Exploring the Role of Prior Beliefs for Argument Persuasion

Esin Durmus
Cornell University
ed459@cornell.edu

Claire Cardie
Cornell University
cardie@cs.cornell.edu

Abstract

Public debate forums provide a common platform for exchanging opinions on a topic of interest. While recent studies in natural language processing (NLP) have provided empirical evidence that the language of the debaters and their patterns of interaction play a key role in changing the mind of a reader, research in psychology has shown that prior beliefs can affect our interpretation of an argument and could therefore constitute a competing alternative explanation for resistance to changing one's stance. To study the actual effect of language use vs. prior beliefs on persuasion, we provide a new dataset and propose a controlled setting that takes into consideration two reader-level factors: political and religious ideology. We find that prior beliefs affected by these reader-level factors play a more important role than language use effects and argue that it is important to account for them in NLP studies of persuasion.

1 Introduction

Public debate forums provide to participants a common platform for expressing their point of view on a topic; they also present to participants the different sides of an argument. The latter can be particularly important: awareness of divergent points of view allows one, in theory, to make a fair and informed decision about an issue; and exposure to new points of view can furthermore possibly persuade a reader to change his overall stance on a topic.

Research in natural language processing (NLP) has begun to study persuasive writing and the role of language in persuasion. Tan et al. (2016) and Zhang et al. (2016), for example, have shown that the language of opinion holders or debaters and their patterns of interaction play a key role in changing the mind of a reader. At the same time, research in psychology has shown that prior beliefs can affect our interpretation of an argument even when the argument consists of numbers and empirical studies that would seemingly belie misinterpretation (Lord et al., 1979; Vallone et al., 1985; Chambliss and Garner, 1996).

We hypothesize that studying the actual effect of language on persuasion will require a more controlled experimental setting — one that takes into account any potentially confounding user-level (i.e., reader-level) factors[1] that could cause a person to change, or keep a person from changing, his opinion. In this paper we study one such type of factor: the prior beliefs of the reader as impacted by their political or religious ideology. We adopt this focus since it has been shown that ideologies play an important role for an individual when they form beliefs about controversial topics, and potentially affect how open the individual is to being persuaded (Stout and Buddenbaum, 1996; Goren, 2005; Croucher and Harris, 2012).

We first present a dataset of online debates that enables us to construct the setting described above in which we can study the effect of language on persuasion while taking into account selected user-level factors. In addition to the text of the debates, the dataset contains a multitude of background information on the users of the debate platform. To the best of our knowledge, it is the first publicly available dataset of debates that simultaneously provides such comprehensive information about the debates, the debaters and those voting on the debates.

With the dataset in hand, we then propose the novel task of studying persuasion (1) at the level of individual users, and (2) in a setting that can control for selected user-level factors, in our case, the prior beliefs associated with the political or

[1] Variables that affect both the dependent and independent variables causing misleading associations.

Proceedings of NAACL-HLT 2018, pages 1035–1045
New Orleans, Louisiana, June 1 - 6, 2018.

religious ideology of the debaters and voters. In particular, previous studies focus on predicting the winner of a debate based on the cumulative change in pre-debate vs. post-debate votes for the opposing sides (Zhang et al., 2016; Potash and Rumshisky, 2017). In contrast, we aim to predict which debater an individual user (i.e., reader of the debate) perceives as more successful, given their stated political and religious ideology.

Finally, we identify which features appear to be most important for persuasion, considering the selected user-level factors as well as the more traditional linguistic features associated with the language of the debate itself. We hypothesize that the effect of political and religious ideology will be stronger when the debate topic is *Politics* and *Religion*, respectively. To test this hypothesis, we experiment with debates on only *Politics* or only *Religion* vs. debates from all topics including *Music*, *Health*, *Arts*, etc.

Our main finding is that prior beliefs associated with the selected user-level factors play a larger role than linguistic features when predicting the successful debater in a debate. In addition, the effect of these factors varies according to the topic of the debate topic. The best performance, however, is achieved when we rely on features extracted from user-level factors in conjunction with linguistic features derived from the debate text. Finally, we find that the set of linguistic features that emerges as the most predictive changes when we control for user-level factors (political and religious ideology) vs. when we do not, showing the importance of accounting for these factors when studying the effect of language on persuasion.

