Leveraging Topic Relatedness for Argument Persuasion

Xinran Zhao[1], **Esin Durmus**[2], **Hongming Zhang**[1], **and Claire Cardie**[2]
[1]The Hong Kong University of Science and Technology
[2]Cornell University
xzhaoar@connect.ust.hk, ed459@cornell.edu
hzhangal@cse.ust.hk, cardie@cs.cornell.edu

Abstract

Argumentation exposes individuals to conflicting viewpoints and can help them make more informed decisions based on the pros and cons of a particular issue. While recent studies of argumentation in Natural Language Processing have mainly focused on understanding the effect of various factors of persuasion (i.e. the source, audience, and language style), the impact of exploiting the relationships among controversial topics when predicting argument persuasiveness remains under-explored. In this paper, we model the relatedness among controversial topics utilizing an embedding-based method based on individuals' stances on the topics. We then leverage these *topic embedding* features and incorporate *topic semantics* features extracted from the arguments along with the previously studied factors of persuasion. We show that incorporating both types of topic relatedness features explicitly leads to significant improvement in predicting persuasiveness and also helps enhance generalization to rare topics, in a few-shot setting.

1 Introduction

Emergence of social media and online argumentative forums provide users with a platform to gain information, express, and form opinions on a diverse set of controversial topics (i.e. issues). The increasing importance of these online platforms has motivated NLP researchers to use these platforms as one of the main domains to study the important factors of persuasion. In particular, prior work has shown that characteristics of the speaker (i.e. source), prior beliefs of the audience (Lukin et al., 2017; Durmus and Cardie, 2018), and language style (Feng and Hirst, 2011; Tan et al., 2016) are important factors in determining persuasiveness of the arguments in online argumentation platforms. Although there has been evidence in previous studies of Social Sciences that people's perceptions on a particular controversial topic may be related to their perceptions on other controversial topics (Judd and Krosnick, 1989; Sapra, 2012), the impact of exploiting this relationship among controversial topics are under-explored in NLP studies of persuasion. In this paper, we explicitly study the effect of incorporating topic relatedness among controversial topics in predicting argument persuasiveness.

To study the impact of involving topic relatedness in argument persuasion, we define two types of features: (1) *topic embedding features* and (2) *topic semantics features*. Prior work has shown that topic is an important factor (Das et al., 2016) to determine whether an emotional vs. a logical argument will be received positively by the audience. We hypothesize that encoding underlying relationship among topics with topic embedding features will be helpful in predicting persuasion since similar strategies may be effective for related topics. We further define topic semantics features to encode how focused vs. divergent each of the arguments made by the debaters is given the discussion topic, similar to Zhang et al. (2016).

We first develop an embedding-based technique inspired by (Barkan and Koenigstein, 2016) to determine the relationship among controversial topics. This methodology leverages users' stances on the topics to determine the relationship among them. We then incorporate the *topic embedding features* and *topic semantics features*, along with the previously studied factors of persuasion. We find that incorporating the topic relatedness features help improve state-of-the-art results in persuasion prediction. Moreover, we conduct experiments in a few-shot setting and show that these features help models achieve significantly better generalization performance for the rare topics.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4401–4407
August 1–6, 2021.

2 Dataset

We use DDO (Durmus and Cardie, 2019) for our study. DDO includes 67,315 debates from 23 different categories, 36,294 users with their background information (e.g. political ideology, and religious ideology), and 198,759 votes from the users when they are as readers of these debates. For each debate, two debaters with different viewpoints express their opinions on a controversial topic in rounds. After the debate, voters evaluate the debaters with respect to various criteria and they share whether any of the debaters changed their stance on the topic. Users also have an opportunity to share their demographic and ideological information such as gender, ethnicity, income level, education level, political ideology, and religious ideology. They also share their stance on a pre-defined list of controversial topics (i.e. **BIG ISSUES**, such as *Abortion*, *Gay Marriage*, and *Global Warming* etc.)[1] that we use to determine semantic relatedness among these controversial topics.

