

# Enhancing Zero-shot and Few-shot Stance Detection with Commonsense Knowledge Graph

Rui Liu<sup>1,2</sup>, Zheng Lin<sup>1\*</sup>, Yutong Tan<sup>1,2</sup>, Weiping Wang<sup>1</sup>

<sup>1</sup>Institute of Information Engineering, Chinese Academy of Sciences

<sup>2</sup>School of Cyber Security, University of Chinese Academy of Sciences

{liurui1995, linzheng, tanyutong, wangweiping}@iie.ac.cn

## Abstract

In this paper, we consider a realistic scenario on stance detection with more application potential, i.e., zero-shot and few-shot stance detection, which identifies stances for a wide range of topics with no or very few training examples. Conventional data-driven approaches are not applicable to the above zero-shot and few-shot scenarios. For human beings, commonsense knowledge is a crucial element of understanding and reasoning. In the absence of annotated data and cryptic expression of users' stance, we believe that introducing commonsense relational knowledge as support for reasoning can further improve the generalization and reasoning ability of the model in the zero-shot and few-shot scenarios. Specifically, we introduce a commonsense knowledge enhanced model to exploit both the structural-level and semantic-level information of the relational knowledge. Extensive experiments demonstrate that our model outperforms the state-of-the-art methods on zero-shot and few-shot stance detection task.

## 1 Introduction

Stance detection aims to identify the text authors' attitudes or positions towards a specific topic as a category label from this set:  $\{Pro, Con, Neutral\}$  (Mohammad et al., 2016b, 2017). Conventionally, this task is designed to learn a target-specific classifier for prediction on the same topic. Afterward, cross-target stance detection comes out as a subclass of the initial generic stance detection, where the classifier is adapted from different but closely related topics (e.g., training classifier on "Hillary Clinton" and predicting on "Donald Trump") (Augenstein et al., 2016a). However, both target-specific and cross-target stance detection models (Du et al., 2017; Wei and Mao, 2019;

\*Zheng Lin is the corresponding author.

**Topic:** Stability **Stance:** Pro

**Text:** Tenure does not mean a teacher cannot lose their job. It requires due process before termination. Before tenure is achieved, a teacher can be fired without due process. In the Atlanta School District administrators, fearing that low test scores would cost them their jobs, instructed teachers to change student test responses. Without tenure and due process, teachers risked being fired if they didn't follow instructions.

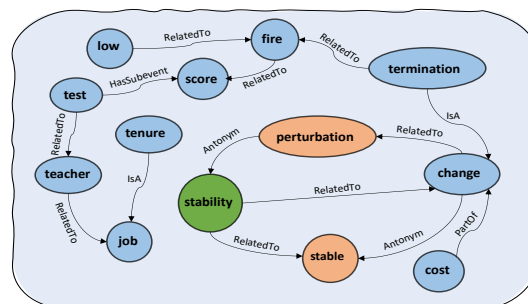


Figure 1: An example where the topic isn't contained in the text. Entity mentions in the text and the topic are highlighted. We omit the reverse edges in the relation graph for clarity.

Augenstein et al., 2016a; Zhang et al., 2020) require a large number of training examples with manual annotation, and annotating data for thousands of new topics is time-consuming and expensive.

In this paper, we focus on zero-shot and few-shot stance detection (Allaway and McKeown, 2020), a task to classify stances for a large number of topics with no or very few training examples. A key challenge for zero-shot and few-shot stance detection is the generalization ability of the models. However, most of the previous approaches (Xu et al., 2018; Augenstein et al., 2016b; Wei and Mao, 2019; Wei et al., 2019) for stance detection have relied on only the training data, which fails to achieve satisfactory results in zero-shot and few-shot scenarios. Another prominent challenge is the implicit expression of the users' stance, where the topic does not always appear in the document, resulting

in the difficulty of directly establishing a connection between the topic and the document. Take Figure 1 as an example, the topic “Stability” is not mentioned in the document, where the relational knowledge can supplement the lack of explicit inferential evidence such as (*stability*, *Antonym*, *perturbation*) and (*perturbation*, *RelatedTo*, *change*) etc. Despite attempting to introduce the external word-level semantic and emotion knowledge (Cambria et al., 2018) about each word of the document, Zhang et al. (2020) neglect the global relationship between the topic and the document.

