Target Generation by Llama

Abstract

Keywords: Zero-shot stance detection, domain generalization.

1 Introduction

2 Literature Review

3 Methodology

Problem Definition

Let the training dataset be represented as $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{t}_i, \mathbf{y}_i)\}_{i=1}^{n}$, where $\mathbf{x}_i$ denotes the input text, $\mathbf{t}_i$ is the corresponding target, and $\mathbf{y}_i$ is the stance label categorized as "*Against*", "*Favor*" or "*Neutral*". The test dataset is represented as $\mathcal{D}' = \{(\mathbf{x}'_j, \mathbf{t}'_j)\}_{j=1}^{n'}$, where $\mathbf{t}'_j$ corresponds to novel targets not encountered during training. The core objective of ZSSD is to construct a model that, trained on $\mathcal{D}$, can predict the stance of $\mathbf{x}'_j$ with respect to unseen targets $\mathbf{t}'_j$ from $\mathcal{D}'$.
ZSSD can be formulated as:

$$f(x_{\text{test}}, t_u) = g(f(E_{x_{\text{train}}}, E_{t_s}), \Psi(E_{x_{\text{test}}}, E_{t_u})) \tag{1}$$

where:

- x_{train} represents the input text during training.
- x_{test} represents the input text during testing.
- $E_{x_{\text{train}}}$ and $E_{x_{\text{test}}}$ are the embeddings of the training and inference texts, respectively.
- E_{t_s} represents the embedding of seen targets (available during training).
- E_{t_u} represents the embedding of unseen targets (avilable during testing).
- $f(E_{x_{\text{train}}}, E_{t_s})$ learns representations from the training text and seen targets.
- $\Psi(E_{x_{\text{test}}}, E_{t_u})$ models the relationships between the inference text and unseen target.
- $g(\cdot)$ transfers the learned knowledge to predict stance for unseen targets.

3.1 Model Architecture

The architecture of the proposed ZSSD framework model is illustrated in Figure 2. The individual components are described in the subsections below.

3.1.1 Input Representation and Tokenization

3.1.2 Transformer-based Encoder

3.1.3 Interaction between *text* and *topic* Representations

3.1.4 Classification and Label Smoothing

3.2 Datasets

Table 1 provides an overview of the datasets used in the model study. Each dataset is designed to evaluate different aspects of stance detection, covering diverse topics.

1. **SemEval-2016**[6]: This dataset consists of five predefined topics: Atheism, Climate Change is a Real Concern, Feminist Movement, Hillary Clinton, and Legalization of Abortion. Each sample is annotated as *Favor*, *Against*, or *None*.
2. **ArgMin**[6]: The ArgMin dataset is centered on argument mining and includes eight topics: Abortion, Cloning, Death Penalty, Gun Control, Marijuana Legalization, Minimum Wage, Nuclear Energy, and School Uniforms. Labels assigned to each sample are *Support*, *Oppose*, or *Neutral*.
3. **COVID-19**[6]: This dataset focuses on four COVID-19-related topics: Wearing a Face Mask, Anthony S. Fauci, School Closures, and

Table 1: Statistics of the datasets used in the study of model.

Dataset	**Train**	**Val**	**Test**	**Targets**
SemEval-2016 [1]	2,160	359	1,080	Atheism, Feminist Movement, Hillary Clinton, Legalization of Abortion
ArgMin [2]	18,341	2,042	5,109	Abortion, Cloning, Death Penalty, Gun Control, Marijuana Legalization, Minimum Wage, Nuclear Energy, School Uniforms
COVID-19 [3]	4,533	800	800	Face Masks, Fauci, Stay at Home Orders, School Closures
P-Stance [4]	17,224	2,193	2,157	Joe Biden, Bernie Sanders, Donald Trump
VAST [5]	13,477	2,062	3,006	Few-shot Topics: 638-train, 114-val, 159-test; Zero-shot Topics: 4,003-train, 383-val, 600-test
VAST with Data Filtering	11,772	2,062	3,006	Few-shot Topics: 638-train, 114-val, 159-test; Zero-shot Topics: 4,003-train, 383-val, 600-test

Stay-at-Home Orders. Samples are labeled as *Favor*, *Against*, or *None*.

