Target-phrase Zero-shot Stance Detection: Where Do We Stand?

Dawid Motyka and Maciej Piasecki[0000−0003−1503−0993]

Departament of Artificial Intelligence, Wrocław University of Science and Technology,
Wrocław, Poland
dawid.motyka@pwr.edu.pl

Abstract. Stance detection, i.e. recognition of utterances in favor, against or neutral in relation to some targets is important for text analysis. However, different approaches were tested on different datasets, often interpreted in different ways. We propose a unified overview of the state-of-the-art stance detection methods in which targets are expressed by short phrases. Special attention is given to zero-shot learning settings. An overview of the available multiple target datasets is presented that reveals several problems with the sets and their proper interpretation. Wherever possible, methods were re-run or even re-implemented to facilitate reliable comparison. A novel modification of a prompt-based approach to training encoder transformers for stance detection is proposed. It showed comparable results to those obtained with large language models, but at the cost of an order of magnitude fewer parameters. Our work tries to reliably show where do we stand in stance detection and where should we go, especially in terms of datasets and experimental settings.

Keywords: stance detection · zero-shot learning · prompt based learning for transformers

1 Introduction

People not only communicate some information or opinions, but also often express their stance against or in favor of the topic.

Stance is orthogonal to sentiment and emotions to a very large extent. While writing in favor is naturally more likely to express positive sentiment, this is not guaranteed. Mohhamad et al. [27] characterise *stance detection* as "the task of automatically determining from text whether the author of the text is in favor of, against, or neutral towards a proposition or target.". Such broad definition results in several variants, e.g. rumour stance [7]. First of all, different target types are considered, e.g. [12]: headline, claim, topic, person. Dealing with different target types, across different domains [12] is challenging. Some datasets use only two labels: *in-favor* and *against*, e.g. p-stance [17]. It may work if targets are clearly identifiable as proper names, but in the case of multiple targets two-label annotation blurs the difference between neutral utterances towards some target,

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_3

and non-related ones. Some authors consider inter-connected targets, like multi-target [30] (stance towards two targets, e.g. Clinton-Sanders), but this is the question of target definition.

Building data sets for supervised learning gets more laborious with the increasing number of targets, while the generalisation becomes harder. Thus, zero-shot-learning approach to stance detection is a good direction. [27] already discussed it, and [2] characterise it as stance detection for targets absent in the training data. [3] considers also different zero-shot perspectives, like language, genre, label set in relation to stance-detection. However, such perspective have not received limited attention so far.

We focus on stance detection tasks with targets expressed by short phrases. This case may be called *target-phrase stance detection* – contrary to tasks with targets represented by sentences or short texts. We focus on problems described with three labels: *in favor*, *against* and *neutral*, as the lack of the neutral category or implicit assumption that all non-labelled samples are neutral, significantly simplifies the problem. Target-phrase stance detection is represented by the most influential stance detection datasets, see Sec 2. Our main motivation was observation that different approaches in different papers were tested with slightly different experimental settings, also in combination with different interpretations of the annotation of datasets. All this causes problems with comparison and reproducibility. Focusing only on target-phrase stance detection may seem to be limitation, but target phrase stance detection is important for applications and enables to confine the overview within a limited size of the paper.

Due to the unclear picture of zero-shot stance detection, we aim at investigating where do we stand in the stance detection, at least in the target phrase subtype. What are the real results of approaches when compared in fairly comparable conditions?

The main novelty of our work is in clarifying the picture of zero-shot target phrase stance detection in relation to:

- available datasets for multiple target stance detection suitable for zero-shot learning together with their limitations,
- comparison of the SOTA approaches to multiple target stance detection performed on a carefully set-up common ground,
- re-running and in some cases even re-implementing several approaches[1],
- significant problems observed in an important benchmark, namely VAST [2], causing issues in comparison,
- supplementing the map of approaches with methods based on prompting transformers that were not tested in this task so far.

Our prompting-based methods express very good results in a zero-shot setting, better than the current SOTA. They are also in many cases better than a prompt-based learning stance detection with gpt-3.5-turbo or FLAN-UL2 models, while much more efficient, that makes them especially interesting.

[1] The source code: https://github.com/dawidm/zssd-iccs-2024

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_3

2 Stance Detection Datasets

For zero-shot target phrase stance detection, it is of primary importance that a dataset includes multiple and diversified targets. In the cases of small number of targets, e.g. six [27], zero-shot methods struggle to achieve good generalisation. Datasets with a small number of targets are not suitable for evaluation and development of such methods, while they dominate in stance detection. There are only two commonly used datasets, namely *SemEval 2016 Task 6* [27] and *VAried Stance Topics* [2] that are useful for target phrase zero-shot stance detection, but not free of some shortcomings, characterised below.

SemEval 2016 Task 6 [27] (henceforth Sem2016T6) is a commonly used stance detection dataset based on Twitter posts. It consists of 4 870 samples with six stance targets labelled by *favor*, *against* and *neither*. The last one is used to mark texts that are not related to a given target, but not necessarily including its mention. It may be used in a zero-shot configuration with the data for five targets used for training and the remaining one reserved for testing. During annotation authors labelled a few samples as *neutral*, i.e. referring to the target, but where stance cannot be deduced. However, finally all of them were assigned to the *neither* class.

