Detecting Dissonant Stance in Social Media: The Role of Topic Exposure

Vasudha Varadarajan[1], **Nikita Soni**[1], **Weixi Wang**[1]
Christian Luhmann[2], **H. Andrew Schwartz**[1] and **Naoya Inoue**[3]
[1]Department of Computer Science, Stony Brook University
[2]Department of Psychology, Stony Brook University
[3]School of Information Science, Japan Advanced Institute of Science and Technology
{vvaradarajan,nisoni,weixiwang,has}@cs.stonybrook.edu
christian.luhmann@stonybrook.edu
naoya-i@jaist.ac.jp

Abstract

We address *dissonant stance detection*, classifying 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 dissonant stance detection to situations where examples of the topic have not been seen at all or have only been seen a few times. We show that dissonant stance detection models trained on only 8 topics, none of which are the target topic, can perform as well as those trained only on a target topic. Further, adding non-target topics boosts performance further up to approximately 32 topics where accuracies start to plateau. Taken together, our experiments suggest dissonant stance detection models can generalize to new unanticipated topics, an important attribute for the social scientific study of social media where new topics emerge daily.

1 Introduction

A prevalent theory about human reasoning, the argumentative theory, is that its primary function is to support argumentation of one's stance or belief (Mercier and Sperber, 2011). New arguments come up on a daily basis and thus new topics for stance emerge. However, most current approaches to stance detection are restricted to well-established topics, and thus are limited in their applications, such as improving educational strategies to facilitate learning (Schwarz and Asterhan, 2010; Scheuer et al., 2010) or tracking political opinions on the latest concerns (Thomas et al., 2006).

As a step toward stance detection, unrestricted to particular topics, we study the problem of identifying (dis)agreement between two statements under pre-chosen as well as unseen topics (Bar-Haim et al., 2017; Xu et al., 2019; Körner et al., 2021) (henceforth, *dissonant stance 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 c_2, or (iii) NEITHER (see Table 1 for examples). This is a challenging task that tries to understand (dis)agreement between two statements where the topic of contention (henceforth, *target topic*) is not explicitly stated. Such instances are found abundantly in comments, replies and responses to videos, news articles and other online media content.

Here, we question the necessity of the target topic by exploring the impact of non-target topics on transformer-based models. Over a corpus of 34 diverse topics, we conduct a large-scale empirical evaluation on the role of exposure to topics. Our **contributions** include: (a) the evaluation of the role of exposure to other topics when detecting statements with dissonant stance for a target topic using transformer-based models; (b) we show that topic-independent (TOPICINDEP) dissonant stance detection models, which are not exposed to the target topic, can perform as well as those trained on a target topic when exposed to as few as 4 non-target topics during training (§3); (c) we show that adding more non-target topics further boosts the performance, beginning to reach a plateau at approximately 24 to 32 non-target topics, evaluating several transformer-based models; (d) we demonstrate that a topic-independence dissonance model, trained only on **pairs of social media posts**, can transfer to a different social media domain and variant of task (finding dissonance within phrase pairs of **a single post**) with a novel small annotated dataset.

2 Related Work

Stance detection is conventionally modeled as identifying the stance expressed by a statement towards a target topic (Küçük and Can, 2020; Hasan and

Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 151 - 156
November 7, 2022 ©2022 Association for Computational Linguistics

Ng, 2013; Mohammad et al., 2016; Xu et al., 2019; Rosenthal and McKeown, 2015; Xu et al., 2019; Körner et al., 2021; Bar-Haim et al., 2017) We generalize conventional stance detection as dissonant stance detection or contrast detection. Beyond generalized stance detection, identification of dissonance in language has other social scientific applications such as detecting cognitive dissonance (Festinger, 1957).

