Topic-independent Detection of Dissonance in Short Stance Text

Anonymous ACL submission

Abstract

We propose *dissonance detection*, the task of detecting conflicting stance between two input statements. Computational models for stance detection have typically been trained for a given target topic (e.g. gun control). In this paper, we aim for building a computational model for dissonance detection without using training data from the topic of test data. We first build a large-scale dataset of topic-controlled arguments from two sources: (i) an online debate platform, consisting of 15k pairs of statements with support, attack, or no relation from 20 diverse topics, and (ii) Twitter, consisting of 5k pairs of statements from 5 topics. We then evaluate a BERT-based dissonance detection model on this dataset in a topic-controlled manner. Our experiments suggest that dissonance detection models learn the topic-independent patterns of language for detecting dissonance and generalize largely to other arguments in unseen topics.

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 propose the task of topic-independent *dissonance detection*. Given two statements s_1, s_2 under topic t, the task is to classify them into either CONSONANCE if the stance suggested by s_1 towards t is the same as that by s_2, DISSONANCE if the stance suggested by s_1 towards t is the opposite to that by s_2, or NEITHER relation otherwise (e.g. *"Vaping is injurious to health"—"Health problems tend to be caused from unregulated vaping products"* classified into DISSONANCE).

We view topic-independent dissonance detection as an expansion of traditional stance detection (Küçük and Can, 2020), which is typically 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). In this way, models can learn key content that is indicative of stance for the given target topic. However, such an approach can only be applied to topics that are pre-specified and which training data is available, and yet one can express stance on endless topics — local, situational, or new — for which training data is not available.

Here, we thus propose a computational model for dissonance detection in a topic-independent manner. Instead of training a model with examples specific to the target topic, we attempt to train a model that can generally detect when two statements are in opposition, agreement, or neither. We **contribute:** (1) a proposed generalization of the stance detection task into topic-independent dissonance detection, (2) transformer language-model based dissonance detection models (§4), and (3) evaluation for topic-generalizability of our models and traditional stance models (§5), demonstrating that our dissonance models trained on datasets with completely different topics from test data do not experience a significant performance degradation from those trained with-in topic datasets (while the same is not true of traditional stance models). We also modify and repurpose two dissonance detection corpora derived from an online debate forum and Twitter in a semi-automatic manner (§3).

2 Related work

Our task is a generalization of stance detection, the automatic classification of the stance expressed by a piece of text, towards a target, into either: Favor, Against, Neither. The input to such a task is a target domain (e.g. *Legalization of Abortion*), and a piece of text or a statement to infer the stance of the author/speaker from (Küçük and Can, 2020). Some recent work has focused on cross-target stance classification (we focus, rather, on topic independent stance), which is similar to our work in the sense that it explores the generalizability of stance to unseen targets (Xu et al., 2018; Kaushal et al., 2021). Similar to our experiments, Stab et al. (2018) collect a corpus of arguments over a smaller 8 topics to investigate the topic-generalizability of stance detection models.

Although we are not identifying a direct semantic relation between statements, 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), Textual Entailment Task (Bowman et al., 2015; Williams et al., 2018) and argumentative relation prediction (Cocarascu and Toni, 2017). However, few studies investigate the topic-generalizability of models.[1]

Our work is particularly pertinent to the argument mining community, where most existing work focuses on argument mining at the discourse level or long-form texts for a limited number of targets or topics (Menini and Tonelli, 2016; 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 and attack, as opposed to a generalized, topic-level approach to identifying support or attack in the text.

3 Data collection

To build topic-generalizable dissonance detection models, we create two corpora of arguments annotated with topic and consonance/dissonance relations from existing resources: (i) KIALO (§3.1), and (ii) SD16 (§3.2). The summary statistics of each corpus is shown in Table 1.

3.1 KIALO: arguments from debate forum

To obtain clean, topic-diverse arguments, we extract arguments from Kialo.[2] Kialo is one of the popular online debate platforms where people debate on claims. The arguments in Kialo are tree-structured: given a topic (i.e. a *thesis topic*, or a starting statement which is being debated upon, such as *Should vaping be banned?*), the users can add *claims*, i.e. supporting and opposing statements as pros and cons for the topic, and then the other users can add more claims as pros and cons arguments for each claim.

Our goal is to collect arguments with diverse topics but to keep a reasonable amount of arguments per topic. For the purpose of our experiments, we also want to have the same number of arguments per topic. To this end, first, we manually choose mutually exclusive 57 topics. We then choose 20 topics with most frequent claims, and then extract pairs of arguments in a parent-child relationship.