VAried Stance Topics (VAST) [2], also popular in research, differs from Sem2016T6. Texts come from the *New York Times Room for Debate* forum and are on average longer and of a less casual character. VAST includes 18 515 samples divided into train, validation and test parts with 5 634 stance targets (called *topics* by the authors). Three labels are used: *pro*, *con* and *neutral*. A subset of the test samples includes mentions of targets not occurring in the training and validation parts, so this gives an opportunity for zero-shot experiment setting. It is worth to notice that only about 9% of samples can be considered to be *ingeniously neutral* stance: a given target occurs, but the sample has neutral stance in respect to it. Other ‘neutral’ samples have been created synthetically by assigning a random target to a sample in which it does not occur (so in a trivial way the stance cannot be other than neutral). We will call them *synthetic neutral* and others *true neutral*. More than half of the sample targets (counted with the exclusion of synthetic neutrals) are automatically extracted noun phrases, which is a distinctive feature of VAST. They tend to be relatively specific and, in some cases, even arguably suitable for a task e.g. *comment sections, healthier, a smart investment*. In this work we use the same label *neutral* for both *neither* samples from Sem2016T6 and *neutral* samples from VAST.

In addition to the two main datasets, several other were proposed, but all of them lack properties required for target phrase zero-shot stance detection.

P-Stance [17] is a large dataset of more than 20 000 samples but only 3 stance targets, which are also highly related (Trump, Biden, Sanders).

The encryption debate [1] consists of 3 000 samples but for only 1 target. Only tweets IDs are available, so there could be a problem with obtaining texts.

A Dataset for Multi-Target Stance Detection [30] incorporates 4 455 samples and 4 stance targets, which are highly related (e.g. Hillary Clinton, Bernie Sanders), but it enables cross-target or multi-target stance detection.

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_3

Stance Detection in COVID-19 Tweets [11] with 6 133 samples and 4 highly related stances targets (e.g. *stay at home orders, school closures*). It uses same annotation instructions as Sem2016T6. Only tweets IDs are available.

BASIL [9] includes only 300 news articles with sentence-level annotations, but only article-level annotations can be considered suitable for detecting the author's stance. The results on article-level annotations are very low even in non-zero-shot settings [24], probably due to the small number of samples.

3 Existing approaches

Existing zero-shot stance detection methods aim at improving language model generalisation to unseen stance targets. For a dataset like VAST, particularly rich in stance targets, only transformer models achieve high results, leaving behind other models based on recurrent (LSTM) neutral networks [2]. This is also true for Sem2016T6 in zero-shot setting, though the performance gap is smaller [4]. BERT-base [8] is most commonly used transformer encoder, followed by RoBERTa [23]. Recently, also the encoder-decoder transformer BART [16] was utilised [32, 18]. Various techniques described below are applied along with the mentioned models to improve the results.

Latent target representations from training datasets are incorporated to relate unseen targets to known ones to obtain better target-aware representations. Allaway et al. [2] used clustering of input representations, while [24] proposed a method with learned topic clusters. Another approach [20] used graph neural networks (GNNs) to link latent representations with the new samples.

External knowledge sources. Conditioning the stance detection model on an external knowledge source may be beneficial for generalisation to new targets, if the source contains relevant information. Also, updating knowledge should be easier with such methods compared to retraining the underlying language model. A general knowledge graph with graph neural networks to obtain graph embeddings was used in [22] and [26]. A different approach is to use knowledge in a form of plain text, which could be used directly as an input to a language model. Wikipedia's texts were used in [13] and [32] for creating stance target definitions. In [36] detailed information were obtained using keyword matching.

Contrastive learning. Contrastive learning was applied in stance detection for different purposes. Liang et al. [20] applied supervised contrastive learning in order to improve generalisation ability. In [35] contrastive approach were combined with word masking to capture target-invariant stance features. [14] proposed a solution able to leverage unlabelled data to acquire better target representations.

Pre-training on an auxiliary task. Liu et al. [24] investigated pretraining RoBERTa language model using a large collection of political texts to improve stance detection in this domain. [33] utilised similarities in textual entailment and stance detection to pre-train RoBERTa with textual entailment task.

Dataset augmentations. In zero-shot stance detection, data augmentation could be used to generate new pairs: target and text to improve generalisation. Such an approach with the help of a large language model (LLM), namely GPT3

[6] was proposed in [33]. A method with a smaller generative model used to extract keywords as potential new stance targets was also developed [18].
Using prompts for transformer encoder models has been shown to increase performance, especially in a few-shot settings for many tasks [28, 29, 10]. [24] apply the RoBERTa-base model and test it on VAST and Sem2016T6 datasets, but does not focus specifically on zero-shot performance.
Opportunities of generative models. Using generative models brings new possibilities in improving learning for stance detection, e.g. [32]. They proposed predicting not only the stance label but also the target and using unlikelihood to leverage samples with assigned opposite labels as an auxiliary tasks.
Large language models (LLM), exhibit an ability to solve various tasks in zero-shot settings [6]. Kocoń et al. [15] tested ChatGPT (a GPT-family model) on numerous tasks, including stance detection, showing results lower compared to fine-tuned transformer encoder models. [34] also provided results for 3 targets from Sem2016T6. Both approaches were conducted using ChatGPT web interface before official OpenAI API was available. More tests are needed to examine the performance and problems of LLMs in stance detection.