To further tackle the above challenges, we propose to bring in commonsense knowledge from external structural knowledge base ConceptNet (Speer et al., 2017). We believe that the relational knowledge graph extracted from ConceptNet can promote the transmission of relational information between the document and the topic as well as the inference of corresponding stances, which can further reduce the dependency on annotated data. Specifically, we introduce a commonsense knowledge enhanced module based on Graph Convolution Networks (Kipf and Welling, 2017; Velickovic et al., 2018; Vashishth et al., 2020) to exploit both the structural-level and semantic-level information of the relation subgraph, which can further strengthen the generalization and reasoning capacities of the model. Extensive experiments show that our method outperforms the state-of-the-art models on the benchmark dataset for zero-shot and few-shot stance detection.

## 2 Methodology

### 2.1 Problem Formulation

Formally,  $D = \{(x_i, t_i, y_i)\}_{i=1}^N$  denotes the zero-shot stance detection dataset which contains  $N$  examples, where  $x_i$  is a document,  $t_i$  is the corresponding topic, and  $y_i$  is the stance label. The goal of the task is to obtain a stance label  $\tilde{y}$  given  $x_i$  and  $t_i$ . To bridge the document and the topic, we introduce an commonsense knowledge subgraph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  extracted from the external KG, where  $\mathcal{V}$  is the subset of the concepts and  $\mathcal{E}$  denotes the relations between concepts.

### 2.2 BERT Encoding

We employ the pre-trained language model BERT (Devlin et al., 2019) to encode the document  $x$  and topic  $t$ . Specifically, we concatenate  $x$  and  $t$  into one input sequence in the following

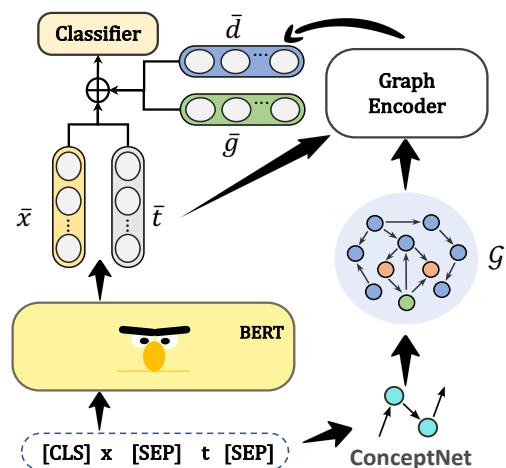


Figure 2: The architecture of our model.

format:  $[\text{CLS}] x [\text{SEP}] t [\text{SEP}]$ . Then, the input sequence is fed into BERT to obtain the contextual representations  $\mathbf{X} = \{x^1, \dots, x^m\}$  for the document and  $\mathbf{T} = \{t^1, \dots, t^n\}$  for the topic, where  $m$  and  $n$  is the length of the document and the topic respectively. Finally, we can get the average representations  $\hat{x}$  and  $\hat{t}$  of the document and the topic, respectively.

### 2.3 Knowledge Graph Encoding with CompGCN

Before introducing our graph encoder, let’s first describe the process of constructing the relational subgraph from the external knowledge graph. We adopt ConceptNet as our knowledge graph base  $\mathcal{G}$ . ConceptNet consists of millions of relation triples, which contains 34 relations in total. Each triple is represented as  $R = (u, r, v)$ , where  $u$  is the head concept,  $r$  is the relation, and  $v$  is the tail concept. We match phrases in **documents** and **topics** to sets of mentioned concepts ( $C_d$  and  $C_t$  respectively) from the ConceptNet. To extract the relational subgraph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  from  $\mathcal{G}$ , we find the two-hop directed paths from concepts in  $C_d$  to concepts in  $C_t$ . All concepts on the paths form the concepts set  $\mathcal{V}$  and  $\mathcal{E}$  is composed of all edges between concepts within  $\mathcal{V}$ . Moreover, we add reverse relations edge between any concept pair to improve the information flow.