Finally, we label the claim-pro statement pairs as CONSONANCE (e.g. for the topic *Vaping sould be banned*: *Vaping is injurious to health.—There is a public health crisis brought on by vaping in the USA.*), and the claim-con statement pairs as DISSONANCE (e.g. for the topic *Is Gender a Social Construct?*: *Gender roles are natural. Gender theory is just a dangerous invention that denies the "order of creation".—Gender is a social construct, but that doesn't mean it's an invention.*).

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

Dataset	# topics	# statement pairs	Source
KIALO	20	15,300 (5,100 / 5,100 / 5,100)	Debate forum
SD16	5	8,051 (2,683 / 2,656 / 2,702)	Tweets

Table 1: Summary of the constructed dataset. The numbers in the parenthesis indicates the instances of CONSONANCE, DISSONANCE, and NEITHER, respectively.

3.2 SD16: arguments from Tweets

To create the topic-annotated corpus of arguments, we also use the dataset of stance detection. We use the dataset from *SemEval 2016 Task 6: Detecting Stance in Tweets* (Mohammad et al., 2016).[3] In the original task, the Task A dataset has five topics such as *atheism, legalization of abortion, climate change*

[1] a notable exception being Williams et al. (2018) who 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, dissonance relations are fundamentally different from logical entailment relations.

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

[3] https://alt.qcri.org/semeval2016/task6/

is a real concern, feminist movement, Hillary Clinton. The Task B is used to measure performance on a set of tweets from the same pool of five topics, but including a new topic *Donald Trump*. There are 4,870 tweets annotated with stance (favor, against, or neither) (e.g. *The pregnant are more than walking incubators, and have rights!*, favor).

For our experiments, for each topic, we extract a pair of statements annotated with the same stance (favor or against) as CONSONANCE (e.g. for the topic *Feminist Movement*: *If Feminism is not hypocritical fake "equality" then manure sprayed in pink is not fecal. #GamerGate #SemST—@elllode BUT SHE RUNS IN HIGH HEELS #SemST*), a pair of statements annotated with opposite stances as DISSONANCE (e.g. for the topic *Atheism*, *Imagine how amazing the world would be without religion. No wars. No hate (religion wise). No extremist. #SemST—I bind and rebuke the angel of light in the name of Jesus -2 Cor. 11:14 #SemST*), and a random pair of statements as NEITHER.

3.3 Spurious cues

Recent studies report that many NLP datasets has spurious cues unrelated to the task (Ribeiro et al., 2020), which would mislead the results of experiments. By definition, the consonance/dissonance relations signify the relation between the input statements, and this should require dissonance detection models to analyze a *pair* of statements. We thus make sure the failure of a dissonance detection model taking only a *single* statement. In this experiment, we use the BERT-based model as described in §4.2. Ideally, we expect an accuracy similar to random prediction (33.3% on both datasets).

Our experiments show that the BERT-based model achieves an accuracy of 43.0% on KIALO and 41.2% on SD16, which are only slightly better than the random performance. This indicates that the constructed dataset rarely contains spurious cues for dissonance detection.

4 Models

4.1 Baseline model

Given a pair of statements s_1, s_2, we create a sentence representation $\mathbf{s}_1, \mathbf{s}_2$ by averaging word embeddings. We then feed it into a three-way linear classifier to predict consonance/dissonance relations:

$$\mathbf{y} = W \cdot [\mathbf{s}_1 \odot \mathbf{s}_2; \mathrm{abs}(\mathbf{s}_1 - \mathbf{s}_2)] + \mathbf{b}, \quad (1)$$

where $W \in \mathbb{R}^{3\times 2d}, \mathbf{b} \in \mathbb{R}^3$ are the model parameters learned from the dataset, d is the dimension of word embeddings, and $\odot$ is element-wise multiplication. Henceforth, we call it *WordEmbAvg*.

4.2 BERT-based model

We use RoBERTA-base (Liu et al., 2019) to obtain a representation of input statement pair. Given a pair of statements s_1, s_2, the input to the model is of the following form:

$$\text{[CLS] } s_1 \text{ [SEP] } s_2 \text{ [SEP]} \quad (2)$$

We then take the contextualized word embedding of [CLS] in the final layer and feed it into the same three-way non-linear classifier as (Devlin et al., 2019).

5 Experiments

5.1 Setup

5.1.1 Setting

To explore the generalizability of the dissonance detection models to topics unseen in the training set, we explore two settings on KIALO and SD16.