4 Experiments

4.1 Datasets and Metrics

We chose the Sem2016T6 and VAST datasets. For Sem2016T6 we use six configurations created by leaving out samples with a given target as a test set and splitting the rest of the samples into 85% and 15% for a train and validation set, respectively. For VAST, we use the default splits.

We apply metrics commonly used for the selected datasets. For Sem2016T6 the average of F_1 for *favor* and *against* classes (F_{1mfa}) [27]. It should be noted that *neutral* samples are included in these calculations. For VAST, the average of F_1 score of all three classes is used, calculated both for the whole test dataset F_{1m} and its zero-shot part F_{1mZS} [2]. We also report the average of F_{1m} for Sem2016T6 in some experiments, for comparison with [33]. All experiments were run 10 times and the average result is reported.

Most neutral samples in VAST are *synthetic* ones. We verified performance on *true neutral* (i.e. related to a target but with no clear stance): test set containing only *true neutral* samples left in the test set and train two models, one with also *synthetic neutrals* and the other without. The limitation of this approach is that there are only 114 and 45 *true neutral* samples in the training and test set respectively.

4.2 Baseline models

Establishing good, reliable baseline results for models of sizes comparable to BERT-base for both considered datasets on the basis of literature appeared to be challenging. Our results for **Sem2016T6** were not consistent with the

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_3

commonly used baseline from [4]. We re-implemented their setup with BERT-base to recreate the results. We also trained RoBERTa-base with our setup for comparison, as it was examined so far only for the non-zero-shot variant [5].

For the **VAST** dataset, due to its rich original data structure, it may be unclear which values to use. We distinguished and tested four possible variants using BERT-base and the best configuration also with RoBERTa-base:
Using unprocessed texts. Lemmatised and with stop words removed texts are stored in `text_s` along with unprocessed ones in `post`. We found it ambiguous which values are used in the works from literature, and we test both variants.
Using unprocessed targets. The similar situation is with stance targets, but there is no specific column where all unprocessed values are valid, and it requires additional effort to extract them. We test both target sets with our baselines.
Discarding type 3 (*list*) samples. It can be concluded from the annotation task that samples of a certain type of *list* samples often have the wrong stance label. Such samples can be removed, not affecting the validation and test sets.
Discarding ambiguous samples. there are some samples in the training set with different labels for the same text and stance target, that can be removed. The training set sizes after different modifications are shown in a Tab. 1.

4.3 Reproduced methods

For fair comparison we tried to reproduce results for all the methods using models of the size comparable to BERT-base running the published source codes. In the **JoinCL** [20] implementation we find out that for Sem2016T6 in zero-shot configuration, the test part is also used as a *validation* set during development (sic!), while it should be a subset of the training set, e.g. [4]. In our reproduction we did not introduce any changes to the method itself and used the provided parameters. We only added an evaluation metric for the full VAST dataset.

In **WS-BERT** [13] target definitions for VAST are automatically retrieved from Wikipedia, but the targets come from `new_topic` column and are not correct for *synthetic neutral* samples, i.e. they have not been changed to random ones in order to make the stance neutral. This makes the obtained definitions highly related to texts, which potentially adds positive bias to classification. For reproduction, we fixed this problem by correcting stance targets in `new_topic` column in VAST, but has not introduced any other changes, and used parameters provided by authors. For Sem2016T6 we used the WS-BERT-Single variant and originally selected articles from Wikipedia as definitions.

Wen et al. [32] (henceforth **VTCG**) used the same definitions as in WS-BERT, so the same concerns are valid for their method. We fixed `new_topic` column and run experiments with the original code and parameters. We also test the method without target definitions (VTCG-NW) and with BART-base without modifications (VTCG-BO).

BS-RGCN [26] was not tested on Sem2016T6 and needed modifications to work for its shorter texts – we let more words be used for knowledge graph embeddings. We run BS-RGCN with the original code and configuration proposed for VAST, and our modified version on Sem2016T6.

4.4 Prompt based methods

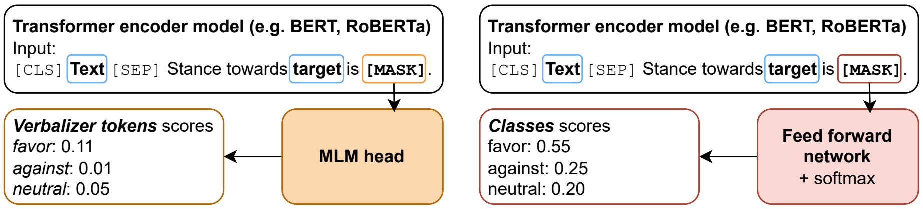
Transformer encoder model (e.g. BERT, RoBERTa)
Input:
[CLS] Text [SEP] Stance towards target is [MASK].
MLM head
Verbalizer tokens scores
favor: 0.11
against: 0.01
neutral: 0.05
Transformer encoder model (e.g. BERT, RoBERTa)
Input:
[CLS] Text [SEP] Stance towards target is [MASK].
Feed forward network + softmax
Classes scores
favor: 0.55
against: 0.25
neutral: 0.20

Fig. 1. Comparison of two approaches for using prompts with transformer encoder language models. Left: common approach proposed in [28]. Right: our modification.