3 Methodology

3.1 Topic Related Embeddings

Topic Embeddings. To capture the underlying relatedness between the debate topics, we learn the embedding for each controversial topics with a method inspired by Barkan and Koenigstein (2016). We hypothesize that the users' opinions (i.e. whether they are SUPPORTING or OPPOSING) on similar topics are related. We treat a set of controversial issues with the same stance from a user as a set of words appearing in the same context and use adapted Skip-gram algorithm proposed by (Mikolov et al., 2013). The embedding vectors are optimized by predicting the topic similarity that is defined as the probability that a pair of topics appearing in the same group with respect to users' opinions. We then can use these vectors to compare the similarity of each pair of the big issues.

Table 1 shows the most similar controversial issues for each of the given issues, where the similarity is calculated by the cosine similarity between the embedding vectors. We observe that some of these associations can be more related to relatively intuitive topic similarity between these issues (i.e., *Capitalism* and *Flax Tax*, and *Environmental Protection* and *Global Warming Exists*). However, in some cases, the similarity between controversial issues may have more complicated motivations such as users' underlying ideologies (i.e., as in the case of *Environmental Protection* and *Abortion* which can be justified by the study conducted by Sapra (2012)). Therefore, this method may help identify relationships among controversial issues that are not as intuitive to come up with.

Issue	Top similar issues: Similarity
Torture	Iran-Iraq War: 0.90 Electoral College: 0.85 Border Fence: 0.85 Military Intervention: 0.74 Racial Profiling: 0.71
Welfare	Minimum Wage: 0.97 Occupy Movement: 0.86 Medicaid & Medicare: 0.85 Labor Union: 0.84 National Health Care: 0.84
Capitalism	Flat Tax: 0.80 Social Programs: 0.73 Electoral College: 0.69 Affirmative Action: 0.69 Stimulus Spending: 0.66
Environment Protection	Medical Marijuana: 0.93 Abortion: 0.85 Global Warming Exists: 0.85 Drug Legalization: 0.82 United Nations: 0.82

Table 1: Most similar issues for *Torture*, *Welfare*, *Capitalism*, and *Environment Protection* issues.

Topic-Centric Attribute Embeddings. There is evidence showing there is a strong association between the users' demographics and their stances towards controversial topics (Sapra, 2012; Tedin et al., 1977). Although prior studies of persuasion has studied the effect of users' attributes on persuasiveness (Lukin et al., 2017; Durmus and Cardie, 2018, 2019), they did not explicitly model the relationship between the users' attributes and their stance towards the controversial topics. To explicitly model this relationship, we create embeddings for users' attributes in a similar way as the issue embeddings introduced in Section 3.1. We optimize the embeddings with the probability of a certain attribute appearing with a given issue-stance pair for each particular user.

Table 2 shows top similar issue-stance combinations for users with given political and religious ideology categories. Similarly, we calculate similarity by the cosine similarity between the embedding vectors. This approach reveals certain associations between users attributes (e.g. Political Ideology) and their stances (e.g. PRO vs. CON) towards certain controversial issues (e.g. *Gay Marriage*). For

[1] See https://www.debate.org/big-issues/ for the full list of the Big Issues.

Ideology	Issue-stance: Similarity
Conservative	Gay Marriage-CON: 0.94 Abortion-CON: 0.93 Global Warming Exists-CON: 0.92 Euthanasia-CON: 0.92 Border Fence-PRO: 0.92 Death Penalty-PRO: 0.89
Liberal	Gun Rights-CON: 0.87 Environmental Protection-PRO: 0.83 Medicaid & Medicare-PRO: 0.83 Affirmative Action-PRO: 0.83 Global Warming Exists-PRO: 0.82 Barack Obama-PRO: 0.81

Table 2: Most similar issue-stance combinations for the given categories of *Political Ideology*.

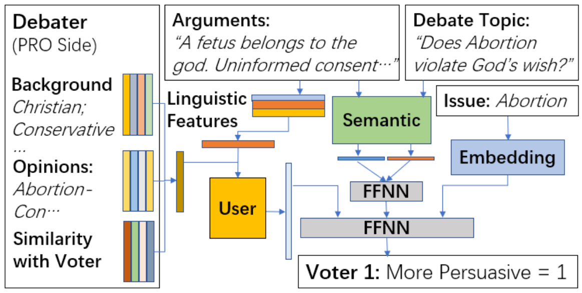

Figure 1: Overall model structure. *User, Semantic, Embedding* blocks denote the encoders for user information, argument semantics, and topic embedding. *FFNN* is a multi-layer feed-forward neural network.

example, we find that users with *Conservative* vs. *Liberal* political ideologies has different views on *Global Warming* (i.e. CON vs. PRO respectively) issue looking at the similarity of corresponding embeddings.

