Same Side Stance Classification

Benno Stein **Yamen Ajjour** **Roxanne El Baff** **Khalid Al-Khatib**
Bauhaus-Universität Weimar
Faculty of Media, Webis Group
`<first>.<last>@uni-weimar.de`

Philipp Cimiano
Bielefeld University
AG Semantic Computing
`cimiano@cit-ec.uni-bielefeld.de`

Henning Wachsmuth
Paderborn University
Department of Computer Science
`henningw@upb.de`

Abstract

This paper introduces the Same Side Stance Classification problem and reports on the outcome of a related shared task, which has been collocated with the Sixth Workshop on Argument Mining at the ACL 2019 in Florence.[1] We have proposed this task as a variant of the well-known stance classification task: Instead of predicting for a single argument whether it has a positive or negative stance towards a given topic, same side classification 'merely' involves the prediction of whether two given arguments share the same stance. The paper in hand provides the rationale for proposing this task, overviews important related work, describes the developed datasets, and reports on the results along with the main methods of the nine submitted systems. We draw conclusions from these results with respect to the suitability of the task as a proxy for measuring progress in the field of argument mining.

1 Introduction

Identifying (i.e., classifying) the stance of an argument towards a particular topic is a fundamental task in computational argumentation and argument mining. The stance of an argument as considered here is a two-valued function: it can either be "pro" a topic (meaning, "yes, I agree"), or "con" the topic ("no, I do not agree").

Here we propose a related though simpler task, called *same side stance classification* (later also referred to as Π_{sameside}). Same side stance classification deals with the problem of classifying two arguments as to whether they (a) share the same stance or (b) have a different stance towards the topic in question.

As an example, consider the following two arguments on the topic "gay marriage", which obviously are on the same side.

Argument 1. Marriage is a commitment to love and care for your spouse till death. This is what is heard in all wedding vows. Gays can clearly qualify for marriage according to these vows, and any definition of marriage deduced from these vows.

Argument 2. Gay Marriage should be legalized since denying some people the option to marry is discriminatory and creates a second class of citizens.

Argument 3 below, however, is neither on the side of Argument 1 nor on the side of Argument 2.

Argument 3. Marriage is the institution that forms and upholds for society, its values and symbols are related to procreation. To change the definition of marriage to include same-sex couples would destroy its function, because it could no longer represent the inherently procreative relationship of opposite-sex pair-bonding.

Same side stance classification is simpler than the "classical" stance classification problem, or at

[1] https://sameside.webis.de/

most equally complex: solving the latter implies solving the former as well.

Aside from the difference in problem complexity a second aspect renders same side stance classification a relevant task of its own right: Stance classification, by definition, requires knowledge about the topic that an argument is meant to address, i.e., stance classifiers must be trained for a particular topic and hence cannot be reliably applied to other (i.e, *across*) topics. In contrast, a same side stance classifier does not necessarily need to distinguish between topic-specific pro- and con-vocabulary; "merely" the argument similarity *within* a stance needs to be assessed. Consequently, same side stance classification is likely to be solvable independently of a topic or a domain—so to speak, in a *topic-agnostic* fashion. Since topic agnosticity is a big step towards application robustness and flexibility, we believe that the development of technologies that tackle this task has game-changing potential.

Last but not least, same side stance classification has a number of useful and important applications related to both argumentation analytics and information retrieval, including but not limited to the following:

- Measuring the strength of bias within an argumentative utterance (analytics).
- Structuring a discussion (analytics).
- Finding out who or what is challenging in a discussion (analytics, retrieval).
- Filtering wrongly-labeled arguments in a large argument corpus, without relying on knowledge of a topic or a domain (retrieval).

To initiate research on same side stance classification, we carried out a first respective shared task in collocation with the Sixth Workshop on Argument Mining at ACL 2019. We report on this shared task and its results in the paper in hand.