For transformer encoder models, a masked language pretraining objective can be used to train the model for a downstream task, as shown by [28, 29, 10]. Compared to standard fine-tuning, this requires a task-specific prompt and verbalizer tokens (verbalizers), i.e. tokens from model's dictionary that would replace the `[MASK]` token following the prompt to indicate given class. Such an approach was not yet examined specifically for zero-shot stance detection, but was shown to be promising [24]. Several methods for selecting verbalizers were proposed in the aforementioned publications.

We use two prompt-based approaches, see Fig. 1. In our modification of [28] we replace the head used in masked language modelling (MLM) pretraining (returning scores for every token), by a feed forward network returning scores corresponding to 3 stance classes which eliminates the need to specify a verbalizer and doesn't restrict classification to predefined class tokens.

As a feed forward head we use 2 layers $\dim h \times \dim h$ and $\dim h \times |C|$ (where C – a set of classes, $\dim h$ – a model hidden state size) with dropout and GELU activation layer between.

We propose slight modifications to the best prompt of [24] (*The stance towards [target] is [MASK].*, **P2**) making it written in the first person and using a word that is an indicator that it concludes the previous text: *Therefore my stance towards [target] is [MASK].* (**P1**).

We run prompt based method with prompts described above. We name the models as follows: `model-type-prompt_name`, where model is underlying encoder model, type is PV (with verbalizer) or PFF (our model with feed forward classification head) e.g. `BERT-PV-P1`.

We use gpt-3.5-turbo (Apr. 30, 2023) from OpenAI, based on GPT-3 architecture[2] and the FLAN-UL2 model – based on encoder-decoder transformer architecture, with 19.5B parameters, fine-tuned to various tasks, with zero-shot ability [31, 25] – as LLM baselines for in zero-shot setting. For each dataset, we propose one prompt based on its annotation task (**P3**, **P4**) and a designed

[2] https://platform.openai.com/docs/models/gpt-3-5

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_3

prompt (**P5**) aimed at improving results (we provide them in our source code repository). We use the online inference API for both models[3], [4]

Table 1. The numbers of samples in VAST train, in *+Discard amb.* – *list* type and ambiguous samples are excluded, see Sec. 4.2

	VAST train	**Discard *list* type**	**+Discard amb.**
Samples	13447	6922	6870
Unique texts	1845	1845	1844
***favor* samples**	5327	2104	2082
***against* samples**	5595	2385	2363
***neutral* samples**	2555	2104	2425
Stance targets	5014	1797	1794

Table 2. Sem2016T6 zero-shot setting with smaller models: * evaluation on the test set in training, † – originally not tested with Sem2016T6, ‡ – modified by us for Sem2016T6.

	Method	**Per-target results (F_{1fa})**						**Average**
		DT	HC	FM	LA	A	CC	(F_{1mfa})
Base (ours)	Random guessing	0.315	0.321	0.360	0.328	0.330	0.254	0.318
	BERT-base	**0.403**	0.549	**0.441**	**0.447**	**0.368**	**0.299**	**0.418**
	RoBERTa-base	0.279	**0.565**	0.327	0.401	0.268	0.248	0.348
Reported results	BERT [4]	0.401	0.496	0.419	0.448	0.552	0.373	0.448
	TOAD [4]	0.495	0.512	0.541	0.462	0.461	0.309	0.463
	JointCL* [20]	0.505	0.548	0.538	0.495	0.545	0.397	0.505
	TarTK [36]	0.508	0.551	0.538	0.487	0.562	0.395	0.507
	PT-HCL [19]	0.501	0.545	0.546	0.509	0.565	0.389	0.509
	FECL [35]	**0.516**	0.556	0.553	0.533	**0.573**	0.418	0.525
	MPCL [14]	0.512	**0.595**	**0.556**	**0.534**	0.567	**0.454**	**0.536**
Reproduced results	JointCL*	**0.453**	**0.551**	**0.451**	**0.470**	**0.464**	**0.280**	**0.445**
	WS-BERT† [13]	0.212	0.356	0.279	0.262	0.420	0.076	0.268
	BS-RGCN†‡ [26]	0.214	0.325	0.257	0.253	0.396	0.100	0.258
	VTCG [32]†	0.400	0.294	0.285	0.362	0.415	0.234	0.332
	VTCG-NW†	0.516	0.501	0.262	0.317	0.325	0.270	0.365
	VTCG-BO†	0.322	0.464	0.324	0.166	0.373	0.120	0.295
Prompt models	BERT-base-PV-P1	0.371	0.587	**0.468**	**0.488**	0.342	0.258	0.419
	BERT-base-PFF-P1	0.369	0.573	0.429	0.478	0.326	0.286	0.410
	RoBERTa-base-PV-P1	**0.599**	**0.710**	0.455	0.457	0.417	**0.375**	**0.502**
	RoBERTa-base-PFF-P1	0.537	0.654	0.426	0.474	**0.423**	0.303	0.470

5 Results

Tab. 2 & 4 show results for methods with models of size comparable to BERT-base, while Tab. 3 & 4 focus on larger models. In Tab. 4 we consider variations of VAST usually neglected in literature.