3.2 Predicting Persuasiveness

We aim to predict which debater, either the PRO or CON side, expresses more persuasive arguments in the debate (i.e., received more votes for the *"Made more convincing arguments"* criterion.).

Debate Topic Representation. We collect debates related to list of issues in BIG ISSUES by using the words of these issues as keywords (e.g. debate topic *"Abortion should be illegal if it pregnancy does not endanger the mother's life and she is adult."* is related to BIG ISSUE "ABORTION"). The dataset includes 2,893 debates and 10,441 votes. We represent the debate topic with the **topic embedding** of the corresponding BIG ISSUE as introduced in Section 3.1. We further encode the text of the debate topic with a fine-tuned BERT (Devlin et al., 2019) taking average embedding of all the tokens to get the representation of the **topic semantics**. We concatenate these two types of embeddings to get the final representation of the debate topic.

Representing User Information. Previous work shows that both characteristics of the debaters and the audience and the linguistic features of the debate arguments are important factors in persuasion studies (Lukin et al., 2017; Durmus and Cardie, 2018, 2019; Longpre et al., 2019). Similar to prior work, to encode the background information, we first represent the user background with one-hot representation (**ONE-HOT**) to capture the users' selections on the categories (e.g., gender, political ideology, religious ideology, and etc.) or the opinion similarity with the voters. However, this representation can be very sparse and not relevant to the topic information. Therefore, we also experiment with the topic-centric embedding-based method (**ATT-EMB**) proposed in Section 3.1. We compute the background similarity as the cosine similarity of the representation vectors for the users (i.e., debaters and voters).

Linguistic Features. Consistent with the prior work (Durmus and Cardie, 2018), to encode the arguments in the debates, we extract linguistic features including the information about the style (i.e., length, links), sentiment polarity, subjectivity (Wilson et al., 2005), and argument lexicon features (Somasundaran and Wiebe, 2010)) etc. Similar to the topic semantics introduced in Section 3.2, we also represent the semantics of the arguments with the same fine-tuned BERT.

Proposed Model. We employ a model that contains separate encoders to represent the debater characteristics, arguments, and topic-related features, as shown in Figure 1. The model encodes the debater's background information and opinions towards the BIG ISSUES, and combines the linguistic features extracted from arguments to represent the users. For the text in the debate, the model consists a siamese network structure (*Semantic* Block + *FFNN* block) (Reimers and Gurevych, 2019) to encode the relation between arguments and debate topics. Then, the model extracts the issue representation from the pretrained issue embedding introduced in Section 3.1. Finally, the user representation, together with the representation for the argument semantics and topic embedding, is passed through a multi-layer feed-forward neural network to predict the voter's perception on the persuasiveness.

Model	F1 (%)
Majority	33.25
Bi-LSTM+Glove	33.41
SBERT	50.05
ONE-HOT+Linguistic+Topic SVM	41.38
ATT-EMB+Linguistic SVM	53.51
ATT-EMB+Linguistic+Topic SVM	59.04
Ours with ONE-HOT	57.03
Ours W/O ARGUMENT	63.81
Ours W/O ATT-EMB	51.33
Ours W/O TOPIC	64.21
Ours	**65.62**

Table 3: Macro F1 scores. **Ours** denotes the model explained in Figure 1. We split the collected debates to train (70%), validation (15%), and test (15%) sets.

4 Baselines

(1) Majority Baseline: We assign the label that appears the most in training set to be the prediction for all test instances. **(2) SVM**: We concatenate debate topic representation, user and linguistic features and classify with SVM with RBF kernel. **(3) Bi-LSTM** (Hochreiter and Schmidhuber, 1997): Following (Durmus et al., 2019; Li et al., 2020), We encode the arguments and topic representation with bidirectional LSTM encoders and use a *FFNN* as classification head[2]. **(4) SBERT** (Reimers and Gurevych, 2019): Sentence-BERT (SBERT) has demonstrated that fine-tuning BERT in a Siamese/Triplet network architecture achieves the state-of-the-art results over various sentence-level classification benchmarks. We use a Siamese network to encode the sentence representation from arguments and debate topics.

