A Corpus for Modeling User and Language Effects in Argumentation on Online Debating

Esin Durmus
Cornell University
ed459@cornell.edu

Claire Cardie
Cornell University
cardie@cs.cornell.edu

Abstract

Existing argumentation datasets have succeeded in allowing researchers to develop computational methods for analyzing the content, structure and linguistic features of argumentative text. They have been much less successful in fostering studies of the effect of "user" traits — characteristics and beliefs of the participants — on the debate/argument outcome as this type of user information is generally not available. This paper presents a dataset of 78,376 debates generated over a 10-year period along with surprisingly comprehensive participant profiles. We also complete an example study using the dataset to analyze the effect of selected user traits on the debate outcome in comparison to the linguistic features typically employed in studies of this kind.

1 Introduction

Previous work from Natural Language Processing (NLP) and Computational Social Science (CSS) that studies argumentative text and its persuasive effects has mainly focused on identifying the content and structure of an argument (e.g. Feng and Hirst (2011)) and the linguistic features that are indicative of effective argumentation strategies (e.g. Tan et al. (2016)). The effectiveness of an argument, however, cannot be determined solely by its textual content; rather, it is important to consider characteristics of the reader, listener or participants in the debate or discussion. Does the reader already agree with the argument's stance? Is she predisposed to changing her mind on the particular topic of the debate? Is the style of the argument appropriate for the individual? To date, existing argumentation datasets have permitted only limited assessment of such "user" traits because information on the background of users is generally unavailable. In this paper, we present a dataset of 78,376 debates from October of 2007 until November of 2017 drawn from *debate.org* along with quite comprehensive user profile information — for debate participants as well as users voting on the debate quality and outcome. Background information on users includes demographics (e.g. education, income, religion) and stance on a variety of controversial debate topics as well as a record of user activity on the debate platform (e.g. debates won and lost). We view this new dataset as a resource that affords the NLP and CSS communities the opportunity to understand the effect of audience characteristics on the efficacy of different debating and persuasion strategies as well as to model changes in user's opinions and activities on a debate platform over time. (To date, part of our *debate.org* dataset has been used in one such study to understand the effect of prior beliefs in persuasion[1] (Durmus and Cardie, 2018). Here, we focus on the properties of the dataset itself and study a different task.)

In the next section, we describe the dataset in the context of existing argumentation datasets. We then provide statistics on key aspects of the collected debates and user profiles (Section 3). Section 4 reports a study in which we investigate the predictive effect of selected user traits (namely, the debaters' and audience's experience, prior debate success, social interactions, and demographic information) vs. standard linguistic features. Experimental results show that features of the user traits are significantly more predictive of a debater's success than the linguistic features that are shown to be predictive of debater success by the previous work (Zhang et al., 2016). This suggests that user traits are important to take into account in studying success in online debating.

[1] That study is distinct from those presented here. See Section 4 for details.

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 602–607
Florence, Italy, July 28 - August 2, 2019.

The dataset will be made publicly available[2].

2 Related Work and Datasets

There has been a tremendous amount of research effort to understand the important linguistic features for identifying argument structure and determining effective argumentation strategies in monologic text (Mochales and Moens, 2011; Feng and Hirst, 2011; Stab and Gurevych, 2014; Guerini et al., 2015). For example, Habernal and Gurevych (2016) has experimented with different machine learning models to predict which of two arguments is more convincing. To understand what kind of persuasive strategies are effective, Hidey et al. (2017) has further annotated different modes of persuasion (ethos, logos, pathos) and looked at which combinations appear most often in more persuasive arguments.