The remainder is organized as follows. Section 2 formalizes the same side stance classification task and relates it to other problems in the field. Section 3 points to relevant research and suggested readings related to stance classification. Section 4 describes the dataset and the experiment settings of the shared task. Section 5 reports on the systems of the nine participating teams and their effectiveness. Section 6 concludes with the lessons learned and the planned follow-up resarch.

2 Argument Decision Problems

The same side stance classification task, Π_{sameside}, is a decision task in the field of computational argumentation. As outlined in Section 1, mastering this task is beneficial in the context of argumentation analytics and information retrieval. This section provides a succinct formalization of the problem.

The syntax of the argument model underlying Π_{sameside} is rather simple but well-accepted: An argument consists of a conclusion, c, and a set (a conjunction) of premises, P.

Both premises and conclusions are considered as propositions to which a truth value can be assigned. For this purpose an interpretation function, $\mathcal{I}$, which maps from premises and conclusion to $\{0, 1\}$ can be stated. Based on $\mathcal{I}$ the premises P and the conclusion c can be connected semantically. Recall in this regard the classical notion of entailment, which bases the concept of logical consequence on all possible interpretation functions: Given two propositional formulas α, β, then α entails β (denoted as $\alpha \models \beta$) if and only if for all $\mathcal{I}$ holds:

$$\mathcal{I}(\alpha) = 1 \text{ implies } \mathcal{I}(\beta) = 1 \tag{1}$$

However, for our argument model (and for argumentation in natural language in general) this notion of entailment is not applicable: human language cannot be stuffed entirely into logical formulas; the detection of semantically equivalent argument units (which is necessary to transform formulas whose atoms correspond to argument units) belongs to the hardest NLP problems; truth entailment in natural language is not restricted to a recursive evaluation of truth values but comes in many different flavors such as argument from authority, analogical argument, or inductive argument; and so forth.

In any way, argumentation theory speaks of *acceptability* rather than truth, since truth is often unknown or not accessible (Wachsmuth et al., 2017a). The acceptability of an argument is subjective, which we capture as follows. Given an interpretation function $\mathcal{I}$, propositional premises P, and a propositional conclusion c, then (c, P) is an *acceptable argument* if and only if holds:

$$\mathcal{I}(\wedge_{p \in P}) = 1 \text{ and } \mathcal{I}(c) = 1 \tag{2}$$

Compared to the classical notion of entailment the universality requirement regarding interpretation functions is relaxed. In this vein, (c, P) may be an argument for an individual, for a group, or for all beholders—depending on the respective $\mathcal{I}$. Also, due to the aforementioned reasons, there is no simple structural means[2] that connects the interpretation of c to the interpretation of P: For participants in a debate the interpretation of the premises may be identical, but their mental models *to determine* the truth value of c, as well as the truth value itself, can differ.

The formalization of argument acceptability via interpretation functions as introduced above illustrates how a *belief semantics* for arguments can be formalized. However, the identification and classification of argument *stance* (as treated here as well as treated by other researchers) does not depend on individual interpretation functions. Arguments are formulated purposefully with respect to a thesis, which means that they are always dedicated to be used either as pro or as con argument—independent of the acceptability of a beholder.

To formalize the interesting argument decision problems will consider a propositional thesis t, also called the "main claim", which encodes a particular "side" of a controversial issue. E.g., when referring to the introductory example, t may encode "Gay marriage is a great achievement.", but t may also encode "Gay marriage cannot be tolerated.".[3]

Let $\mathbf{A} = \{(c_1, P_1), (c_2, P_2), \ldots, (c_n, P_n)\}$ be a set of arguments related to t, then we are also given an (implicitly defined) function σ, called "stance", which maps each argument $A \in \mathbf{A}$ either to pro or to con: σ encodes for which side of a controversial issue an argument is devised. A pro argument *supports* t; likewise, a con argument *attacks* t. Two arguments A_1 and A_2 have the same stance iff $\sigma(A_1) = \sigma(A_2)$.