In the remainder of the paper, we describe the debate dataset (Section 2) and the prediction task (Section 3) followed by the experimental results and analysis (Section 4), related work (Section 5) and conclusions (Section 6).

2 Dataset

For this study, we collected 67,315 debates from debate.org[2] from 23 different topic categories including *Politics*, *Religion*, *Health*, *Science* and *Music*.[3] In addition to text of the debates, we collected 198,759 votes from the readers of these debates. Votes evaluate different dimensions of the debate.

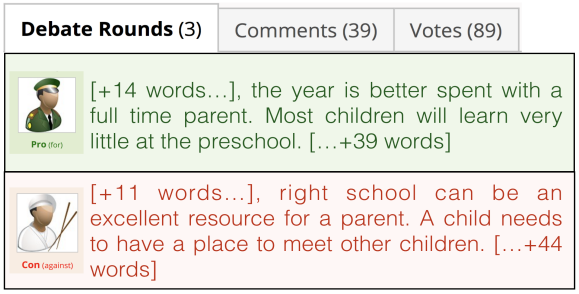


Figure 1: ROUND 1 for the debate claim "PRESCHOOL IS A WASTE OF TIME".

To study the effect of user characteristics, we collected user information for 36,294 different users. Aspects of the dataset most relevant to our task are explained in the following section in more detail.

2.1 Debates

Debate rounds. Each debate consists of a sequence of ROUNDS in which two debaters from opposing sides (one is supportive of the claim (i.e., PRO) and the other is against the claim (i.e., CON)) provide their arguments. Each debater has a single chance in a ROUND to make his points. Figure 1 shows an example ROUND 1 for the debate claim "PRESCHOOL IS A WASTE OF TIME". The number of ROUNDS in debates ranges from 1 to 5 and the majority of debates (61,474 out of 67,315) contain 3 or more ROUNDS.

Votes. All users in the *debate.org* community can vote on debates. As shown in Figure 2, voters share their stances on the debate topic before and after the debate and evaluate the debaters' conduct, their spelling and grammar, the convincingness of their arguments and the reliability of the sources they refer to. For each such dimension, voters have the option to choose one of the debaters as better or indicate a tie. This fine-grained voting system gives a glimpse into the reasoning behind the voters' decisions.

2.1.1 Determining the successful debater

There are two alternate criteria for determining the successful debater in a debate. Our experiments consider both.

Criterion 1: Argument quality. As shown in Figure 2, debaters get points for each dimension of the debate. The most important dimension — in

[2] www.debate.org

[3] The dataset will be made publicly available at http://www.cs.cornell.edu/ esindurmus/.

	Debater 1	Debater 2	Tied	
Agreed with before the debate:	✔	-	-	*0 points*
Agreed with after the debate:	-	✔	-	*0 points*
Who had better conduct:	-	✔	-	*1 point*
Had better spelling and grammar:	-	✔	-	*1 point*
Made more convincing arguments:	-	✔	-	*3 points*
Used the most reliable sources:	-	✔	-	*2 points*
Total points awarded:	**0**	**7**		

Figure 2: An example post-debate vote. Convincingness of arguments contributes to the total points the most.

that it contributes most to the point total — is making convincing arguments. *debate.org* uses Criterion 1 to determine the winner of a debate.

Criterion 2: Convinced voters. Since voters share their stances before and after the debate, the debater who convinces more voters to change their stance is declared as the winner.

2.2 User information

On *debate.org*, each user has the option to share demographic and private state information such as their age, gender, ethnicity, political ideology, religious ideology, income level, education level, the president and the political party they support. Beyond that, we have access to information about their activities on the website such as their overall success rate of winning debates, the debates they participated in as a debater or voter, and their votes. An example of a user profile is shown in Figure 3.

Opinions on the *big issues*. *debate.org* maintains a list of the most controversial debate topics as determined by the editors of the website. These are referred to as *big issues*.[4] Each user shares his stance on each *big issue* on his profile (see Figure 3): either PRO (in favor), CON (against), N/O (no opinion), N/S (not saying) or UND (undecided).

3 Prediction task: which debater will be declared as more successful by an individual voter?

In this section, we first analyze which dimensions of argument quality are the most important for determining the successful debater. Then, we analyze whether there is any connection between selected user-level factors and users' opinions on the *big issues* to see if we can infer their opinions from these factors. Finally, using our findings from these analyses, we perform the task of predicting which debater will be perceived as more successful by an individual voter.

