Empirical Evaluation of Topic Zero- and Few-Shot Learning for Stance Dissonance Detection

Anonymous ACL submission

Abstract

We address *stance dissonance detection*, the task of detecting conflicting stance between two input statements. Computational models for traditional stance detection have typically been trained to indicate pro/con for a given target topic (e.g. gun control) and thus do not generalize well to new topics. In this paper, we systematically evaluate the generalizability of this task to situations where examples of the topic has not been seen at all (zero-shot) or only a few times (few-shot). We first build a large-scale dataset of stance dissonance detection from an online debate platform, consisting of 23.8k pairs of statements from 34 diverse topics. We show that stance dissonance detection models trained only on a small number of non-target topics already perform as well as those trained on a target topic. We also show that adding more non-target topics further boosts the performance, indicating the generalizability of non-target topics to a target topic in the stance dissonance detection task.

1 Introduction

It has been suggested that the main point of human reasoning is to support stance argumentation (Mercier and Sperber, 2011). Techniques to better capture stance and argumentation have wide ranging applications from an educational strategy for facilitating learning (Schwarz and Asterhan, 2010; Scheuer et al., 2010) to tracking political opinions (Thomas et al., 2006).

We address the problem of identifying (dis)agreement between utterances that express stances towards a topic (Bar-Haim et al., 2017; Xu et al., 2019; Körner et al., 2021) (henceforth, *stance dissonance detection*). Given two claims c_1, c_2 under topic t, the task is to classify them into either (i) CONSONANCE if the stance suggested by c_1 towards t is the same as that by c_2, (ii) DISSONANCE if the stance suggested by c_1 towards t is the opposite to that by s_2, or (iii) NEITHER (see Table 1 for examples). This is a challenging task that tries to understand (dis)agreement between consecutive utterances where the topic of contention (henceforth, *target topic*) is not always explicitly stated. Such instances are found abundantly in comments, replies and responses to videos, news articles and other online media content.

Stance detection (Küçük and Can, 2020) is conventionally modeled as a single document (topic-dependent) classification task, whereby models are trained for each potential target topic (e.g. gun control, abortion, etc.) (Hasan and Ng, 2013; Mohammad et al., 2016; Xu et al., 2019). However, such an approach can only be applied to topics that are pre-specified and for which training data is available. Yet one can express stance on endless topics — local, situational, or new — for which training data is not available.

To address this issue, stance detection has also been studied in the *topic-zero-shot (*TZS*)* and *topic-few-shot (*TFS*)* settings (Xu et al., 2018; Stab et al., 2018; Zhang et al., 2020; Allaway and McKeown, 2020; Allaway et al., 2021; Hardalov et al., 2021; Reuver et al., 2021), where only training data from non-target topics is available (TZS), or only a small number of training data from target topics is available (TFS). The main limitation of these studies is that they experiment on a small number of non-target topics (up to eight topics): it is still unclear how a larger number of non-target topics is generalized in the stance detection task.

Here, we conduct a large-scale empirical study on how training data from non-target topics impacts on TZS and TFS stance dissonance detection models. Our contributions are:

- We evaluate the topic generalizability of stance dissonance detection models on a topic-diverse, large-scale corpus. Our dataset (§3) consists of 23.8k claim pairs from 34 diverse topics, larger than existing studies using, e.g.,

eight (Reuver et al., 2021), five (Xu et al., 2019) or two topics (Körner et al., 2021)).

- We show that TZS stance dissonance detection models trained on a small number (only 4) of non-target topics already perform as well as those trained on a target topic (§4).

- We show that adding more non-target topics further boosts the performance, indicating the generalizability of non-target topics to unseen topics in the stance dissonance detection task.

2 Related work

Stance detection is the problem of identifying the stance expressed by a text towards a target (Küçük and Can, 2020), which is often generalized as stance dissonance detection (Rosenthal and McKeown, 2015; Xu et al., 2019; Körner et al., 2021). Bar-Haim et al. (2017) introduce contrast detection between two short, concise topic phrases, whereas more recent work (Xu et al., 2019; Körner et al., 2021) examines the claim-pair stance in the context of a topic and examines cross-topic transferability among two topics.