Using these definitions, among others the following decision problems can be stated. Given are a thesis t and a set of related arguments $\mathbf{A}$.

- Π_{sameside}. Decide for two arguments, A_1, A_2 in $\mathbf{A}$ whether or not they have the same stance.

- Π_{stance}. Decide for an argument A in $\mathbf{A}$ whether it has a pro or a con stance, i.e., whether $\sigma(A) = \mathrm{pro}$ or $\sigma(A) = \mathrm{con}$.

Algorithmic stance classification as treated here means to learn the function σ from a set of examples.

3 Related Work

We have first mentioned same side stance classification as a potential task in the context of argument search (Ajjour et al., 2019). Some related previous research has been concerned with the agreement of different texts on a given topic (Menini et al., 2017). In computational argumentation, the task is new to our knowledge, which is why we restrict our view to the most related task in the following: stance classification.

Stance classification has drawn a wide interest in the last decade. The problem has been studied for various linguistic genres including online debates (Somasundaran and Wiebe, 2009; Hasan and Ng, 2013; Ranade et al., 2013), political debates (Vilares and He, 2017), tweets (Addawood et al., 2017; Mohammad et al., 2017), and spontaneous speech (Levow et al., 2014). Stance classification approaches have been motivated by different goals, such as fact checking (Bourgonje et al., 2017; Baly et al., 2018; Nadeem et al., 2019), enthymeme reconstruction (Rajendran et al., 2016), and knowledge graph building (Toledo-Ronen et al., 2016). The underlying methods concentrate on supervised learning. Among these, Bar-Haim et al. (2017) employ a support vector machine with multiple linguistic features, similar to those used in sentiment analysis. Iyyer et al. (2014) apply recursive neural networks, Augenstein et al. (2016) use a bidirectional LSTM, and Chen et al. (2018) implement a hybrid neural attention model. Unlike stance classification, the task we consider here does widely abstract from the topic on which stance is expressed.

4 Dataset and Experiments

In the shared task we carried out, we have devised two types of same side stance classification experiments: *within* a single topic and *across* two topics. The latter experiment type models the situation of a domain transfer and addresses the question of topic-agnostic classification. As topics we chose "gay marriage" and "abortion", and we sampled the respective argument datasets from the corpus underlying the argument search engine *args.me* (Wachsmuth et al., 2017b). The following subsections provide details about the dataset construction and the experiment setup.

[2] Except for the trivial case where $c \in P$.

[3] Given a thesis t we can consider its opposite as antithesis.

	Training			Test		
Class	**Gay**	**Abortion**	**Σ**	**Gay**	**Abortion**	**Σ**
Sameside	13 277	20 834	34 111	63	63	126
Diffside	9 786	20 006	29 792	63	63	126
Σ	23 063	40 840	63 903	126	126	252

Table 1: Number of argument pairs in the training sets and test sets of the within-topic experiments.

4.1 Dataset

Because of its size and the balanced stance distribution, the args.me corpus provides a rich source for our experiments. At the time of the shared task the corpus consisted of 387 606 arguments that collected from 59 637 debates; a detailed description can be found in (Ajjour et al., 2019).[4]

An argument in args.me is modeled as a conclusion along with a set of supporting premises. In addition, each premise is labeled with a stance, indicating whether it is "pro" or "con" the conclusion. The stances originate from the debates where the arguments are used in. Debates can be started from different viewpoints, for instance, a debate may discuss the viewpoint "abortion should be legalized" while another may discuss "abortion should be banned"). Therefore, the stance of an argument has to interpreted in relation to the arguments in the same debate. During the acquisition process of the data for the shared task we followed this constraint by ensuring that the arguments of an argument pair stem always from the same debate.

The count of debates that treat "abortion" and "gay marriage" is 1567 and 712 respectively. We filtered out those arguments whose premises are shorter than four words since they are often meta statements such as "I win" or "I accept". As a result, we kept 9426 arguments on abortion and 4480 arguments on gay marriage for the task.