[4] http://www.debate.org/big-issues/

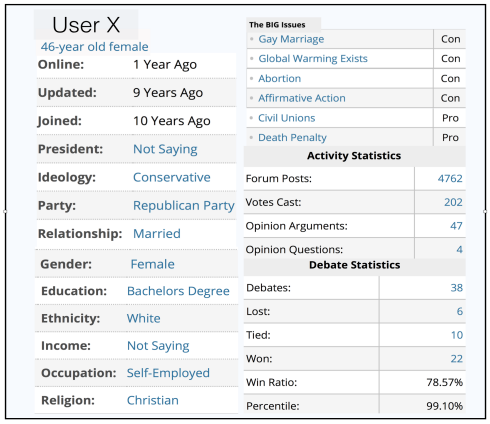


Figure 3: An example of a (partial) user profile. Top right: Some of the *big issues* on which the user shares his opinion are included. The user is against (CON) abortion and gay marriage and in favor of (PRO) the death penalty.

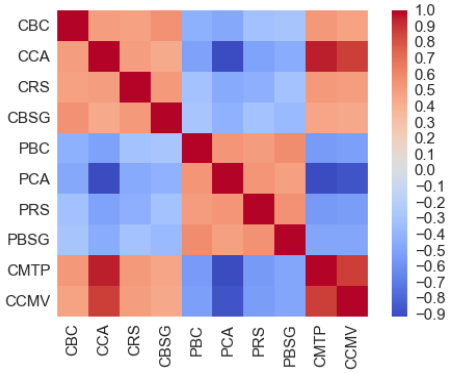


Figure 4: The correlations among argument quality dimensions.

3.1 Relationships between argument quality dimensions

Figure 4 shows the correlation between pairs of voting dimensions (in the first 8 rows and columns) and the correlation of each dimension with (1) getting more points (row or column 9) and (2) convincing more people as a debater (final row or column). Abbreviations stand for (on the CON side): has better conduct (CBC), makes more convincing arguments (CCA), uses more reliable sources (CRS), has better spelling and grammar (CBSG), gets more total points (CMTP) and convinces more voters (CCMV). For the PRO side we

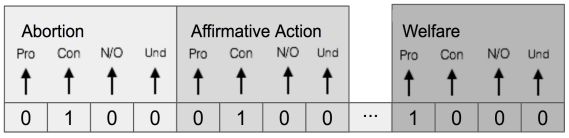


Figure 5: The representation of the BIGISSUES vector derived by this user's decisions on *big issues*. Here, the user is CON for ABORTION and AFFIRMATIVE ACTION issues and PRO for the WELFARE issue.

use PBC, PCA, and so on.

From Figure 4, we can see that making more convincing arguments (CCA) correlates the most with total points (CMTP) and convincing more voters (CCMV). This analysis motivates us to identify the linguistic features that are indicators of more convincing arguments.

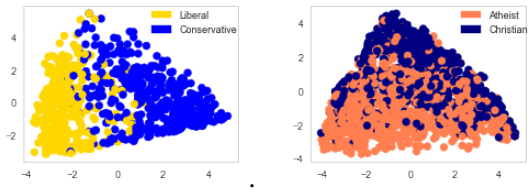
3.2 The relationship between a user's opinions on the *big issues* and their prior beliefs

We disentangle different aspects of a person's prior beliefs to understand how well each correlates with their opinions on the *big issues*. As noted earlier, we focus here only on prior beliefs in the form of self-identified political and religious ideology.

Representing the *big issues*. To represent the opinions of a user on a *big issue*, we use a four-dimensional one-hot encoding where the indices of the vector correspond to PRO, CON, N/O (no opinion), and UND (undecided), consecutively (1 if the user chooses that value for the issue, 0 otherwise). Note that we do not have a representation for N/S since we eliminate users having N/S for at least one *big issue* for this study. We then concatenate the vector for each *big issue* to get a representation for a user's stance on all the *big issues* as shown in Figure 5. We denote this vector by BIGISSUES.

We test the correlation between the individual's opinions on *big issues* and the selected user-level factors in this study using two different approaches: clustering and classification.