While conventional stance detection models are trained for a specific target topic, recent work has explored topic-general stance detection models in a cross-target setting (Xu et al., 2018; Stab et al., 2018; Zhang et al., 2020; Hardalov et al., 2021; Kaushal et al., 2021; Reuver et al., 2021) and TZS/TFS settings (Allaway and McKeown, 2020; Allaway et al., 2021). For example, Xu et al. (2018) examine the cross-target applicability and transferability of traditional stance detection tasks. Zhang et al. (2020) incorporate external emotion, sentiment lexicon, and world knowledge to achieve better performance on the cross-target setting. Allaway et al. (2021) eliminate any topic-specific information from models using adversarial training. However, most of these studies use a corpus comprising a small number of topics, such as the SemEval-2016 Task 6 corpus (Mohammad et al., 2016) which contains six topics. In contrast, we explore the impact of non-target topics on a corpus of 32 topics, many more than prior studies.

An exception is Allaway and McKeown (2020), who introduce a corpus for TZS/TFS stance detection consisting of 23.5k statements over (possibly semantically overlapping) 5,634 topics (4 statements per topic on avg.). In contrast, our dataset is designed to have diverse 34 topics and is more topic-dense (700 claim pairs per topic; see §3), which provides us a more direct route to explore the generalizability of non-target topics at scale.

Our task is also similar to a broad range of NLP tasks seeking to identify some type of relation between spans of text. Notable instantiations of this problem include Discourse Relation Identification (Prasad et al., 2008; Bosc et al., 2016), Semantic Textual Similarity (Cer et al., 2017), and Textual Entailment Task (Bowman et al., 2015; Williams et al., 2018). However, few studies have investigated generalizability of models.[1] Our work is particularly pertinent to the argument mining community, where most existing work focuses on discourse level or long-form texts for a limited number of targets or topics (Menini and Tonelli, 2016; Cocarascu and Toni, 2017; Menini et al., 2018). Some work has sought to annotate and classify discourse arguments in tweets that support or attack each other (Bosc et al., 2016), but focused on argumentation-level support/attack, as opposed to a generalized, topic-level approach.

3 Data collection

3.1 Source data

To build a dataset for stance dissonance detection with a large number of diverse topics, we extract arguments from Kialo[2], one of the popular online debate platforms. The arguments on Kialo are tree-structured: given a topic claim (i.e. a starting statement which is being debated, such as *Should vaping be banned?*), users write claims, explicitly labeling their stance (either pro or con) on the topic claim. Other users can also reply to each claim with pro/con labels. At the time of submission, Kialo has 16,884 topic claims and 637,383 pro/con claims.

3.2 Extracting claim pairs

Our goal is to collect claim pairs from diverse topics. To this end, we manually choose seed 72 topics which are semantically dissimilar to each other, and then extract any claim pairs in a parent-child relationship. Given a claim pair c_1, c_2, we label them as (i) CONSONANCE if c_1 is a pro claim for c_2, or (ii) DISSONANCE if c_1 is a con claim for c_2.

[1] Williams et al. (2018) created a large-scale corpus of textual entailment from diverse sources of texts including government websites and telephone conversations, and analyzed the domain-generalizability of textual entailment models. However, domain here is a type of document rather than a topic.

[2] https://www.kialo.com/

Label	# topics	# claim pairs	Example (topic: *Should Zoos be banned?*)
CONSONANCE	34	7,559	c_1: Animals in zoos are not usually trained to do tricks, and when they are in contemporary zoos, it is done through a reward based system as an enrichment exercise. c_2: Trainers reinforce desirable behavior with a variety of rewards, and do not draw attention to undesirable behavior.
DISSONANCE	34	8,289	c_1: Zoos cause suffering and harm to animals. c_2: We are unable to understand how, or even if, animals feel pain in a way that is remotely similar to how humans do. We should therefore prioritise quantifiable human utility.
NEITHER	34	7,952	c_1: Dogs were created by humans selectively breeding wolves. c_2: Humans do not have a right to breed, capture and confine other animals, even if they are endangered.

Table 1: Summary of the constructed dataset. Our dataset has a diverse, larger number of topics, and each topic has 700 labeled claim pairs.

To ensure that the absence of a relation between any two unrelated claims is also captured by stance dissonance detection models, we artificially created pairs of claims randomly chosen from the same topic and labeled them as NEITHER.

3.3 Postprocessing

To ensure a reasonable number of claims for each topic, we eliminate topics consisting of fewer than 700 claim pairs. To balance the number of claim pairs per topic, we randomly sample 700 claim pairs from each topic for use in our experiments.

Our final dataset consists of 34 topics, each of which consists of 700 claim pairs. This enables a large-scale empirical study on the impact of non-target topics for TZS/TFS stance dissonance detection models. The summary statistics of our dataset along with examples are shown in Table 1.