4. **P-Stance**[**6**]: The P-Stance dataset examines three topics related to the 2020 U.S. presidential election: Donald Trump, Joe Biden, and Bernie Sanders. Each instance is labeled as *Favor* or *Against*.
5. **TSE**[**6**]: In [6], the authors utilized four datasets—SemEval-2016, ArgMin, COVID-19, and P-Stance—for the task of target stance extraction. Collectively, these datasets are referred to as the TSE dataset.
6. **VAST** [**5**]: The VAST dataset is specifically designed for zero-shot stance detection, covering a wide range of topics, including politics, education, and public health. It is derived from New York Times article comments and provides diverse, realistic topics to support the development of more generalizable stance detection models.

4 Experimental Setup

4.1 Training Settings

We carried out all model variant experiments on a single NVIDIA RTX A4000 16 GB GPU and used the pre-trained BART-large[1] model, fine-tuning only its encoder while excluding the decoder to manage memory constraints effectively and also BERT[2] encoder.

The learning rates were configured as 2e-5 for the BART and BERT encoders and 1e-3 for the fully connected layers, with AdamW serving as the optimizer. We set the mini-batch size to 64 and limited the maximum sequence lengths to 200 for texts and 10 for targets. Training was performed over 4 epochs, incorporating early stopping with a patience of 5 to prevent overfitting. A dropout rate of 0.1 was applied, and all experiments were run using a single seed value of 0

4.2 Baselines

We evaluate our proposed model model against three baselines from the target-based stance detection framework by Li et al. [7] are BERT-joint, TTS-base, and TTS-p.

4.2.1 BERT-Joint

BERT Joint architecture implemented with BERT transformer as proposed in [7] and [5] by considering both textual content and topic information in a joint representation as [CLS] text [SEP] topic [SEP]

4.2.2 TTS-Base

TTS-base is the teacher model in the target-based teacher-student framework[7] . It is implemented

[0] https://ai.google.dev/gemini-api/docs

[2] https://huggingface.co/facebook/bart-large-mnli

[2] https://huggingface.co/google-bert/bert-base-uncased

Table 2: Ambiguous samples found from the VAST dataset.

Text	**Target**	**Stance**
A small restaurant near us, affiliated with a state university's hospitality program, did away with tipping early this year. It has worked quite well – I rather enjoy tipping generously for good service – that sort of work, after all, is a livelihood for many and should be more widely regarded as a profession. But the no-tipping system does remove considerable uncertainty as to whom and how much to tip, which I think is a source of resentment for some people.	Tip	AGAINST
	Tip	NONE
	Tip	FAVOR

Table 3: Comparative Analysis of model performance using base data (VAST), Augmented data, and Data Filtering (VAST-DF) with BERT and BART encoders in terms of macro-averaged F1 Scores.

Model	**100% Training Data**				**10% Training Data**			
	Against	***Favor***	***Neutral***	**Average**	***Against***	***Favor***	***Neutral***	**Average**
BERT [7]	0.658	0.559	0.895	0.704	0.552	0.496	0.888	0.645
TTS-base [7]	0.719	0.698	0.923	0.780	0.653	0.708	0.910	0.757
TTS-p [7]	0.751	0.725	0.925	0.801	0.721	0.719	0.913	0.784
model-Base+BERT	0.653	0.623	0.898	0.725#	0.563	0.563	0.893	0.673#
model-Aug+BERT	0.636	0.642	0.905	0.728#	0.527	0.604	0.891	0.674#

# : Significant compared to BERT ($p < 0.05$), * : Significant compared to TTS-base ($p < 0.05$), † : Significant compared to TTS-p ($p < 0.05$)

with BART encoder and trained on the VAST data without target-based data augmentation

4.2.3 TTS-p

TTS-p is the student model in the target-based teacher-student framework[7]. It is implemented with BART encoder and trained on the VAST data along with target-based data augmentation.