Understanding argumentation strategies in conversations and the effect of interplay between the language of the participants has also been an important avenue of research. Tan et al. (2016), for example, has examined the effectiveness of arguments on ChangeMyView[3], a debate forum website in which people invite others to challenge their opinions. They found that the interplay between the language of the opinion holder and that of the counterargument provides highly predictive cues of persuasiveness. Zhang et al. (2016) has examined the effect of conversational style in Oxford-style debates and found that the side that can best adapt in response to opponents' discussion points over the course of the debate is more likely to be more persuasive. Although research on computational argumentation has mainly focused on identifying important linguistic features of the text, there is also evidence that it is important to model the debaters themselves and the people who are judging the quality of the arguments: multiple studies show that people perceive arguments from different perspectives depending on their backgrounds and experiences (Correll et al., 2004; Hullett, 2005; Petty et al., 1981; Lord et al., 1979; Vallone et al., 1985; Chambliss and Garner, 1996). As a result, we introduce data from a social media debate site that also includes substantial information about its users and their activity and interaction on the website. This is in contrast to the datasets commonly employed in studies of argument strategies (Johnson and Goldman, 2009; Walker et al., 2012; Zhang et al., 2016; Wang et al., 2017; Cano-Basave and He, 2016; Al Khatib et al., 2016). Lukin et al. (2017) is the closest work to ours as it studies the effect of OCEAN personality traits (Roccas et al., 2002; T. Norman, 1963) of the audience on how they perceive the persuasiveness of monologic arguments. Note that, in our dataset, we do not have information about users' personality traits; however, we have extensive information about their demographics, social interactions, beliefs and language use.

3 Dataset[4]

Debates. The dataset includes $78,376$ debates from 23 different topic categories including *Politics*, *Religion*, *Technology*, *Movies*, *Music*, *Places-Travel*. Each debate consists of different rounds in which opposing sides provide their arguments. An example debate along with the user information for PRO and CON debaters and corresponding comments and votes are shown in Figure 1. The majority of debates have three or more rounds; *Politics*, *Religion*, and *Society* are the most common debate categories. Each debate includes comments as well as the votes provided by other users in the community. We collected all the comments and votes for each debate with 606,102 comments and 199,210 votes in total. Voters evaluate each debater along diverse set of criteria such as convincingness, conduct during the debate, reliability of resources cited, spelling and grammar. With this fine-grained evaluation scheme, we can study the quality of arguments from different perspectives.

User Information. The dataset also includes self-identified information for $45,348$ users participating in the debates or voting for the debates: demographic information such as age, gender, education, ethnicity; prior belief and personal information such as political, religious ideology, income, occupation and the user's stance on a set of 48 controversial topics chosen by the website. The controversial debate topics[5] include ABORTION, DEATH PENALTY, GAY MARRIAGE, and AFFIRMATIVE ACTION. Information about user's activity is also provided and includes their debates, votes, comments, opinion questions they ask, poll

[2]Link to the dataset: http://www.cs.cornell.edu/ esindurmus/.

[3]https://www.reddit.com/r/changemyview/.

[4]Data is crawled in accordance to the terms and conditions of the website.

[5]Full list of topics: https://www.debate.org/big-issues/.

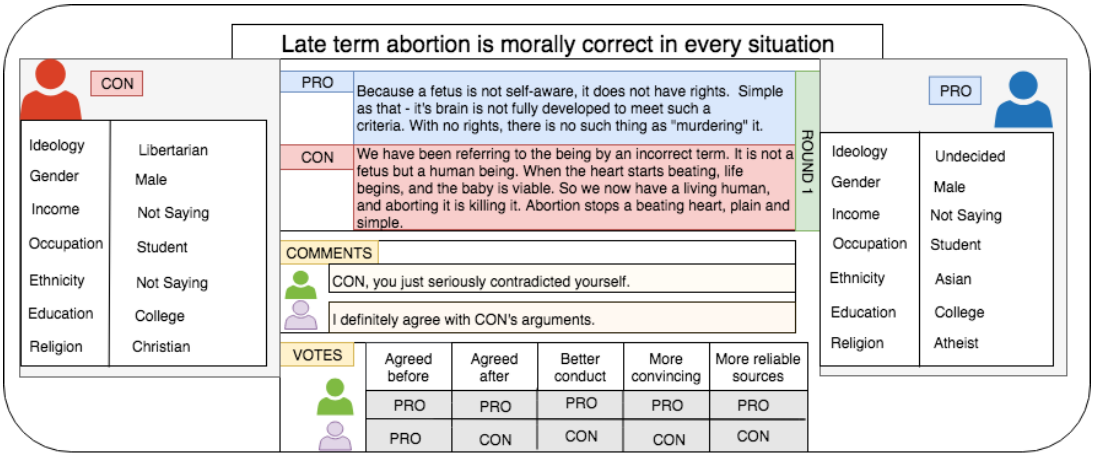

Figure 1: Example debate along with the user profile information for PRO and CON debaters and the corresponding comments and votes. The full information for this debate can be found at https://www.debate.org/debates/Late-term-abortion-is-morally-correct-in-every-situation/1/.

votes they participated in, overall success in winning debates as well as their social network information.