Most of the existing research on GCNs mainly focuses on non-relational graphs. Thus, to incorporate commonsense relational knowledge, we utilize CompGCN (Vashishth et al., 2020), a variant of Graph Convolution Networks (GCNs), which jointly embeds both nodes and relations of the subgraph  $\mathcal{G}$ . The graph encoder consists of  $L$ -stacked

CompGCN layers. The features of nodes and relations are all initialized with **TransE** (Bordes et al., 2013) embeddings. We update node representations by aggregating the information from their neighbors and their relational edges. Formally, the update equation of nodes is defined as:

$$\mathbf{h}_v^{l+1} = f\left(\frac{1}{|\mathcal{N}(v)|} \sum_{(u,r) \in \mathcal{N}(v)} \mathbf{W}_u^l \phi(\mathbf{h}_u^l, \mathbf{h}_r^l)\right), \quad (1)$$

where  $f$  is an activation function,  $\mathcal{N}(v)$  is the neighbors of the node  $v$ ;  $\mathbf{h}_u$ ,  $\mathbf{h}_r$ , and  $\mathbf{h}_v$  are the representations of node  $u$ , node  $v$  and relation  $r$ .

Here,  $\phi$  is an entity-relation composition operation based on the translational theory (Bordes et al., 2013) in the form of subtraction:

$$\phi(\mathbf{h}_u, \mathbf{h}_r) = \mathbf{h}_u - \mathbf{h}_r. \quad (2)$$

The relation embeddings are transformed as follows:  $\mathbf{h}_r^{l+1} = \mathbf{W}_r^l \mathbf{h}_r^l$ . After that, we obtain the node representations  $\mathbf{H}_d$  and  $\mathbf{H}_t$  of  $C_d$  and  $C_t$ , respectively. To aggregate reasonable relational information, we compute the average relational representation  $\hat{\mathbf{d}}$  for  $C_d$  by performing scaled dot-product attention (Vaswani et al., 2017), with  $\hat{\mathbf{t}}$  as the key and  $\mathbf{H}_d$  as the query and value. Similarly, we get the average relational representation  $\hat{\mathbf{g}}$  for  $C_t$ .

## 2.4 Stance Classification

We concatenate the representations of plain texts (i.e.,  $\hat{\mathbf{x}}$  and  $\hat{\mathbf{t}}$ ) with the relational representations (i.e.,  $\hat{\mathbf{d}}$  and  $\hat{\mathbf{g}}$ ) to make full use of the textual information and the graph structural information. Afterward, the concatenated representations are fed into a two-layer multi-layer perception (MLP) with a softmax function to predict the stance label:

$$\hat{y} = \text{softmax}(\text{MLP}([\hat{\mathbf{x}}; \hat{\mathbf{t}}; \hat{\mathbf{d}}; \hat{\mathbf{g}}])), \quad (3)$$

where  $[\cdot]$  is vector concatenation operation.

Finally, the parameters of the network are trained using multi-class cross-entropy loss.

## 3 Experiments

### 3.1 Datasets and Evaluation Metrics

Different from the existing stance datasets (Mohammad et al., 2016a; Kobbe et al., 2020; Vamvas and Sennrich, 2020) including few topics (ranging from 6 to 194 topics), Allaway and McKeown (2020) present a new dataset for the zero-shot and

Statistics	Train	Dev	Test
# Examples	13477	2062	3006
# Documents	1845	682	786
# Zero-shot Topics	4003	383	600
# Few-shot Topics	638	114	159

Table 1: Detailed statistics for VAST.

few-shot stance detection, VARied Stance Topics (VAST), which consists of thousands of topics. The statistics of VAST are demonstrated in Table 1. Note that a document only belongs to one partition, which means that documents in the training set do not appear in the validation set or the test set, and vice versa. In addition, the zero-shot topics in the test set never appear in the training set, and the few-shot topics only contains few training data. Following the previous work (Allaway and McKeown, 2020), the macro average of F1-score is adopted as the evaluation metric.