4.3 model Models

The model models are built based on the architecture illustrated in Figure 2. and configured in different ways to evaluate the impact of various transformers

5 Results & Discussion

However, The model models, particularly when coupled with BART encoder, augmentation, and data filtering, surpass baseline models, indicating the effectiveness of data augmentation and filtering strategies in improving stance classification. We performed statistical significance tests to evaluate the performance improvements of the model models. The models were compared against three baselines—BERT, TTS-base, and TTS-p. Significant differences ($p < 0.05$) were marked with symbols [#, *, †] to highlight those improvements.

Table 4. shows the performance of our model on the TSE dataset, viz. SemEval-2016, ArgMin, Covid-19, and P-Stance, using the macro-averaged F1 score. In the single-task framework [6] focused on stance detection, our model achieved the highest scores across all datasets.

We also explore whether our proposed model could be used to improve other existing ZSSD baselines, including BERTweet from [13] , BERT, TTS-p, and TTS-base models from [7] on various datasets. model outperforms BERTweet in stance detection, with improvements on SemEval-2016 (+0.84%), ArgMin (+19.89%), Covid-19

Table 4: Comparative analysis of model across various models from [6] on the TSE dataset for stance detection. Results are reported in terms of macro-averaged F1 scores.

Model	**SemEval-2016**	**ArgMin**	**Covid-19**	**P-Stance**	**Average**
BiLSTM [8]	53.05	45.7	53.34	73.62	56.43
BiCond [9]	52.63	46.96	58.73	74.56	58.22
TAN [10]	55.26	50.85	56.83	74.67	59.40
CrossNet [11]	61.06	50.79	65.89	75.08	63.21
TGA-Net [12]	63.74	58.71	64.70	77.70	66.21
BERTweet [13]	68.03	64.31	72.99	81.47	71.70
model (Proposed)	**68.60**	**77.10**	**78.80**	**83.60**	**77.03**

(+7.96%), and P-Stance (+2.62%). In VAST (100%), it achieves a +14.49% gain over BERT but shows smaller improvements over TTS-base (+3.33%) and TTS-p (+0.62%) due to the diverse number of targets.

Table 5: Label-wise performance comparison of model on TSE dataset in terms of macro-averaged F1 score for *Against*, *Favor*, and *Neutral* labels.

Model	**Dataset**	***Against***	***Favor***	***Neutral***	**Average**
model	SemEval-2016	0.785	0.622	0.651	0.686
model	ArgMin	0.737	0.729	0.847	0.771
model	Covid-19	0.741	0.803	0.821	0.788
model	P-Stance	0.848	0.825	—	0.836

The Table 5. represents the performance of model evaluated across all the lables of the datasets, viz. *Against*, *Favor* and *Neutral* labels.

5.1 Ablation Study

To understand how different parts of our model contribute to its performance, we conducted a series of experiments. We focused on two key aspects: the attention mechanism and the type of loss function used for training.

5.1.1 Attention Mechanism

m

5.1.2 Loss Function

We tested two types of loss functions: standard Cross-Entropy loss (L_{CE}) and

5.2 Cross-domain performance analysis

Table 6: Performance of model-Aug+BART+DF model in cross-domain settings in terms of macro-averaged F1 scores.

Train Data	**Test Data**	***Against***	***Favor***	***Neutral***	**Avg**
TSE	VAST	0.752	0.750	0.881	0.794
VAST	TSE	0.629	0.459	0.562	0.550
	SemEval-2016	0.595	0.459	0.495	0.516
	ArgMin	0.541	0.457	0.552	0.517
	Covid-19	0.513	0.379	0.751	0.548

5.3 Error Analysis

The confusion matrix of model-

5.3.1 False Positives (*Favor* Misclassified as *Against*)

A detailed review of the model model's predictions shows that several factors lead to incorrect stance classification. One main issue is misunderstanding negations, where the model does not properly recognize words like "not," which can completely change the stance. Conditional statements also create confusion when support is given with conditions, making it unclear whether the stance is truly in *Favor* or *Against*. Another challenge is sarcasm and irony [14], where a supportive statement may appear negative if the model fails to detect the intended meaning. Mixed sentiments—where both support and opposition appear in the same statement—can also mislead the model. Lastly, keyword biases cause errors when the model relies too much on certain words instead of understanding the full meaning of the sentence. Some examples of such misclassifications are shown in Table 7.