4 Task: What makes a debater successful?

To understand the effect of user characteristics vs. language features, and staying consistent with majority of previous work, we conduct the task of predicting the winner of a debate by looking at accumulated scores from the voters. We model this as a binary classification task and experiment with a logistic regression model, optimizing the regularizer ($\ell 1$ or $\ell 2$) and the regularization parameter C (between 10^{-5} and 10^{5}) with 3-fold cross validation.

4.1 Data preprocessing

Controlling for the debate text. We eliminate debates where a debater forfeits before the debate ends. From the remaining debates, we keep only the ones with three or more rounds with at least 20 sentences by each debater in each round to be able to study the important linguistic features [6].

Determining the winner. For this particular dataset, the winning debater is determined by the votes of other users on different aspects of the arguments as outlined in Section 3, and the debaters are scored accordingly[7]. We determine the winner by the total number of points the debaters get from the voters. We consider the debates with at least 5 voters and remove the debates resulting in a tie.

4.2 Features

Experience and Success Prior. We define the **experience** of a user during a debate d_t at time t as the total number of debates participated as a debater by the user before time t. The **success prior** is defined as the ratio of the number of debates the user won before time t to the total number of debates before time t.

Similarity with audience's user profile. We encode the similarity of each of the debaters and the voters by comparing each debaters' opinions on controversial topics, religious ideology, genders, political ideology, ethnicity and education level to same of the audience. We include the features that encode the similarity by counting number of voters having the same values as each of the debaters for each of these characteristics. We also include features that corresponds to cosine distance between the vectors of each debater and each voter where the user vector is one-hot representation for each user characteristic.

Social Network. We extract features that represent the debaters' social interactions before a particular debate by creating the network for their commenting and voting activity before that debate. We then computed the degree, centrality, hub and authority scores from these graphs and include them as features in our model.

Linguistic features of the debate. We perform ablation analysis with various linguistic features shown to be effective in determining

[6] After all the eliminations, we have 1635 debates in our dataset.

[7] Having better conduct: 1 point, having better spelling and grammar: 1 point, making more convincing arguments: 3 points, using the most reliable sources: 2 points.

	Accuracy
Majority baseline	57.23
User features	
Debate experience	63.54
Success prior	65.78
Overall similarity with audience	62.52
Social network features	62.93
All user features	**68.43**
Linguistic features	
Length	58.45
Flow features	58.66
All linguistic features	**60.28**
User+Linguistic Features	**71.35**

Table 1: Ablation tests for the features.

persuasive arguments including argument lexicon features (Somasundaran et al., 2007), politeness marks (Danescu-Niculescu-Mizil et al., 2013), sentiment, connotation (Feng and Hirst, 2011), subjectivity (Wilson et al., 2005), modal verbs, evidence (marks of showing evidence including words and phrases like "evidence" ,"show", "according to", links, and numbers), hedge words (Tan and Lee, 2016), positive words, negative words, swear words, personal pronouns, type-token ratio, tf-idf, and punctuation. To get a text representation for the debate, we concatenated all the turns of each of the participants, extracted features for each and finally concatenated the feature representation of each participant's text.
We also experimented with *conversational flow features* shown to be effective in determining the successful debaters by (Zhang et al., 2016) to track how ideas flow between debaters throughout a debate. Consistent with (Zhang et al., 2016), to extract these features, we determine the *talking points* that are most discriminating words for each side from the first round of the debate applying the method introduced by (Monroe et al.) which estimates the divergence between the two sides word-usage.