Clustering the users' decisions on *big issues*. We apply PCA on the BIGISSUES vectors of users who identified themselves as CONSERVATIVE vs. LIBERAL (740 users). We do the same for the users who identified themselves as ATHEIST vs. CHRISTIAN (1501 users). In Figure 6, we see that there are distinctive clusters of CONSERVATIVE vs. LIBERAL users in the two-dimensional representation



(a) LIBERAL VS. CONSERVATIVE (b) ATHEIST VS. CHRISTIAN.

Figure 6: PCA representation of decisions on *big issues* color-coded with political and religious ideology. We see more distinctive clusters for CONSERVATIVE vs. LIBERAL users suggesting that people's opinions are more correlated with their political ideology.

Prior belief type	Majority	BIGISSUES
Political ideology	57.70%	92.43%
Religious Ideology	52.70%	82.81%

Table 1: Accuracy using majority baseline vs. BIGISSUES vectors as features.

while for ATHEIST VS. CHRISTIAN, the separation is not as distinct. This suggests that people's opinions on the *big issues* identified by *debate.org* correlate more with their political ideology than their religious ideology.

Classification approach. We also treat this as a classification task[5] using the BIGISSUES vectors for each user as features and the user's religious and political ideology as the labels to be predicted. So the classification task is: Given the user's BIGISSUES vector, predict his political and religious ideology. Table 1 shows the accuracy for each case. We see that using the BIGISSUES vectors as features performs significantly better[6] than majority baseline[7].

This analysis shows that there is a clear relationship between people's opinions on the *big issues* and the selected user-level factors. It raises the question of whether it is even possible to persuade someone with prior beliefs relevant to a debate claim to change their stance on the issue. It may be the case that people prefer to agree with the individuals having the same (or similar) beliefs regardless of the quality of the arguments and the

[5]For all the classification tasks described in this paper, we experiment with logistic regression, optimizing the regularizer ($\ell 1$ or $\ell 2$) and the regularization parameter C (between 10^{-5} and 10^{5}).

[6]We performed the McNemar significance test.

[7]The majority class baseline predicts CONSERVATIVE for political and CHRISTIAN for religious ideology for each example, respectively.

particular language used. Therefore, it is important to understand the relative effect of prior beliefs vs. argument strength on persuasion.

3.3 Task descriptions

Some of the previous work in NLP on persuasion focuses on predicting the winner of a debate as determined by the change in the number of people supporting each stance before and after the debate (Zhang et al., 2016; Potash and Rumshisky, 2017). However, we believe that studies of the effect of language on persuasion should take into account other, extra-linguistic, factors that can affect opinion change: in particular, we propose an experimental framework for studying the effect of language on persuasion that aims to control for the prior beliefs of the reader as denoted through their self-identified political and religious ideologies. As a result, we study a more fine-grained prediction task: for an individual voter, predict which side/debater/argument the voter will declare as the winner.

Task 1 : Controlling for religious ideology. In the first task, we control for religious ideology by selecting debates for which each of the two debaters is from a different religious ideology (e.g., debater 1 is ATHEIST, debater 2 is CHRISTIAN). In addition, we consider only voters that (a) self-identify with one of these religious ideologies (e.g., the voter is either ATHEIST or CHRISTIAN) and (b) changed their stance on the debate claim post-debate vs. pre-debate. For each such voter, we want to predict which of the PRO-side debater or the CON-side debater did the convincing. Thus, in this task, we use *Criterion 2* to determine the winner of the debate from the point of view of the voter. Our hypothesis is that the voter will be convinced by the debater that espouses the religious ideology of the voter.

In this setting, we can study the factors that are important for a particular voter to be convinced by a debater. This setting also provides an opportunity to understand how the voters who change their minds perceive arguments from a debater who is expressing the same vs. the opposing prior belief.

To study the effect of the debate topic, we perform this study for two cases — debates belonging to the *Religion* category and then all the categories. The *Religion* category contains debates like "IS THE BIBLE AGAINST WOMEN'S RIGHTS?" and "RELIGIOUS THEORIES SHOULD NOT BE TAUGHT IN SCHOOL". We want to see how strongly a user's religious ideology affects the persuasive effect of language in such a topic as compared to the all topics. We expect to see stronger effects of prior beliefs for debates on *Religion*.