Generalized stance has previously studied between two short concise statements and without evaluation for the amount of topic exposure (Allaway and McKeown, 2020; Allaway et al., 2021).[1] On the other hand, other work has considered stance detection models in a cross-target settings (Xu et al., 2018; Stab et al., 2018; Hardalov et al., 2021; Kaushal et al., 2021; Reuver et al., 2021; Xu et al., 2019; Körner et al., 2021). Some approaches achieve this through incorporation of external lexical or world knowledge (Zhang et al., 2020) or using adversarial training to eliminate topic-specific information (Allaway et al., 2021). However, most of these studies use a corpus comprising a small number of topics, such as the six topics of SemEval-2016 Task 6 (Mohammad et al., 2016). Importantly, despite these promising results, the question of whether the topic needs to be included at all has remained opened as well as the degree of non-target topic exposure.

3 Methodology

3.1 Dataset

To build a dataset for dissonant stance detection with a large number of diverse topics, we extract arguments from Kialo[2], a popular online debate platforms. Kialo arguments are tree-structured: given a topic claim (i.e. a statement being debated, such as *Should vaping be banned?*), users write claims, explicitly labeling their stance (either pro or con) on the topic statement. Users can 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.

We started with 72 seed 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. Neutral or absent-relations were also captured by dissonant stance detection models, we randomly paired separate claims from the same larger topic and labeled them as (iii) NEITHER- in these pairs, one claim is not a pro or a con to the other. To ensure a reasonable diversity of observations for each topic, we eliminate topics consisting of fewer than 700 claim pairs. We then balance the number of claim pairs by randomly sampling 700 claim pairs from each topic.

The final dataset resulted in 34 topics, each with 700 claim pairs. Existing studies of stance detection typically use a small number of topics, e.g., eight (Reuver et al., 2021), five (Xu et al., 2019) or two topics (Körner et al., 2021)). Our work is focused on large-scale empirical study of the impact of non-target topics (topic-independence) for dissonant stance detection models. The summary statistics and examples are shown in Table 1.

3.2 Model

We use Transformer models to obtain a representation of each input claim pair. In our experiments, we used BERT-base (Devlin et al., 2019), RoBERTA-base (Liu et al., 2019), and ALBERT-base (Lan et al., 2020). Given a pair of claims 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 the model weights) 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 (patience of 5) with a macro-averaged F1.

3.3 Target topics

To explore the generalizability of topics in the dissonant stance detection task, we select a diverse set of target topics that are dissimilar to each other. To ensure the dissimilarity, 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.

[1] The dataset is annotated with "topic-phrase" stance rather than dissonant stance. See §3.1 for details on our dataset.

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

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

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

Label	# topics	# claim pairs	Example (topic: *Should Zoos be banned?*)
CONSONANCE	34	7,559	c_1: Zoos are, by nature, restricted in the space they provide their animals. For many animals, it is much more cramped than the wild. c_2: For some captive animals, the small enclosures provided by zoos are directly related to the infant mortality rate.
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.

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.

3.4 Training configurations

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

TOPICINDEP 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.

INTOPICFEW 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 (INTOPICFEW-20) or 50 instances (INTOPICFEW-50) from the target topic.

INTOPIC To estimate the baseline performance, we train the model *only* on the target topic. This roughly corresponds to conventional stance detection models.

ALLTOPICS To estimate a performance upper bound, we also train on all topics including both the target topic and 32 non-target topics.

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

Approach	F1-co	F1-di	F1-na	F1$_{mac}$
Random	0.325	0.367	0.325	0.339
Majority	0.000	0.519	0.000	0.173
BERT (TOPICINDEP)	0.586	0.673	0.710	0.656
ALBERT (TOPICINDEP)	0.598	0.673	0.726	0.666
RoBERTa (TOPICINDEP)	0.659	0.728	0.756	0.717
RoBERTa (INTOPIC)	0.524	0.637	0.776	0.653
RoBERTa (ALLTOPICS)	0.673	0.742	0.824	0.745

Table 2: Evaluation of approaches for topic independent dissonant stance detection versus baselines and an upperbound of witnessing the topic (INTOPIC, ALLTOPICS).