4.3 Results and Analysis

Table 1 shows the results for the user and linguistic features. We find that combination of the debater experience, debater success prior, audience similarity features and debaters' social network features performs significantly better[8] than the majority baseline and linguistic features achieving the best accuracy (68.43%). We observe that experience and social interactions are positively correlated with success. It suggests that as debaters spend more time on the platform, they probably learn strategies and adjust to the norms of the platform and this helps them to be more successful. We also find that success prior is positively correlated with success in a particular debate. In general, the debaters who win the majority of the debates when first join the platform, tend to be successful in debating through their lifetime. This may imply that some users may already are good at debating or develop strategies to win the debates when they first join to the platform. Moreover, we find that similarity with audience is positively correlated with success which shows that accounting for the characteristics of the audience is important in persuasion studies (Lukin et al., 2017).

Although the linguistic features perform better than the majority baseline, they are not able to achieve as high performance as the features encoding debater and audience characteristics. This suggest that success in online debating may be more related to the users' characteristics and social interactions than the linguistic characteristics of the debates. We find that use of argument lexicon features and subjectivity are the most important features and positively correlated with success whereas *conversational flow features* do not perform significantly better than length. This may be because debates in social media are much more informal compare to Oxford style debates and therefore, in the first round, the debaters may not necessarily present an overview of their arguments (*talking points*) they make through the debate.

We observe that (44%) of the mistakes made by the model with user features are classified correctly by the linguistic model. This motivated us to combine the user features with linguistic features which gives the best overall performance (71.35%). This suggests that user aspects and linguistic characteristics are both important components to consider in persuasion studies. We believe that these aspects complement each other and it is crucial to account for them to understand the actual effect of each of these components. For future work, it may be interesting to understand the role of these components in persuasion further and think about the best ways to combine the information from these two components to better represent

[8] We measure the significance performing t-test.

a user.

Acknowledgments

This work was supported in part by NSF grants IIS-1815455 and SES-1741441. 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, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443. The COLING 2016 Organizing Committee.

Amparo Elizabeth Cano-Basave and Yulan He. 2016. A study of the impact of persuasive argumentation in political debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1413. Association for Computational Linguistics.

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

Joshua Correll, Steven J Spencer, and Mark P Zanna. 2004. An affirmed self and an open mind: Self-affirmation and sensitivity to argument strength. *Journal of Experimental Social Psychology*, 40(3):350–356.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259. Association for Computational Linguistics.

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.

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-Volume 1*, pages 987–996. Association for Computational Linguistics.

Marco Guerini, Gozde Ozbal, and Carlo Strapparava. 2015. Echoes of persuasion: The effect of euphony in persuasive communication.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *EMNLP*, pages 1214–1223.

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. Association for Computational Linguistics.

Craig R Hullett. 2005. The impact of mood on persuasion: A meta-analysis. *Communication Research*, 32(4):423–442.

Timothy R. Johnson and Jerry Goldman. 2009. *A good quarrel: America's top legal reporters share stories from inside the supreme court*. University of Michigan Press.

Charles G Lord, Lee Ross, and Mark R Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098.

Stephanie M Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder : Audience effects in persuasion. *arXiv preprint arXiv:1708.09085*.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. 19:1–22.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372403.

Richard E Petty, John T Cacioppo, and Rachel Goldman. 1981. Personal involvement as a determinant of argument-based persuasion. *Journal of personality and social psychology*, 41(5):847.

Sonia Roccas, Lilach Sagiv, Shalom H. Schwartz, and Ariel Knafo. 2002. The big five personality factors and personal values. *Personality and Social Psychology Bulletin*, 28(6):789–801.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, volume 6.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *EMNLP*.

Warren T. Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of abnormal and social psychology*, 66:574–83.

Chenhao Tan and Lillian Lee. 2016. Talk it up or play it down? (un)expected correlations between (de-)emphasis and recurrence of discussion points in consequential u.s. economic policy meetings. Presented in Text as Data.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. *CoRR*, abs/1602.01103.

Robert P Vallone, Lee Ross, and Mark R Lepper. 1985. The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the beirut massacre. *Journal of personality and social psychology*, 49(3):577.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 812–817, Istanbul, Turkey. European Language Resources Association (ELRA). ACL Anthology Identifier: L12-1643.

Lu Wang, Nick Beauchamp, Sarah Shugars, and Kechen Qin. 2017. Winning on the merits: The joint effects of content and style on debate outcomes. *TACL*, 5:219–232.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 347–354. Association for Computational Linguistics.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in oxford-style debates. *arXiv preprint arXiv:1604.03114*.