### 3.2 Experimental Settings

We employ the base version of BERT as the backbone. The graph encoder has two layers of CompGCN. We train our model on 1 GPU (Nvidia RTX TITAN, 24G) using Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 4e-5 and a batch size of 64. All documents are kept the first 200 words and the topic is the first 5 words. The best checkpoints are selected according to the evaluation metrics on the validation set. We repeat our model three times using different random seeds and report the averaged results. Our code will be released on Github.

We compare our model with the several state-of-the-art baselines: **BiCond** (Augenstein et al., 2016b), **CrossNet** (Xu et al., 2018), **SEKT** (Zhang et al., 2020), **BERT-joint** (Allaway and McKeown, 2020) and **TGA-Net** (Allaway and McKeown, 2020). The first three models are based on BiLSTM for cross-target stance detection. When training the latter two BERT-based models, Allaway and McKeown (2020) fixed the parameters of the BERT module. Hence, we extend two models **BERT-joint-ft** and **TGA-Net-ft**, in which BERT has been fine-tuned during the training process. Besides, we compare our model with **BERT-GCN**, which applies the conventional GCN (Kipf and Welling, 2017) only considering the node information aggregation.

Model	F1 Zero-Shot				F1 Few-Shot				F1 All			
	pro	con	neu	all	pro	con	neu	all	pro	con	neu	all
BiCond	.459	.475	.349	.427	.454	.463	.259	.392	.457	.468	.306	.410
Cross-Net	.462	.434	.404	.434	.508	.505	.410	.474	.486	.471	.408	.455
SEKT	.504	.442	.308	.418	.510	.479	.215	.474	.507	.462	.263	.411
BERT-joint	.546	.584	.853	.660	.543	.597	.796	.646	.545	.591	.823	.653
TGA-Net	.554	.585	.858	.666	.589	.595	.805	.663	.573	.590	.831	.665
BERT-joint-ft	.579	.603	.875	.685	.595	.621	.831	.684	.588	.614	.853	.684
TGA-Net-ft	.568	.598	<b>.885</b>	.684	.628	.601	.834	.687	.599	.599	<b>.859</b>	.686
BERT-GCN	.583	.606	.869	.686	.628	<b>.634</b>	.830	.697	.606	<b>.620</b>	.849	.692
CKE-Net(Ours)	<b>.612</b>	<b>.612</b>	.880	<b>.702</b>	<b>.644</b>	.622	<b>.835</b>	<b>.701</b>	<b>.629</b>	.617	.857	<b>.701</b>

Table 2: Macro-averaged F1 on the test set. The suffix “ft” means the bert model is fine-tuned.

Model	Imp	mIT	mIS	Qte	Sarc
BERT-joint	.571	.590	.524	.634	.601
TGA-Net	.594	.605	.532	.661	.637
BERT-joint-ft	.617	.621	.547	.647	.668
TGA-Net-ft	.615	.625	.546	.664	.675
BERT-GCN	.619	.627	.547	.668	.673
CKE-Net	<b>.625</b>	<b>.634</b>	<b>.553</b>	<b>.695</b>	<b>.682</b>

Table 3: Accuracy on five challenging phenomena in the test set.

### 3.3 Results and Discussions

**Results of Different Scenarios** The overall results of our model and baselines are shown in Table 2. To evaluate the effectiveness of our method on different scenarios, we categorize the results into three subsets: *Zero-Shot*, *Few-shot* and *All*. Our model outperforms all baselines by a large margin, which can demonstrate the importance of incorporating the rich commonsense knowledge in the form of relational graphs. Additionally, we observe that all BERT-based baselines perform worse on *pro* examples than on *con* examples for zero-shot topics. A possible explanation might be that there are more negative words in the *con* examples, which is easier to identify in terms of semantics. Conversely, our model brings a significant improvement on average for both zero-shot and few-shot topics, which indicates that the relational information from the external knowledge base can boost the generalization and reasoning ability. Compared to BERT-GCN only modeling the node aggregation, our model takes full advantage of the relational information to contribute much to the overall performance.