[3] https://platform.openai.com/docs/api-reference

[4] https://huggingface.co/inference-api

Table 3. Evaluation on Sem2016T6 zero-shot setting with BERT-large and larger.

Method	**Per-target results (F_{1mfa})**						**Average**	**Average**
	DT	HC	FM	LA	A	CC	(F_{1mfa})	(F_{1m})
BERT-large	**0.422**	0.514	0.404	0.416	**0.382**	**0.399**	**0.423**	**0.463**
RoBERTa-large	0.233	**0.573**	**0.482**	**0.431**	0.252	0.112	0.347	0.404
OpenStance [33]	-	-	-	-	**-**	-	-	0.637
ChatGPT [34]	-	0.780	0.690	0.593	-	-	-	-
gpt-3.5-turbo-P3	0.684	0.821	0.715	0.547	0.126	0.732	0.588	0.589
gpt-3.5-turbo-P5	0.661	0.821	0.724	0.692	0.539	**0.707**	0.697	**0.670**
FLAN-UL2-P3	**0.700**	**0.824**	**0.729**	0.682	0.687	0.543	0.694	**0.670**
FLAN-UL2-P5	0.630	0.748	0.706	**0.742**	**0.763**	0.692	**0.713**	0.634
RoBERTa-large-PV-P1	0.631	**0.788**	**0.679**	0.612	0.622	0.268	0.600	0.624
RoBERTa-large-PFF-P1	**0.641**	0.777	0.655	**0.623**	**0.716**	**0.365**	**0.634**	**0.653**
RoBERTa-large-PV-P2	0.624	0.779	0.646	0.612	0.687	0.196	0.591	0.622
RoBERTa-large-PFF-P2	0.634	0.761	0.656	0.598	0.606	0.308	0.594	0.629

5.1 Baseline results

We failed to reproduce the BERT baseline of [4] (Tab. 2) for Sem2016T6 zero-shot dataset. The average score is slightly lower, even if some targets scored higher. The most substantial difference is for *Atheism* (A) target. RoBERTa model scored even lower. It is worth to notice that both results are not much higher than random guessing. The *large* models are marginally better when fine-tuned in a standard way (Tab. 3). In Tab. 4, we found the differences between dataset configurations to be substantial, and the best are achieved with all text unprocessed and with discarding a large amount of training samples. The *base* size models achieved comparable results with both the full test set and the zero-shot part.

5.2 Reproduced results

Our reproduced results are generally lower for Sem2016T6 for all methods, with only two of the approaches better than random guessing. From the validation results we see that models actually learn stance detection, ($F_{1fa-val}$ was always higher than 0.65), but fail to generalise for a zero-shot target. On VAST (Tab. 4) our reproduced results are mostly lower than reported by the authors. Methods that use unprocessed texts tend to score higher, which is consistent with our analysis of the possible VAST configurations. Methods based on Wikipedia definitions fail to achieve results better than baselines when corrected definitions are provided. Just one tested method, VTCG, but without target definitions (VTCG-NW) achieved better results than the language model itself (BART-base, VTCG-BO).

5.3 Prompt models results

Prompting RoBERTa models shows significant improvement in comparison to fine-tuning on Sem2016T6. Our prompting RoBERTa-base achieved the highest

Table 4. Smaller models on VAST: *UP* – unprocessed texts, *UT* – unprocessed stance targets, *DA* – without ambiguous samples, *DA* – no *list* samples, see Sec. 4.2. * "?" no information or lacking source code. † only if explicitly stated in publication. ‡ corrected Wikipedia target definitions (Sec. 4.3), § [2].

	Method	**Dataset variant**				**Result**	**ZS Result**
		UP*	**UT†**	**DA†**	**DL†**	(F_{1m})	(F_{1mZS})
Base (ours)	BERT-base	-	-	-	-	0.695	0.694
	BERT-base	-	-	+	+	0.716	0.717
	BERT-base	+	-	-	-	0.716	0.719
	BERT-base	+	+	-	-	0.707	0.707
	BERT-base	+	-	+	-	0.717	0.720
	BERT-base	+	+	+	-	0.719	0.722
	BERT-base	+	-	+	+	0.733	0.736
	BERT-base	+	+	+	+	0.735	0.739
	RoBERTa-base	+	+	+	+	**0.757**	**0.768**
Reported results	TGA-NET§	-	-	-	-	0.665	0.666
	CKE-Net [22]	?	?	-	-	0.701	0.702
	JointCL [20]	-	-	-	-	-	0.723
	TarTK [36]	?	?	-	-	-	0.736
	WS-BERT [13]	-	-	-	-	0.745	0.753
	DTCL [21]	?	?	-	-	0.712	0.708
	BS-RGCN [26]	+	-	-	-	0.713	0.726
	PT-HCL [19]	?	?	-	-	-	0.716
	FECL [35]	?	?	-	-	-	0.725
	MPCL [14]	?	?	-	-	-	0.724
	POLITICS [24]	?	?	-	-	0.763	-
	VTCG [32]	+	-	-	-	**0.773**	**0.764**
Reproduced results	JointCL	-	-	-	-	0.701	0.700
	VTCG‡	+	-	-	-	0.730	0.739
	VTCG-NW	+	-	-	-	**0.731**	**0.747**
	VTCG-BO	+	-	-	-	0.723	0.742
	WS-BERT‡	-	-	-	-	0.677	0.685
	BS-RGCN	+	-	-	-	0.694	0.716
Prompt models	BERT-base-PV-P1	+	+	+	+	0.730	0.732
	BERT-base-PFF-P1	+	+	+	+	0.720	0.728
	RoBERTa-base-PV-P1	+	+	+	+	**0.764**	**0.776**
	RoBERTa-base-PFF-P1	+	+	+	+	0.762	0.770