5 Results

Model Ablations. The full proposed model contains three parts: user-based features (which can be represented by one-hot vectors or topic-centric user embeddings), argument-based features, and topic-based features. To understand the contribution of each component to prediction performance, we conduct ablation studies for the settings where (1) Using ATT-EMB (i.e. topic-centric user embeddings proposed in this work) vs. ONE-HOT representation to encode the user background, (2) Removing the representation for user background (W/O USER), (3) Removing the argument features and (W/O ARGUMENT) (4) Removing the topic-related features (W/O TOPIC)[3].

[2]We choose the hidden embedding dimension to be 200 and use Glove (Pennington et al., 2014) as the pre-trained word embeddings.

Model	Frequent	Rare	Δ
Ours with ATT-EMB	63.31	60.12	**-3.19**
Ours w/o ATT-EMB	56.25	52.65	-3.60
Ours w/o TOPIC-EMB	64.20	58.37	-5.83

Table 4: Results (%macro F1) for the few-shot setting experiment. Δ denotes the performance difference when testing the Rare debates.

Table 3 demonstrates the macro F1 scores for the baselines and the ablations for our model. We observe that our model outperforms the feature-based baselines significantly. For both the SVM model and the deep models, we observe a large performance drop when we use one-hot embedding representation features instead of topic-centric user embeddings. This shows that encoding the relationship between the user background and their opinions on the topics explicitly improves the prediction performance significantly. The experiments on Bi-LSTM and SBERT show that although large-scale pretrained language representation model helps achieve better performance than the baseline, encoding the semantics with deep neural network encoders alone is not as effective as our proposed method. Comparing to the baselines, our proposed method that utilizes the information from different components (i.e. users, language and topic) is more effective. Ablation study shows that components that encode the topic semantics (i.e. ATT-EMB and TOPIC) play an important role to achieve the best performance.

Few-shot Setting. We study whether the topic embeddings also enhance the generalizability across different issues. We split the debates in the test set into frequent and rare categories looking at how often debates with the same topic appear in the training set (more than 200 vs. less than 20). Table 4 shows the results comparing to the baselines for corresponding 324 frequent and 131 rare debates. We see that the gap between the prediction performance is significantly more when we remove the attribute embedding and the topic embeddings, which indicates that the topic-related embeddings benefit the knowledge transferring among debates with different topics.

6 Related Work

Persuasion Studies. Understanding the characteristics of persuasive language has been an im-

[3]Implementation details are described in Appendix A.1.

portant area of study in the Sociology, Psychology (Kelman, 1961; Burgoon et al., 1975; Chaiken, 1987; Tykocinskl et al., 1994; Chambliss and Garner, 1996) and NLP communities (Hasan and Ng, 2014; Habernal and Gurevych, 2016a,b; Fang et al., 2016; Al-Khatib et al., 2017; Wang et al., 2019). The emergence of social media and argumentative forums has further attracted researchers to study the dynamics of persuasion on these platforms, including Twitter (Tan et al., 2014), ChangeMyView community on Reddit (Tan et al., 2016; Hidey et al., 2017) and DDO (Durmus and Cardie, 2019). In this work, we use DDO since it includes a wide-range of user information including users' opinions on various controversial topics as well as well-structured debates with audience votes.

Topic Aware Argument Mining. Farra et al. (2015) studied the effect of topic relevancy or consistency on essay scoring. Bosc et al. (2016) proposes a dataset of Social Media data with coarse topic labels extracted from the hashtags (e.g., #AppleWatch). Zeng et al. (2020) designed a model to encode the latent topics of argumentative conversations. Unlike the previous work, our work studies the effect of argument topic for structured debates explicitly for predicting persuasion.

7 Conclusion

In this paper, we study the impact of topic-relatedness in debate persuasion and find that involving the semantics and features of the debate topics will achieve the best performing model. Moreover, we find that using pretrained embeddings that jointly encode the issues related to the topics and people's characteristics will largely benefit the training process and generalizability. Finally, we find that focusing on the debate topic in making arguments can be an effective strategy in online debates.