4 Experiments

4.1 Model

We use RoBERTA-base (Liu et al., 2019) to obtain a representation of each input claim pair. Given a pair of claim c_1, c_2, the input to the model is of the following form: "[CLS]c_1 [SEP] c_2 [SEP]". We then take the contextualized word embedding $\mathbf{x} \in \mathbb{R}^d$ of [CLS] in the final layer and feed it into the linear classifier: $y = \text{softmax}(W\mathbf{x} + \mathbf{b})$, where $W \in \mathbb{R}^{d \times 3}, \mathbf{b} \in \mathbb{R}^3$ is a learned model parameter.

We trained the model parameters (along with all weights in RoBERTa) with a cross entropy loss for 10 epochs, using AdamW with the learning rate of 3×10^{-5}, the batch size of 16 and warm up ratio of 0.1.[3] To avoid overfitting, we use early stopping (patient of 5) with a macro-averaged F1.

[3]We used huggingface's transformer https://github.com/huggingface/transformers.

4.2 Target topics

To explore the generalizability of topics in the stance dissonance detection task, we select a diverse set of target topics that are dissimilar to each other. In our experiment, we encode all topics into sentence embeddings with Sentence Transformers (Reimers and Gurevych, 2019)[4] and apply k-means clustering ($k = 5$). We then identify one topic closest to the centroid of each cluster.

This yields the following five, mutually exclusive target topics: (i) *Should Zoos Be Banned?*, (ii) *Was Donald Trump a Good President?*, (iii) *Free Will or Determinism*, (iv) *Should "women-only" spaces be open to anyone identifying as a woman?*, and (v) *Should European Monarchies Be Abolished?*. As a final result, we report an average of Macro-F1s for each target topic.

4.3 Training configurations

For each target topic, we train stance dissonance detection models with the following configurations.

Topic-Zero-Shot (TZS) To explore the pure generalizability of non-target topics, we use only training data from 33 (=34-1) non-target topics and do *not use any* training data from the target topic.

Topic-Few-Shot (TFS) In practice, it is not difficult to create a small number of training instances for a given target topic. We train on *a small number of* claim pairs from the target topic in addition to pairs from 33 non-target topics. In our experiments, we randomly sample 20 (TFS-20) or 50 instances (TFS-50) from the target topic.

Full-Shot (FL) To estimate the baseline performance, we train the model *only* on the target topic (FL-0). This roughly corresponds to conventional

[4]all-mpnet-base-v2 at https://www.sbert.net/.

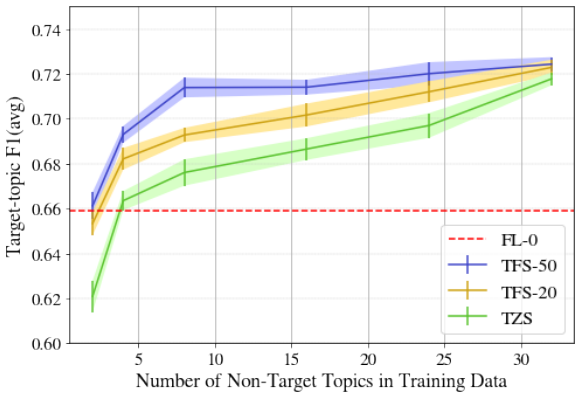

Figure 1: Effect of non-target topics in the topic-zero-/few-shot setting. The models trained only on a small number of non-target topics (TZS,TFS-20/50) already perform as well as those trained only on the target topic (FL-0). Adding more non-target topics further boosts the performance of TZS/TFS models. The shaded area is the standard error of 25 trials (5 targets $\times$ 5 trials).

Setting	# non-target topics	# target samples	Target-topic F1 (avg.)
FL-32	32	560	0.747
TFS-50	32	50	**0.732** (↓ **0.015**)
TFS-20	32	20	0.729 (↓ 0.018)
TZS	32	0	0.718 (↓ 0.029)

Table 2: Performance loss of TZS/TFS models from the full-shot model (FL-32) under 32 non-target topics. The few-shot models trained on only 20 or 50 examples from a target topic (TFS-20/50) has a significantly small loss from the full-shot model (FL-32). Standard error for all these settings is 0.003.

stance detection models. To estimate the upper bound performance, we also train the model on all topics including both the target topic and 32 non-target topics (FL-32).

To see the effect of non-target topics, we vary the number of non-target topics from 2 to 32. For each size k, we create a five random set of k topics and average Macro F1s over these trials.

