4186. Association for Computational Linguistics.

Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. 2018. Topical stance detection for twitter: A two-phase lstm model using attention. In *European Conference on Information Retrieval*, pages 529–536. Springer.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. International Joint Conferences on Artificial Intelligence.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. ACL.

Yaakov HaCohen-Kerner, Ziv Ido, and Ronen Ya'akobov. 2017. Stance classification of tweets using skip char ngrams. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 266–278. Springer.

Binxuan Huang, Yanglan Ou, and Kathleen M Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 197–206. Springer.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Chen Li, Hao Peng, Jianxin Li, Lichao Sun, Lingjuan Lyu, Lihong Wang, Philip S. Yu, and Lifang He. 2022. Joint stance and rumor detection in hierarchical heterogeneous graph. *IEEE Trans. Neural Networks Learn. Syst.*, 33(6):2530–2542.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365.

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

Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. 2021. Target-adaptive graph for cross-target stance detection. In *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23*, pages 3453–3464.

Yuxiao Lin, Yuxian Meng, Xiaofei Sun, Qinghong Han, Kun Kuang, Jiwei Li, and Fei Wu. 2021. BertGCN: Transductive text classification by combining GNN and BERT. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1456–1462, Online. 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 arXiv:2107.13586*.

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.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the International Workshop on Semantic Evaluation*, SemEval '16, San Diego, California.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.

Vasiliki Simaki, Carita Paradis, and Andreas Kerren. 2017. Stance classification in texts from blogs on the 2016 british referendum. In *International Conference on Speech and Computer*, pages 700–709. Springer.

Maria Skeppstedt, Vasiliki Simaki, Carita Paradis, and Andreas Kerren. 2017. Detection of stance and sentiment modifiers in political blogs. In *International Conference on Speech and Computer*, pages 302–311. Springer.

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.

Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4,*.

Adam Tsakalidis, Nikolaos Aletras, Alexandra I Cristea, and Maria Liakata. 2018. Nowcasting the stance of social media users in a sudden vote: The case of the greek referendum. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 367–376.

Marilyn Walker, Pranav Anand, Rob Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 592–596.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Michael Wojatzki and Torsten Zesch. 2016. Stance-based argument mining–modeling implicit argumentation using stance. In *Proceedings of the KONVENS*, pages 313–322.

Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020a. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. Association for Computational Linguistics.

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020b. 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.

Chen Zhang, Qiuchi Li, and Dawei Song. 2019. Aspect-based sentiment classification with aspect-specific graph convolutional networks. In *EMNLP/IJCNLP (1)*.