Cross-topic To test the topic-generalizability of the dissonance detection models, we first split each dataset into 5 folds based on the topic of statement pairs and conduct cross validation. For KIALO, each fold has 16 training topics and 4 test topics, where each topic has 765 corresponding statement pairs. For SD16, each fold has 4 training topics and 2 test topics (the topic *Donald Trump* is always used in the test set, similar to the SemEval 2016 Task-6 dataset). The original dataset has a variable number of tweets, and thus a variable amount of potential training data, per topic. To maintain the distribution in the training and test set, we set training set size to 5,175 statement pairs from all topics.

In-topic To estimate the upper bound performance of dissonance detection models, we allow the dissonance detection models to learn clues for dissonance detection from the same-topic arguments (RoBERTa (In-T)). In this setting, we conduct five-fold cross-validation where the split is purely based on instance-level (not topic-based).

5.1.2 Models

For the word-average model (§4.1), we use GloVe (Pennington et al., 2014)-pretrained 300 di-

	Acc	F1-Co	F1-Di	F1-Ne
Kialo				
Random	0.333	0.325	0.339	0.333
Majority†	0.334	0.207	0.323	0.000
WordEmbAvg	0.367	0.004	0.190	0.519
RoBERTa	**0.786**	**0.870**	**0.730**	**0.760**
RoBERTa (In-T)	0.835	0.930	0.780	0.790
SD16				
Random	0.334	0.333	0.334	0.332
Majority	0.334	0.000	**0.501**	0.000
WordEmbAvg	0.350	0.171	0.310	0.457
RoBERTa	**0.587**	**0.536**	0.387	**0.778**
RoBERTa (In-T)	0.635	0.564	0.466	0.828

Table 2: Performance of dissonance detection task in cross-topic settings. RoBERTa outperforms baseline models and performs as well as RoBERTa trained in the in-topic setting (i.e. upper bound performance), indicating that cross-topic dissonance detection is successful. †: The majority baseline has two non-zero F1s because we report an average F1 across five folds, where the majority class is different.

mensional word embeddings.[4] To fine-tune the BERT-based model (§4.2), we set the learning rate to 2e-5, the batch size to 16 and trained each model for 5 epochs.[5] The other hyperparameter settings were the same as those used in RoBERTA-base.

To estimate the integrity of the dissonance detection models, we also show the performance of random classifier (Random) and majority class-based classifier (Majority).

5.2 Results

The results of both cross-topic and in-topic settings are shown in Table 2. Surprisingly, on both datasets, the cross-topic performance is close to the in-topic performance, despite that the training topics are completely different from test topics. This suggests that the dissonance detection models are not only learning domain-specific clues, but also topic-independent clues that are generalizable to other topics.

To see how much topics enable topic-generalizability, we evaluate the BERT-based dissonance detection model on Kialo in the cross-topic setting, trained on various sizes of topics. Specifically, we start from one topic, and incrementally

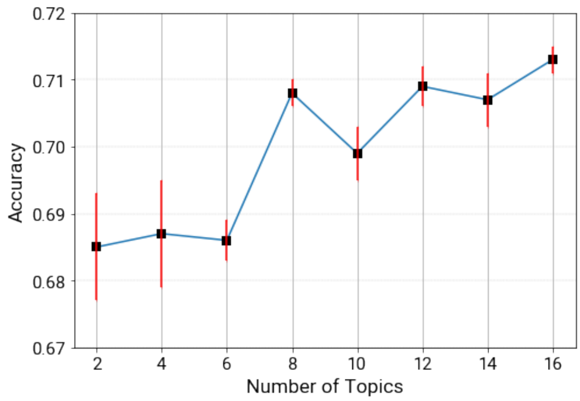

Figure 1: The performance of the model with increasing number of topics in the cross-topic setting. The topics are incrementally added to capture the effects of increase in the number of topics alone.

add more topics. All the models are evaluated on the same test set of 3,060 statement pairs. The results are shown in Fig. 1.

The results show that the dissonance detection model can be reasonably generalized to unseen topics even with the small number of training topics. This indicates that underlying patterns of arguments to signify the dissonance between them is somewhat limited and that the cross-topic model can successfully capture these signals.

6 Conclusions

We have proposed dissonance detection, a generalization of the stance detection task which seeks to detect conflicting stance between two input statements. To build a computational model for dissonance detection without using target test topic at all in the training data, we have built a large-scale dataset of topic-controlled arguments from an online debate platform and Twitter. Our experiments on these datasets have suggested that, while challenging, topic independent stance detection is possible. Our dissonance detection models demonstrated the ability to learn topic-independent patterns for detecting dissonance and generalize largely to other arguments in unseen topics.

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.

[4] CommonCrawl-840B-300d at https://nlp.stanford.edu/projects/glove/.

[5] We used an implementation of huggingface's transformer https://github.com/huggingface/transformers.

References

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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.

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.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. 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).

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. 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.

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.