4.2 Experiments

Starting from the arguments in a debate, we generated all possible argument pairs. An argument pair was labeled as "Sameside" if both arguments are either "pro" or "con" the viewpoint of the debate, otherwise the pair is labeled as "Diffside". Pairs with identical arguments were removed.

Within-Topic Experiments The within-topic experiments treat the two topics "Abortion" and "Gay marriage" independently of each other. The training sets each contain 67% of the argument pairs of one topic, which were randomly chosen. The test sets were formed from the remaining 33% for the respective topic. Among others, it was ensured that a label for an argument pair in the test set cannot be transitively deduced.[5] Note in this regard that the "same side" relation forms an equivalence relation. See Table 1 for the within-topic dataset statistics.

Class	Training: Abortion	Test: Gay
Sameside	31 195	3 028
Diffside	29 853	3 028
Σ	61 048	6 056

Table 2: Number of argument pairs in the training and test set of the cross-topics experiment.

Cross-Topics Experiment The cross-topics experiment provides a different topic for training from the one for testing. In particular, the training set contains argument pairs from the "abortion" debates only, while the test set contains argument pairs from "gay marriage" debates only. "Sameside" pairs and "Diffside" pairs are balanced. See Table 2 for the Cross-Topics dataset statistics.

5 Submitted Systems and Results

Overall, nine teams participated in the first shared task on same side stance classification. This section provides a brief overview of the systems that the teams submitted, along with their results.

Düsseldorf University The system submitted by Düsseldorf University relies on a Siamese network trained to predict the similarity of two arguments on top of a small BERT (Devlin et al., 2018). As the maximum token length for BERT is 512 tokens, a relevance selection component to rank sentences by relevance is integrated, cutting the ranked input

[4] The entire args.me corpus can be accessed here: https://webis.de/data.html#args-me

[5] With transitive deduction we mean: $\mathit{SameSide}(A_1, A_2) \wedge \mathit{SameSide}(A_3, A_2) \vdash \mathit{SameSide}(A_1, A_3)$

	Within-Topic									Cross-Topics		
	Gay			Abortion			All					
Team	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc	Pre	Rec	Acc
Trier University†	**0.90**	0.73	**0.83**	**0.79**	0.59	0.71	**0.85**	0.66	**0.77**	**0.73**	**0.72**	**0.73**
Leipzig University	0.80	0.78	0.79	0.78	0.68	**0.75**	0.79	0.73	**0.77**	0.72	**0.72**	0.72
IBM Research	0.73	0.63	0.70	0.64	0.54	0.62	0.69	0.59	0.66	0.62	0.49	0.60
TU Darmstadt	0.74	0.56	0.68	0.63	0.48	0.60	0.68	0.52	0.64	0.64	0.59	0.63
Düsseldorf University	0.76	0.35	0.62	0.65	0.32	0.57	0.70	0.33	0.60	0.72	0.53	0.66
Trier University†	0.64	0.25	0.64	0.67	0.22	0.56	0.65	0.24	0.56	0.70	0.11	0.53
LMU	0.53	**1.00**	0.55	0.53	**1.00**	0.55	0.53	**1.00**	0.55	0.67	0.53	0.63
MLU Halle‡	0.54	0.57	0.54	0.53	0.57	0.53	0.53	0.57	0.54	0.50	0.57	0.50
Paderborn University	0.55	0.17	0.52	0.62	0.21	0.54	0.59	0.19	0.53	0.60	0.38	0.56
University of Potsdam	0.46	0.54	0.45	0.56	0.62	0.56	0.51	0.58	0.51	0.51	0.52	0.51
MLU Halle‡	0.47	0.11	0.49	0.54	0.11	0.51	0.50	0.11	0.50	0.46	0.00	0.50

Table 3: The results of the submissions for the within-topic experiments and the cross-topics experiment in terms of precision (Pre), recall (Rec), and accuracy (Acc). For both Trier University† and MLU Halle‡, the best and the worst result are reported since they submitted multiple systems.

at 512 tokens. The system achieved an accuracy of 60% on the within-topic task and 66% across topics.