average score for all targets. RoBERTa-large variants (Tab. 3) achieved markedly better results than the BERT and RoBERTa baselines. Our approach of prompting with a feed-forward classification head shows the best average results between all tested prompting variants and the best results for most targets. We observe only a small but statistically insignificant ($p > 0.05$) improvement with prompt models and VAST with *base* models, and practically no difference with *large* ones. There is a visible difference in favor of an approach with verbalizers with *base* models, but feed-forward classification gave better with *large* models.

We show in Tab. 6 that the model learned from the full VAST train part score much lower on *true neutral* samples compared to a model trained on a set with *synthetic neutrals* excluded, despite that there are only 114 training samples left and the training set is highly unbalanced. This points to a limitation of training sets with lacking *true neutrals*.

Table 5. Results for VAST with models larger or equal to BERT-large.

Method	**Result** (F_{1m})	**ZS result** (F_{1mZS})
BERT-large	0.750	0.759
RoBERTa-large	**0.811**	**0.833**
RoBERTa-large[33]	**0.780**	-
TTS [18]	-	**0.801**
gpt-3.5-turbo-P4	0.643	-
gpt-3.5-turbo-P5	**0.772**	-
FLAN-UL2-P4	0.652	-
FLAN-UL2-P5	0.707	-
RoBERTa-large-PV-P1	0.813	0.827
RoBERTa-large-PFF-P1	0.812	**0.832**
RoBERTa-large-PV-P2	**0.815**	0.827
RoBERTa-large-PFF-P2	0.811	0.830

Table 6. Results for VAST for *true neutral* samples with RoBERTa-base-PV-P1 trained with *synthetic neutrals* included or excluded.

	Precision	**Recall**	F_1
full training set	0.050	0.222	0.081
synthetic neutrals excluded	0.400	0.267	**0.320**

5.4 Large language models

There is significant variance in Sem2016T6 per target results of gpt-3.5-turbo (Tab. 3). Results for P3 are very low for *atheism* (A) and *legalization of abortion* (LA), but substantially rise with changing the prompt. We think that this may be due to the specific tunning of the model to not produce harmful content, which may interfere with the classification of controversial topics, but its influence depends on the prompt used. We observe a lower variation of results between targets and prompts used for FLAN-UL2. For VAST (Tab. 5), there is a visible advantage of prompt P5 for both gpt-3.5-turbo and FLAN-UL2. It leads to the conclusion that prompts based on a definition from an annotation task (P3, P4) may not be the best candidates for stance detection with LLMs. Comparing to our prompting encoders LLMs have a slight advantage on Sem2016T6, but their results for VAST are lower. It should be noted that comparing both approaches is problematic since LLMs have no opportunity to learn annotators bias for a given dataset, which could be significant for a relatively subjective task such as stance detection.

5.5 Comparison with the state of the art

Our results shed new light on the current SOTA for VAST. As seen in Tab. 4 results for BERT-base with a subset of the training dataset are higher than other BERT-base methods. Also, our results with RoBERTa-large are higher than those of any other model with >140M parameters, including the current SOTA [18].

Regarding Sem2016T6 in zero-shot configuration (2), we show that prompting RoBERTa-base comes close to the highest average result in [14]. One target, (*climate change is a real concern*), seems to be problematic, but it is less compatible with the used prompts due to its character of a claim. Among >140M models, we could only compare to [33] (the current SOTA) and our poposed prompting approach with RoBERTa-large (RoBERTa-PFF-P1) scored higher (Tab. 3).

6 Conclusions

We touched on many aspects in target-phrase zero shot stance detection, focusing on the two most relevant datasets.

In literature, transformer-based models are the most effective ones due to their ability to jointly encode both target and text. In addition to BERT-base, the most popular one, other approaches like RoBERTa or BART with comparable sizes, were shown to be equally or more effective. BERT should perform better, due to the next sentence prediction task [8], but instead RoBERTa, trained on single sentences, achieves generally better results, especially when applied in a prompt-based approach.

There were many attempts to improve performance of transformers-only solutions, but from both literature and especially our careful reproductions we see that the improvements are small or marginal. Integrating additional knowledge into a transformer model naturally facilitates zero-shot stance detection, because good knowledge about the stance target is often needed. Current knowledge-enhanced methods do not present significant improvements, but this may signal that more work is still needed. We showed that simply introducing Wikipedia target definitions into the model's input actually worsens results for both datasets. We think that this is more easily explained for the Sem2016T6 dataset, when learning to utilize additional knowledge from just five stance target definitions may be too much to expect. On the other hand, this is not the case for VAST. We hypothesize that simply using just target definitions and not knowledge about classified text may not be enough, and often the text may be crucial to disambiguating the target (limiting misleading definitions). Still, utilizing additional knowledge requires understanding it and linking it to the sample's text, which makes it a complex language-understanding task, especially for considered models. Knowledge graph embedding utilizing whole text is used in BS-RGCN, but also introduces additional model parameters (GNN) that were not pretrained on a broad corpus. We think that this may be the main reason why it fails to generalise for Sem2016T6 considering that the training set focuses on just five themes. Including knowledge embedding by concatenation with transformer output is a limiting factor in aggregating it with pertaining knowledge.