4 Evaluation

4.1 Results

The results of topic-independence dissonant stance detection models are shown in Table 2. It shows that all the variants of topic-independent dissonant stance detection models significantly outperformed the INTOPIC model. In addition, surprisingly, the RoBERTa(TOPICINDEP) model shows a similar performance to the ALLTOPICS model trained on 32 non-target samples and target-topic samples (i.e. an upperbound). This indicates the great potential of non-target topic samples: there are a large amount of topic-independent cues in dissonant stance detection, which are seemingly captured by the model.

Fig. 1 shows the effect of increasing number of non-target topics under the TOPICINDEP/INTOPICFEW setting. As the number of non-target topics increases, the performance improves: even TOPICINDEP significantly outperforms INTOPIC at 32 topics.

Surprisingly, the INTOPICFEW-50 trained on *only two non-target topics and 50 target-topic samples* has already F1 comparable to the INTOPIC

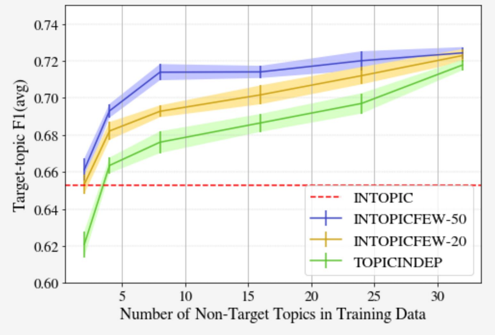

Figure 1: Effect of non-target topics in the topic-independent setting. The models trained only on a small number of non-target topics (TOPICINDEP,INTOPICFEW-20/50) already perform as well as those trained only on the target topic (INTOPIC). Adding more non-target topics boosts the performance of TOPICINDEP/INTOPICFEW models. The shaded area is the standard error of 25 trials (5 targets × 5 trials).

model. The other models also outperform the INTOPIC model when trained on a sufficient number of non-target topics (≥ 4).

This begs the question of how well these approaches compare to training with all the topics, including the target topic. The performance loss of TOPICINDEP/INTOPICFEW models compared to the ALLTOPICS model using 32 non-target samples is shown in Table 3. Surprisingly, the drop in performance observed when cutting down the target-specific training samples from 560 (ALLTOPICS) to 50 samples (INTOPICFEW-50) is comparable to further reducing target-specific samples to 20 (INTOPICFEW-20).

The results show that the dissonant stance 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 underlying patterns of arguments to signify the dissonance between claims can be successfully captured with non-target topics.

4.2 Dissonance generalizability to other domain

We show that the model does not only generalize well over unseen topics, but captures dissonant language in a new domain. To this end, we test the model on a dissonance dataset annotated on a set of tweets parsed into discourse units using (Wang et al., 2018).[5] The annotation is carried out in two stages. First, each unit is annotated as THOUGHT or OTHER. A THOUGHT constitutes of all forms of knowing and awareness: a fact, claim, or statement is a thought. Anything not considered to be a THOUGHT is marked as an OTHER. Second, pairs of THOUGHT units from each tweet are extracted, and then annotated to be either in CONSONANCE, DISSONANCE or NEITHER. The annotations were carried out by a team of three annotators for stage 1 and a team of four annotators for stage 2. The final annotations were extracted by using majority vote and a tiebreaker. To balance the dataset, we choose a test set with 19 pairs of DISSONANCE, 19 pairs of CONSONANCE and 19 pairs of NEITHER. The inputs to the model are not from the training domain, they are tweet discourse units, not entire claims. Thus, this dataset would test the extent to which the model captures dissonance in a single tweet.

Table 4 shows that transferring the ALLTOPICS model trained on Kialo to this domain, without any finetuning, surprisingly still captures DISSONANCE

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

Table 3: Performance loss of TOPICINDEP/INTOPICFEW models from the ALLTOPICS model under 32 non-target topics. The INTOPICFEW models trained on only 20 or 50 examples from a target topic (INTOPICFEW-20/50) has a significantly small loss from the ALLTOPICS model. Standard error for all these settings is 0.003.