5.3.2 False Negatives (*Against* Misclassified as *Favor*)

A major source of false negatives is subtle criticism, *Neutral* language, or polite disagreement, which makes opposition less explicit. The model struggles with indirect opposition, where disagreement is implied rather than directly stated. Additionally, positive wording within critical statements can create confusion, leading the model to misinterpret opposition as support. Examples of such misclassifications are shown in Table 8.

6 Conclusion

In this paper, we try to solve the ZSSD problem through improved generalization capabilities in the model architecture. We utilize Transformer-based pretrained language models to capture the linguistic features and context. The relationship between text and target is extracted through the proposed interaction between their embedding vectors. We further improve the generalization ability and efficiency through neural collapse by incorporating label smoothing Cross Entropy loss. Extensive experiments on multiple datasets and ablation studies show that our method - model greatly improves stance prediction ability in ZSSD settings. Besides, with the task specific data augmentation approach, model is suitable for real world applications with limited amount of labeled data. In addition, the error analysis identified the kind of mistakes the model commits, offering a direction for future research to extend the model capabilities in understanding the nuances present in text.

References

[1] Mohammad, S. M., Kiritchenko, S., Sobhani, P., Zhu, X. & Cherry, C. Semeval-2016 task 6: Detecting stance in tweets. *Proceedings of SemEval* 31–41 (2016).

[2] Stab, C., Miller, T., Schiller, B., Rai, P. & Gurevych, I. *Cross-topic argument mining from heterogeneous sources*, 3664–3674 (2018).

[3] Glandt, K., Khanal, S., Li, Y., Caragea, D. & Caragea, C. *Stance detection in covid-19 tweets*, Vol. 1 (2021).

[4] Li, Y. *et al.* *P-stance: A large dataset for stance detection in political domain*, 2355–2365 (2021).

[5] Allaway, E. & Mckeown, K. *Zero-shot stance detection: A dataset and model using generalized topic representations*, 8913–8931 (2020).

[6] Li, Y., Garg, K. & Caragea, C. Rogers, A., Boyd-Graber, J. & Okazaki, N. (eds) *A new direction in stance detection: Target-stance extraction in the wild.* (eds Rogers, A., Boyd-Graber, J. & Okazaki, N.) *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 10071–10085 (Association for Computational Linguistics, Toronto, Canada, 2023). URL https://aclanthology.org/2023.acl-long.560.

[7] Li, Y., Zhao, C. & Caragea, C. *Tts: A target-based teacher-student framework for zero-shot stance detection*, 1500–1509 (2023).

[8] Schuster, M. & Paliwal, K. K. *Bidirectional recurrent neural networks*, 2673–2676 (IEEE, 1997).

[9] Augenstein, I., Hassel, R., Swayamdipta, S. & Vlachos, A. *Target-dependent sentiment classification with conditional encoding*, 673–683 (Association for Computational Linguistics, 2016).

[10] Du, J., Shao, P. & Qu, Y. *Tan: Target-aware network for stance detection*, 317–327

Table 7: Examples of *Favor* misclassified as *Against*.

Text	**Target**	**Actual Stance**	**Predicted Stance**	**Possible Reason**
I don't see why anyone would oppose gun control. It's necessary.	Gun Control	Favor	Against	Misinterpreted negation
I support climate change policies if they are fair to all.	Climate Change	Favor	Against	Confusion due to conditional phrasing
Oh sure, let's just ignore science and keep polluting. Brilliant idea!	Climate Change	Favor	Against	Sarcasm misunderstood
Gun control laws are beneficial, but criminals still find ways around them.	Gun Control	Favor	Against	Mixed sentiment misleads the model
Renewable energy is essential, yet we should not dismiss nuclear power.	Renewable Energy	Favor	Against	Counterpoint mention creates confusion

Table 8: Examples of *Against* misclassified as *Favor*.