4.4 Results

Fig. 1 shows the effect of increasing number of non-target topics under the TZS/TFS setting. Surprisingly, the TFS-50 trained on *only two non-target topics and 50 target-topic samples* has already F1 comparable to the full-shot model (FL-0). The other models also outperform the full-shot model when trained on a sufficient number of non-target topics ($\geq$ 4). As the number of non-target topics increases, the performance further improves: even TZS significantly outperforms FL-0 at 32 topics. This indicates the great potential of non-target topic samples: there are a large amount of topic-independent cues in stance dissonance detection, which are seemingly captured by the model.

This begs the question of how well these approaches compare to training with all the topics, including the target topic. The performance loss of TZS/TFS models compared to the full-shot model using 32 non-target samples is shown in Table 2. Surprisingly, the drop in performance observed when cutting down the target-specific training samples from 560 (FL-32) to 50 samples (TFS-50) is comparable to further reducing target-specific samples to 20 (TFS-20).

The results show that the stance dissonance detection models trained on a small number of topics exhibit an impressive ability to generalize to previously unseen target topics and exhibit further performance gains when exposed to a small number of samples from the target topic. This indicates that the model learns topic-independent cues, and that underlying patterns of arguments to signify the dissonance between claims can be successfully captured with non-target topics.

5 Conclusions

This paper weighs in on a key problem as NLP expands more and more from word-level models into models describing semantic discourse relations: the role of topic diversity at training time for generalizing to new topics. To this end, we have addressed the problem of stance dissonance detection in the TZS/TFS setting. To investigate the impact of non-target topics for stance dissonance detection, we have built a large-scale dataset of stance dissonance detection from an online debate platform, consisting of 23.8k claim pairs from 34 diverse topics. In the case of consonance and dissonance of stance, we find that models continue to improve under a "topic independent setting" (i.e. with zero- or few-shots of the topic) all the way up to having learned from 32 non-target topics. Our experiments also revealed that TZS/TFS stance dissonance detection models trained on only a small number of non-target topics already perform as well as those trained on a target topic, and that adding more non-target topics further boosts performance, indicating the generalizability of non-target topics to unseen topics in the stance dissonance detection task.

Ethical Considerations

To create the dataset (§3), we use publicly available dataset on the web. We are restricted to only document-level information; No user-level information is used.

References

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

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

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Tom Bosc, Elena Cabrio, and Serena Villata. 2016. DART: a dataset of arguments and their relations on Twitter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1258–1263, Portorož, Slovenia. European Language Resources Association (ELRA).

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379, Copenhagen, Denmark. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Ayush Kaushal, Avirup Saha, and Niloy Ganguly. 2021. tWT–WT: A dataset to assert the role of target entities for detecting stance of tweets. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3879–3889, Online. Association for Computational Linguistics.

Erik Körner, Gregor Wiedemann, Ahmad Dawar Hakimi, Gerhard Heyer, and Martin Potthast. 2021. On classifying whether two texts are on the same side of an argument. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10130–10138, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. 53(1).

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

S. Menini, Elena Cabrio, Sara Tonelli, and S. Villata. 2018. Never retreat, never retract: Argumentation analysis for political speeches. In *AAAI*.

Stefano Menini and Sara Tonelli. 2016. Agreement and disagreement: Comparison of points of view in the political domain. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2461–2470, Osaka, Japan. The COLING 2016 Organizing Committee.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74.

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

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn Discourse TreeBank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Myrthe Reuver, Suzan Verberne, Roser Morante, and Antske Fokkens. 2021. Is stance detection topic-independent and cross-topic generalizable? - a reproduction study. In *Proceedings of the 8th Workshop on Argument Mining*, pages 46–56, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sara Rosenthal and Kathy McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177, Prague, Czech Republic. Association for Computational Linguistics.

Oliver Scheuer, Frank Loll, Niels Pinkwart, and Bruce M McLaren. 2010. Computer-supported argumentation: A review of the state of the art. *International Journal of Computer-supported collaborative learning*, 5(1):43–102.

Baruch B Schwarz and Christa SC Asterhan. 2010. Argumentation and reasoning. *International handbook of psychology in education*, pages 137–176.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 327–335, Sydney, Australia. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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 (Volume 2: Short Papers)*, pages 778–783, Melbourne, Australia. Association for Computational Linguistics.

Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks. 2019. Recognising agreement and disagreement between stances with reason comparing networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4665–4671, Florence, Italy. 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.