Acknowledgements

We thank all the anonymous reviewers for their helpful feedback. This work was supported in part by NSF grant IIS-1815455. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

References

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1351–1357, Copenhagen, Denmark. Association for Computational Linguistics.

Oren Barkan and Noam Koenigstein. 2016. ITEM2VEC: neural item embedding for collaborative filtering. In *Proceedings of MLSP 2016.*

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).

Michael Burgoon, Stephen B. Jones, and Diane Stewart. 1975. Towards a message-centered theory of persuasion: Three empirical investigation of language intensity. *Human Communication Research*, 1(3):240–256.

Shelly Chaiken. 1987. The heuristic model of persuasion. In *Social influence: the ontario symposium*, volume 5, pages 3–39.

Marilyn J. Chambliss and Ruth Garner. 1996. Do adults change their minds after reading persuasive text? *Written Communication*, 13(3):291–313.

Abhimanyu Das, Sreenivas Gollapudi, Emre Kıcıman, and Onur Varol. 2016. Information dissemination in heterogeneous-intent networks. In *Proceedings of the 8th ACM Conference on Web Science*, pages 259–268.

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-HLT 2019.*

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. 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 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2019. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607, Florence, Italy. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019. The role of pragmatic and discourse context in determining argument impact. 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 5668–5678, Hong Kong, China. Association for Computational Linguistics.

Hao Fang, Hao Cheng, and Mari Ostendorf. 2016. Learning latent local conversation modes for predicting comment endorsement in online discussions. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 55–64, Austin, TX, USA. Association for Computational Linguistics.

Noura Farra, Swapna Somasundaran, and Jill Burstein. 2015. Scoring persuasive essays using opinions and their targets. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 64–74, Denver, Colorado. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 987–996, Portland, Oregon, USA. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathy McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21, Copenhagen, Denmark. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Charles M Judd and Jon A Krosnick. 1989. The structural bases of consistency among political attitudes: Effects of political expertise and attitude importance. *Attitude structure and function*.

Herbert C. Kelman. 1961. Processes of Opinion Change. *Public Opinion Quarterly*, 25(1):57–78.

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8905–8912, Online. Association for Computational Linguistics.

Liane Longpre, Esin Durmus, and Claire Cardie. 2019. Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th Workshop on Argument Mining*.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS 2013*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543.

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.

Sonalini Sapra. 2012. Feminist perspectives on the environment. In *Oxford Research Encyclopedia of International Studies*.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, CAAGET '10, pages 116–124, Stroudsburg, PA, USA. Association for Computational Linguistics.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 175–185, Baltimore, Maryland. Association for Computational Linguistics.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Kent L. Tedin, David W. Brady, and Arnold Vedlitz. 1977. Sex differences in political attitudes and behavior: The case for situational factors. *The Journal of Politics*.

Orit Tykocinskl, E. Tory Higgins, and Shelly Chaiken. 1994. Message framing, self-discrepancies, and yielding to persuasive messages: The motivational significance of psychological situations. *Personality and Social Psychology Bulletin*, 20(1):107–115.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis.

Jichuan Zeng, Jing Li, Yulan He, Cuiyun Gao, Michael Lyu, and Irwin King. 2020. What changed your mind: The roles of dynamic topics and discourse in argumentation process. In *Proceedings of The Web Conference 2020*, WWW '20, page 1502–1513, New York, NY, USA. Association for Computing Machinery.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

A Appendix

A.1 Implementation Details

We initialize the semantic encoder for arguments and debate topics with the BERT-base model with 110M parameters. We pad the input sentences with BERT start and end symbols (i.e., [CLS] and [SEP]). For each round of the debate, we take at most three sentences (as (Li et al., 2020)) to take average and represent the arguments. The hidden size for the Bi-LSTM layers is 300. The size for the hidden states for *FFNN* blocks are 256. During training, we use cross entropy as the loss function and stochastic gradient decent (SGD) as the optimizer. We initialize all parameters randomly and train all the model with 10 epochs. The best performing models on the validation set are evaluated on the test set.