Text	**Target**	**Actual Stance**	**Predicted Stance**	**Possible Reason**
I understand the rationale behind gun control, but it won't stop violence.	Gun Control	Against	Favor	Subtle opposition misclassified
Climate change action is crucial, yet current policies are ineffective.	Climate Change	Against	Favor	*Neutral* tone causes confusion
Banning plastic straws is a well-intentioned move, but it won't fix the bigger problem.	Plastic Ban	Against	Favor	Positive phrasing within a critical statement
Electric cars are promising, but their limitations cannot be ignored.	Electric Cars	Against	Favor	Balanced criticism misclassified
I'm unsure if universal healthcare is the best approach.	Universal Healthcare	Against	Favor	Weakly stated opposition misunderstood

(Association for Computational Linguistics, 2017).

[11] Xu, L., Li, N., Li, L., Ji, H. & Chen, Y. *Cross-net: Enhancing targeted stance detection with attention mechanisms*, 550–560 (Association for Computational Linguistics, 2018).

[12] Allaway, E. & McKeown, K. *Tga-net: A topic-grouped attention network for stance detection*, 1234–1244 (Association for Computational Linguistics, 2020).

[13] Li, X., Han, P., Xu, J. *et al. Bertweet: A pre-trained language model for english tweets*, 5000–5011 (Association for Computational Linguistics, 2021).

[14] Chia, Z. L., Ptaszynski, M., Masui, F., Leliwa, G. & Wroczynski, M. Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. *Information Processing & Management* **58**, 102600 (2021).

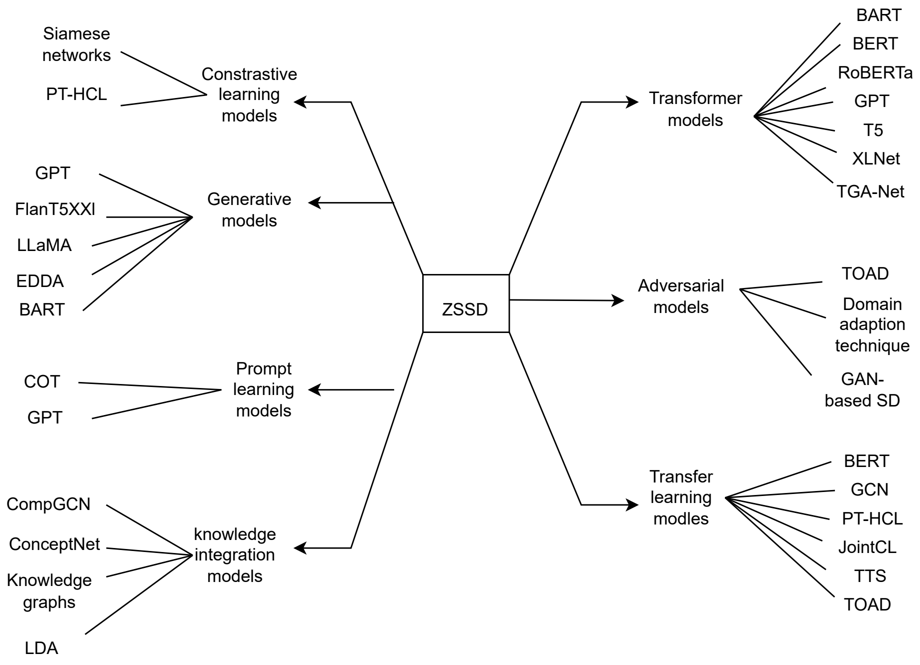

Fig. 1: Literature review of models used in ZSSD.

figures/model_Arch.pdf

Fig. 2: Architecture of model for the Zero-Shot Stance Detection model.

figures/model Confusion matrix (1).jpeg

Fig. 3: Confusion matrix for the model-Aug+BART model on the VAST-DF ataset. The labels are: 0 - *Against*, 1 - *Favor*, and 2 - *Neutral*.