IBM Research The system submitted by IBM is based on a small vanilla BERT model and has been first fine-tuned to perform standard binary pro/con stance classification on data extracted from the IBM Debater project. On top of this model, another model is initialized and fine-tuned on the same side classification task. The system obtained results inverse to the ones of Düsseldorf University: 66% accuracy in the within-topic setting 60% in the cross-topics setting.

Leipzig University The system submitted by Leipzig University uses a pre-trained BERT model that is fine-tuned on the same side stance classification task. In addition, a binary classification layer with one output and cross entropy loss function is used instead of a multilabel classification layer. To embed an argument, the first 254 tokens of an argument are fed through the BERT model. Then, the last 254 tokens of an argument are embedded. The concatenation of both embeddings is fed into the classification layer. The system achieved an accuracy of 77% in the within-topic setting and 72% on the cross-topics setting.

LMU The system submitted by the Ludwig Maximilian University (LMU) relies on a vanilla pre-trained BERT base model that is fine-tuned to the shared task. The data is organized in a graph with one graph per topic. Nodes represent arguments, and edges are labeled with the confidence that the associated arguments agree with each other. This graph-based approach has the benefit that more training data can be generated by a transitive closure. Its accuracy was 55% in the within topic setting and 63% in the cross-topic setting.

MLU Halle The system submitted by the Martin-Luther-University (MLU) of Halle-Wittenberg consists of three system. The first system uses a tree-based learning algorithm as classifier using standard bag-of-words features. The second is a rule-based approach that reduces the task to sentiment classification relying on rules defined over lists of words with their polarity taken from a sentiment lexicon. The third is a re-implementation of the stance classification approach of Bar-Haim et al. (2017). The best system achieves an accuracy of 54% on the within-topics setting and 50% on the cross-topics setting.

Paderborn University The system submitted by Paderborn University relies on a Siamese Neural Network to map arguments to a new space where arguments with the same stance are closer to each other, and other arguments are less close. Arguments are represented by the contextual word embeddings provided by the Flair library (Akbik et al., 2018). A final sigmoid activation function produces the output used for same side stance classification. The system achieved an accuracy of 53% within topics and 56% across topics.

Trier University The system submitted by Trier University relies on a pre-trained BERT base model fine-tuned to the shared task. It was submitted with different configurations. The best yielded an accuracy of 77% in the within-topics setting and 73% on the cross-topics setting, the worst 56% and 53% respectively.

TU Darmstadt The system submitted by the TU Darmstadt relies on a multi-task deep network on the basis of the pre-trained large BERT model. The network is trained on a number of pro/con stance classification datasets in addition to the shared task dataset. The system achieved an accuracy of 64% in the within-topics setting and 63% in the cross-topic setting.

University of Potsdam The system submitted by the University of Potsdam relies on bidirectional LSTMs to encode the arguments. The embeddings of both arguments are concatenated, multiplied in an element-wise fashion, substracted, and fed into a two-layer MLP as a classification layer. The system achieved 51% accuracy both within and across topics.

6 Discussion and Outlook

The results of the shared task license a number of interesting conclusions. First of all, the results have validated our hypothesis that a topic-agnostic approach to same side stance classification is feasible. This is clearly conveyed by the fact that the within-topic and the cross-topics setting seem to be of a similar complexity. Also, the differences in accuracy on both tasks are less than 5–6% points, additionally corroborating the hypothesis.

A second conclusion is that the effectiveness of most systems clearly improves over a random baseline, showing that the task is generally feasible. At the same time, however, the results show that there is potential for improvement.