Our analysis highlighted several aspects of VAST that are important for proper interpretation of the previous works, and that must be carefully considered in all future works to make the results reliable and comparable. Besides that, we also showed that discarding certain samples from the training set, up to half of it, leads to better results! This proofs the importance of training data quality. It let us achieve SOTA results using just RoBERTa-large, but also the best results with BERT-base.

We proposed a modification of an approach based on prompting a transformer encoder model. As a result designing a prompt is simpler (no need for verbal-

DOI: 10.1007/978-3-031-63751-3_3

izer tokens), it works especially well with *large* variants of models and achieves the best average result for Sem2016T6 zero-shot (excluding LLMs). We showed that the gap between the results of transformers with prompts in comparison to standard fine-tuning is visible for Sem2016T6. However, we did not notice statistically significant improvements on VAST, that may signal that prompting is especially effective in the case of fewer stance targets. Anyway, transformers with prompts may be a new strong baseline. They improve results by better utilising MLM pretraining task.

LLMs appeared to be very effective in zero-shot stance detection, especially in Sem2016T6 zero-shot with small number of targets. However, our prompting approach presented only slightly worse results, but utilising an order of magnitude fewer parameters. Both tested LLMs could not beat RoBERTa-large in our experiments on VAST. However, we spent only limited time on tuning the prompts, while correct prompts may be crucial for a high performance. It is also interesting that prompts based on annotation tasks were not as good candidates in comparison to slightly more complex ones.

In all datasets, we observe shortage of ambiguous/neutral samples related to targets (*true neutrals*). During tests on VAST a model trained on *synthetic neutrals* practically does not recognise *true neutral* samples, that is a major problem. Considering VAST, higher results are usually obtained for the zero-shot setting. This may suggest that a more challenging dataset with zero-shot targets more distinct from the training ones is needed. A large set of diverse stance targets with high-quality annotation is crucial for the further development.

We experimented with performance of stance detection models for different types of *neutral* samples, but due to the small number of *true neutral* samples in VAST, our results are only estimation and starting point for future experiments.

Acknowledgements. The work was financed as part of the investment: "CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024-2026) funded by the Polish Ministry of Science (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), agreement number 2024/WK/01."

References

1. Addawood, A., Schneider, J., Bashir, M.: Stance classification of twitter debates. In: Proc. of the 8th Int. Conf. on Social Media. ACM Press (2017)
2. Allaway, E., McKeown, K.: Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In: Proc. of the 2020 EMNLP. pp. 8913–8931. ACL, Online (Nov 2020)
3. Allaway, E., McKeown, K.: Zero-shot stance detection: Paradigms and challenges. Frontiers in Artificial Intelligence **5** (2023)
4. Allaway, E., Srikanth, M., McKeown, K.: Adversarial learning for zero-shot stance detection on social media. In: Proc. of the 2021 NAACL: Human Language Technologies. pp. 4756–4767. ACL, Online (Jun 2021)

5. Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., Neves, L.: TweetEval: Unified benchmark and comparative evaluation for tweet classification. In: Findings of the ACL: EMNLP 2020. pp. 1644–1650. ACL, Online (Nov 2020)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., et al.: Language models are few-shot learners. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
7. Derczynski, L., Bontcheva, K., Liakata, M., Procter, R., Wong Sak Hoi, G., Zubiaga, A.: SemEval-2017 task 8: RumourEval: Determining rumour veracity and support for rumours. In: Proc. of the 11th (SemEval-2017). pp. 69–76. ACL, Vancouver, Canada (Aug 2017)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. of the 2019 NAACL: Human Language Technologies. pp. 4171–4186. ACL, Minneapolis, Minnesota (Jun 2019)
9. Fan, L., White, M., Sharma, E., Su, R., Choubey, P.K., Huang, R., Wang, L.: In plain sight: Media bias through the lens of factual reporting. In: Proc. of the 2019 EMNLP-IJCNLP. pp. 6343–6349. ACL, Hong Kong, China (Nov 2019)
10. Gao, T., Fisch, A., Chen, D.: Making pre-trained language models better few-shot learners. In: Proce. of the 59th ACL and the 11th IJCNLP. pp. 3816–3830. ACL, Online (Aug 2021)
11. Glandt, K., Khanal, S., Li, Y., Caragea, D., Caragea, C.: Stance detection in COVID-19 tweets. In: Proc. of the 59th ACL and the 11th Inter. Joint Conf. on Natural Language Processing. pp. 1596–1611. ACL, Online (Aug 2021)
12. Hardalov, M., Arora, A., Nakov, P., Augenstein, I.: Cross-domain label-adaptive stance detection. In: Proc. of the 2021 EMNLP. pp. 9011–9028. ACL, Online and Punta Cana, Dominican Republic (Nov 2021)
13. He, Z., Mokhberian, N., Lerman, K.: Infusing knowledge from Wikipedia to enhance stance detection. In: Proc. of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis. pp. 71–77. ACL, Dublin, Ireland (May 2022)
14. Jiang, Y., Gao, J., Shen, H., Cheng, X.: Zero-shot stance detection via multi-perspective contrastive learning with unlabeled data. Information Processing & Management **60**(4), 103361 (2023). https://doi.org/10.1016/j.ipm.2023.103361
15. Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., et al.: ChatGPT: Jack of all trades, master of none. Information Fusion **99**, 101861 (nov 2023). https://doi.org/10.1016/j.inffus.2023.101861
16. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proc. of the 58th ACL. pp. 7871–7880. ACL, Online (Jul 2020)
17. Li, Y., Sosea, T., Sawant, A., Nair, A.J., Inkpen, D., Caragea, C.: P-stance: A large dataset for stance detection in political domain. In: Findings of the ACL: ACL-IJCNLP 2021. pp. 2355–2365. ACL, Online (Aug 2021)
18. Li, Y., Zhao, C., Caragea, C.: Tts: A target-based teacher-student framework for zero-shot stance detection. In: Proc. of the ACM Web Conf. 2023. p. 1500–1509. WWW '23, ACM, New York, NY, USA (2023)
19. Liang, B., Chen, Z., Gui, L., He, Y., Yang, M., Xu, R.: Zero-shot stance detection via contrastive learning. In: Proc. of the ACM Web Conf. 2022. p. 2738–2747. WWW '22, ACM, New York, NY, USA (2022)