Furthermore, all BERT-based models perform better than other baseline methods. In presents

that the pre-trained models possess more strong generalization capability because it learns from a large-scale unsupervised corpus. Besides, SEKT does not achieve effective improvement on VAST, probably because they only introduce the external semantic knowledge at the token level without explicitly considering the overall relationship between the topic and the document. And the token-level approach is difficult to transplant to BERT.

**Results of Different Phenomena** To further analyse the effectiveness of our model, we test it under five challenging phenomena in the VAST following (Allaway and McKeown, 2020): (1) **Imp**: examples with non-neutral labels, where the topic does not appear in the document, (2) **mIT**: documents having multiple examples with different topics, (3) **mIS**: documents having multiple examples with different and non-neutral labels, (4) **Qte**: documents with quotations, (5) **Sarc**: documents with sarcasm (Habernal et al., 2018). As shown in table 3, our model achieves the best performance on all difficult phenomena. In particular, the improvement on **Imp** demonstrates that introducing external relational knowledge can help the model better understand the relationship between the topic and the article. Besides, the external semantic-level information from the relational subgraph makes our model perform better on the special rhetorics (**Qte** and **Sarc**).

## 4 Conclusion

In this paper, we interpret the necessity of introducing commonsense knowledge for zero-shot and few-shot stance detection. We present a commonsense knowledge enhanced method, which facilitates the integration of the relational knowledge to

further strengthen the generalization and reasoning capacities of the stance detection model. Extensive experiments show that our proposed model achieved state-of-the-art results.

## Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 61976207, No. 61906187).

## 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.
- Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016a. [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.
- Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016b. [Stance detection with bidirectional conditional encoding](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 876–885. The Association for Computational Linguistics.
- Antoine Bordes, Nicolas Usunier, Alberto García-Durán, Jason Weston, and Oksana Yakhnenko. 2013. [Translating embeddings for modeling multi-relational data](#). In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pages 2787–2795.
- Erik Cambria, Soujanya Poria, Devamanyu Hazarika, and Kenneth Kwok. 2018. [Sentinet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings](#). In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1795–1802. AAAI Press.
- 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. [Stance classification with target-specific neural attention](#). In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3988–3994.
- Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. [The argument reasoning comprehension task: Identification and reconstruction of implicit warrants](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 1930–1940. Association for Computational Linguistics.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. OpenReview.net.
- Thomas N. Kipf and Max Welling. 2017. [Semi-supervised classification with graph convolutional networks](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Jonathan Kobbe, Ioana Hulpus, and Heiner Stuckenschmidt. 2020. [Unsupervised stance detection for arguments from consequences](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 50–60. Association for Computational Linguistics.
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiao-Dan Zhu, and Colin Cherry. 2016a. [Semeval-2016 task 6: Detecting stance in tweets](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 31–41. The Association for Computer Linguistics.
- Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. [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.
- Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. [Stance and sentiment in tweets](#). *ACM Trans. Internet Techn.*, 17(3):26:1–26:23.

- Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. [Conceptnet 5.5: An open multilingual graph of general knowledge](#). In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.
- Jannis Vamvas and Rico Sennrich. 2020. [X -stance: A multilingual multi-target dataset for stance detection](#). In *Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing, SwissText/KONVENS 2020, Zurich, Switzerland, June 23-25, 2020 [online only]*, volume 2624 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Shikhar Vashishth, Soumya Sanyal, Vikram Nitin, and Partha P. Talukdar. 2020. [Composition-based multi-relational graph convolutional networks](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- 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.
- Penghui Wei and Wenji Mao. 2019. [Modeling transferable topics for cross-target stance detection](#). In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 1173–1176. ACM.
- Penghui Wei, Wenji Mao, and Guandan Chen. 2019. [A topic-aware reinforced model for weakly supervised stance detection](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7249–7256. AAAI Press.
- Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. [Cross-target stance classification with self-attention networks](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 778–783. 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.