As for other tasks in the field of argumentation, such as the Argument Reasoning Comprehension Task, ARCT (Habernal et al., 2018), encoder-based models seem to reach top results. In fact, all of the top-5 performing systems on our task (Trier University, Leipzig University, IBM Research, TU Darmstadt, and Düsseldorf University) rely on a BERT model. They differ mainly in the way the input is encoded. As the length of input arguments exceeds the maximum input length for BERT models, the participants explored and proposed different approaches, such as encoding the beginning and end of the arguments separately and then concatenating these encodings or implementing a relevance ranking system to encode only the most relevant sentences of the argument. In any case, the encoding strategy seems to have a clear impact on the results and thus deserves further investigation.

For related tasks, e.g. the ARCT, it has been found recently that encoder-based models seem to pick up surface cues and artifacts of the dataset and that they are not really able to learn a model that shows deeper understanding of how arguments work. It is up to further investigation whether also the same side stance classification task bears the potential for such artifacts that can be picked up by system. It would be interesting to investigate which task the encoder-based models actually learn to solve.

References

Aseel Addawood, Jodi Schneider, and Masooda Bashir. 2017. Stance classification of Twitter debates: The encryption debate as a use case. In *8th International Conference on Social Media and Society*, ACM International Conference Proceeding Series. Association for Computing Machinery.

Yamen Ajjour, Henning Wachsmuth, Johannes Kiesel, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Data Acquisition for Argument Search: The args.me corpus. In *42nd German Conference on Artificial Intelligence (KI 2019)*. Springer.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885. Association for Computational Linguistics.

Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. 2018. Integrating Stance Detection and Fact Checking in a Unified Corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27. 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. Association for Computational Linguistics.

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. 2017. From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 84–89. Association for Computational Linguistics.

Di Chen, Jiachen Du, Lidong Bing, and Ruifeng Xu. 2018. Hybrid Neural Attention for Agreement/Disagreement Inference in Online Debates. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 665–670. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. SemEval-2018 task 12: The argument reasoning comprehension task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 763–772, New Orleans, Louisiana. 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. Asian Federation of Natural Language Processing.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. 2014. Political Ideology Detection Using Recursive Neural Networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122. Association for Computational Linguistics.

G. Levow, V. Freeman, A. Hrynkevich, M. Ostendorf, R. Wright, J. Chan, Y. Luan, and T. Tran. 2014. Recognition of stance strength and polarity in spontaneous speech. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 236–241.

Stefano Menini, Federico Nanni, Simone Paolo Ponzetto, and Sara Tonelli. 2017. Topic-based agreement and disagreement in us electoral manifestos. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2938–2944. Association for Computational Linguistics.

Saif M. Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and Sentiment in Tweets. *ACM Trans. Internet Technol.*, 17(3).

Moin Nadeem, Wei Fang, Brian Xu, Mitra Mohtarami, and James Glass. 2019. FAKTA: An Automatic End-to-End Fact Checking System. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 78–83. Association for Computational Linguistics.

Pavithra Rajendran, Danushka Bollegala, and Simon Parsons. 2016. Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 31–39. Association for Computational Linguistics.

Sarvesh Ranade, Rajeev Sangal, and Radhika Mamidi. 2013. Stance Classification in Online Debates by Recognizing Users' Intentions. In *Proceedings of the SIGDIAL 2013 Conference*, pages 61–69. Association for Computational Linguistics.

Swapna Somasundaran and Janyce Wiebe. 2009. Recognizing Stances in Online Debates. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 226–234. Association for Computational Linguistics.

Orith Toledo-Ronen, Roy Bar-Haim, and Noam Slonim. 2016. Expert Stance Graphs for Computational Argumentation. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 119–123. Association for Computational Linguistics.

David Vilares and Yulan He. 2017. Detecting Perspectives in Political Debates. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1573–1582. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017a. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187. Association for Computational Linguistics.

Henning Wachsmuth, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017b. Building an argument search engine for the web. In *Proceedings of the 4th Workshop on Argument Mining*, pages 49–59. Association for Computational Linguistics.