Approach	F1-co	F1-di	F1-na	$F1_{macro}$
Majority	0.000	0.519	0.000	0.173
RoBERTa (TOPICINDEP-32)	**0.458**	**0.595**	0.207	0.420

Table 4: Evaluation of the generalization of our approach to hand-annotated Twitter phrases. The topic-independent model trained over the Kialo data still performs substantially better than chance when evaluated over dissonance within (much shorter) Twitter posts.

[5] Tweets are sampled from 2019-2020. The frequency of tweets with dissonant discourse units was found to be about 2.5%.

and CONSONANCE fairly well: the RoBERTA-based model trained on Kialo generalizes well by capturing topic-independent cues.

5 Conclusions

This paper weighs in on a key problem as NLP is increasingly used for studies of social science: the role of exposure to a diverse set of social or political topics and the ability to generalize to new topics. To this end, we have proposed and studied the problem of dissonant stance detection in the TOPICINDEP/INTOPICFEW setting. We find that models continue to improve under a "topic independent setting" but start plateauing at around 8 non-target topics. Our experiments also revealed that TOPICINDEP/INTOPICFEW dissonant stance 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. Further, we find the model trained on the debate forum, where statements are from distinct users, generalizes to a new domain and finding dissonant statements from the same person. Taken together, these results suggest transformer-based dissonant stance detection model can generalize to unseen topics and domains.

6 Ethical Considerations

To create the datasets (§3.1 and §4.2), we use publicly available data on the web. The detection of dissonance has many beneficial applications such as understanding belief trends study of mental health from consenting individuals. But, it also could be used toward manipulating people such via targeted messaging without users' consent. All of our work is restricted to document-level information; No user-level information is used.

Acknowledgements

This work was supported by DARPA via Young Faculty Award grant #W911NF-20-1-0306 to H. Andrew Schwartz at Stony Brook University; the conclusions and opinions expressed are attributable only to the authors and should not be construed as those of DARPA or the U.S. Department of Defense. This work was also supported in part by NIH R01 AA028032-01 and Stony Brook University's Institute for AI-Driven Discovery and Innovation.

References

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of EMNLP*, pages 8913–8931.

Emily Allaway, Malavika Srikanth, and Kathleen McKeown. 2021. Adversarial learning for zero-shot stance detection on social media. In *Proceedings of NAACL: Human Language Technologies*, pages 4756–4767.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of EACL: Volume 1, Long Papers*, pages 251–261.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Leon Festinger. 1957. *A theory of cognitive dissonance*, reissued by stanford univ. press in 1962, renewed 1985 by author, [nachdr.] edition. Stanford Univ. Press, Stanford. OCLC: 255286887.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of EMNLP*, pages 9011–9028.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of IJCNLP*, pages 1348–1356.

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 NAACL: Human Language Technologies*, pages 3879–3889.

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 EMNLP*, pages 10130–10138.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *Association for Computing Machinery*, 53(1).

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. *arXiv*, page 1909.11942.

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. *arXiv*, page 1907.11692.

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 SemEval-2016*, pages 31–41.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pages 3982–3992.

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.

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 SIGDIAL*, pages 168–177.

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 EMNLP*, pages 3664–3674.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of EMNLP*, pages 327–335.

Yizhong Wang, Sujian Li, and Jingfeng Yang. 2018. Toward fast and accurate neural discourse segmentation. In *Proceedings of EMNLP*, pages 962–967.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of ACL (Volume 2: Short Papers)*, pages 778–783.

Chang Xu, Cecile Paris, Surya Nepal, and Ross Sparks. 2019. Recognising agreement and disagreement between stances with reason comparing networks. In *Proceedings of ACL*, pages 4665–4671.

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 ACL*, pages 3188–3197.