20. Liang, B., Zhu, Q., Li, X., Yang, M., Gui, L., He, Y., et al.: JointCL: A joint contrastive learning framework for zero-shot stance detection. In: Proc. of the 60th ACL. pp. 81–91. ACL, Dublin, Ireland (May 2022)
21. Liu, R., Lin, Z., Fu, P., Liu, Y., Wang, W.: Connecting targets via latent topics and contrastive learning: A unified framework for robust zero-shot and few-shot stance detection. In: ICASSP 2022 - 2022 IEEE Inter. Conf. on Acoustics, Speech and Signal Processing (ICASSP). pp. 7812–7816 (2022)
22. Liu, R., Lin, Z., Tan, Y., Wang, W.: Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In: Findings of the ACL: ACL-IJCNLP 2021. pp. 3152–3157. ACL, Online (Aug 2021)
23. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019)
24. Liu, Y., Zhang, X.F., Wegsman, D., Beauchamp, N., Wang, L.: POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In: Findings of the ACL: NAACL 2022. pp. 1354–1374. ACL, Seattle, United States (Jul 2022)
25. Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H.W., et al.: The flan collection: Designing data and methods for effective instruction tuning (2023)
26. Luo, Y., Liu, Z., Shi, Y., Li, S.Z., Zhang, Y.: Exploiting sentiment and common sense for zero-shot stance detection. In: Proc. of the 29th COLING. pp. 7112–7123. International Committee on Computational Linguistics, Gyeongju, Republic of Korea (Oct 2022)
27. Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., Cherry, C.: SemEval-2016 task 6: Detecting stance in tweets. In: Proc. of the 10th (SemEval-2016). pp. 31–41. ACL, San Diego, California (Jun 2016)
28. Schick, T., Schütze, H.: Exploiting cloze-questions for few-shot text classification and natural language inference. In: Proc. of the 16th Conf. of EACL. pp. 255–269. ACL, Online (Apr 2021)
29. Schick, T., Schütze, H.: It's not just size that matters: Small language models are also few-shot learners. In: Proc. of the 2021 NAACL: Human Language Technologies. pp. 2339–2352. ACL, Online (Jun 2021)
30. Sobhani, P., Inkpen, D., Zhu, X.: A dataset for multi-target stance detection. In: Proc. of the 15th EACL. pp. 551–557. ACL, Valencia, Spain (Apr 2017)
31. Tay, Y., Dehghani, M., Tran, V.Q., Garcia, X., Wei, J., Wang, X., et al.: (2023)
32. Wen, H., Hauptmann, A.: Zero-shot and few-shot stance detection on varied topics via conditional generation. In: Proc. of the 61st ACL. pp. 1491–1499. ACL, Toronto, Canada (Jul 2023)
33. Xu, H., Vucetic, S., Yin, W.: OpenStance: Real-world zero-shot stance detection. In: Proc. of the 26th CoNLL. pp. 314–324. ACL, Abu Dhabi (Dec 2022)
34. Zhang, B., Ding, D., Jing, L.: How would stance detection techniques evolve after the launch of chatgpt? (2023)
35. Zhao, X., Zou, J., Zhang, Z., Xie, F., Zhou, B., Tian, L.: Feature Enhanced Zero-Shot Stance Detection via Contrastive Learning, pp. 900–908 (2023)
36. Zhu, Q., Liang, B., Sun, J., Du, J., Zhou, L., Xu, R.: Enhancing zero-shot stance detection via targeted background knowledge. In: Proc. of the 45th Inter. ACM SIGIR Conf. on Research and Development in IR. p. 2070–2075. SIGIR '22, ACM, New York, NY, USA (2022)

ICCS Camera Ready Version 2024
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-63751-3_3