Task 2: Controlling for political ideology. Similar to the setting described above, Task 2 controls for political ideology. In particular, we only use debates where the two debaters are from different political ideologies (CONSERVATIVE vs. LIBERAL). In contrast to Task 1, we consider all voters that self-identify with one of the two debater ideologies (regardless of whether the voter's stance changed post-debate vs. pre-debate). This time, we predict whether the voter gives more total points to the PRO side or the CON side argument. Thus, Task 2 uses *Criterion 1* to determine the winner of the debate from the point of view of the voter. Our hypothesis is that the voter will assign more points to the debater that has the same political ideology as the voter.

For this task too, we perform the study for two cases — debates from the *Politics* category only and debates from all categories. And we expect to see stronger effects of prior beliefs for debates on *Politics*.

3.4 Features

The features we use in our model are shown in Table 2. They can be divided into two groups — features that describe the prior beliefs of the users and linguistic features of the arguments themselves.

User features

We use the cosine similarities between the voter and each of the debaters' *big issue* vectors. These features give a good approximation of the overall similarity of two user's opinions. Second, we use indicator features to encode whether the religious and political beliefs of the voter match those of each of the debaters.

Linguistic features

We extract linguistic features separately for both the PRO and CON side of the debate (combining all the utterances of PRO across different turns and doing the same for CON). Table 2 contains a list of these features. It includes features that carry information about the style of the language (e.g., usage of modal verbs, length, punctuation), represent different semantic aspects of the argu-

User-based features	Description
Opinion similarity.	For $userA$ and $userB$, the cosine similarity of BIGISSUES$_{userA}$ and BIGISSUES$_{userB}$.
Matching features.	For $userA$ and $userB$, 1 if $userA_f$==$userB_f$, 0 otherwise where $f \in$ {political ideology, religious ideology}. We denote these features as *matching political ideology* and *matching religious ideology*.
Linguistic features	**Description**
Length.	Number of tokens.
Tf-idf.	Unigram, bigram and trigram features.
Referring to the opponent.	Whether the debater refers to their opponent using words or phrases like "opponent, my opponent".
Politeness cues.	Whether the text includes any signs of politeness such as "thank" and "welcome".
Showing evidence.	Whether the text has any signs of citing any other sources (e.g., phrases like "according to"), or quotation.
Sentiment.	Average sentiment polarity.
Subjectivity (Wilson et al., 2005).	Number of words with negative strong, negative weak, positive strong, and positive weak subjectivity.
Swear words.	# of swear words.
Connotation score (Feng and Hirst, 2011).	Average # of words with positive, negative and neutral connotation.
Personal pronouns.	Usage of first, second, and third person pronouns.
Modal verbs.	Usage of modal verbs.
Argument lexicon features. (Somasundaran et al., 2007).	# of phrases corresponding to different argumentation styles.
Spelling.	# of spelling errors.
Links.	# of links.
Numbers.	# of numbers.
Exclamation marks.	# of exclamation marks.
Questions.	# of questions.

Table 2: Feature descriptions

ment (e.g., showing evidence, connotation (Feng and Hirst, 2011), subjectivity (Wilson et al., 2005), sentiment, swear word features) as well as features that convey different argumentation styles (argument lexicon features (Somasundaran and Wiebe, 2010). Argument lexicon features include the counts for the phrases that match with the regular expressions of argumentation styles such as assessment, authority, conditioning, contrasting, emphasizing, generalizing, empathy, inconsistency, necessity, possibility, priority, rhetorical questions, desire, and difficulty. We then concatenate these features to get a single feature representation for the entire debate.

4 Results and Analysis

For each of the tasks, prediction accuracy is evaluated using 5-fold cross validation. We pick the model parameters for each split with 3-fold cross validation on the training set. We do ablation for each of user-based and linguistic features. We report the results for the feature sets that perform better than the baseline.

We perform analysis by training logistic regression models using only user-based features, only linguistic features and finally combining user-based and linguistic features for both the tasks.

	Accuracy
Baseline	
Majority	56.10%
User-based Features	
Matching religious ideology	65.37 %
Linguistic features	
Personal pronouns	57.00 %
Connotation	61.26 %
All two features above	65.37 %
User-based+linguistic features	
USER*+ Personal pronouns	65.37%
USER*+ Connotation	66.42%
USER*+ LANGUAGE*	64.37%

Table 3: Results for Task 1 for debates in category *Religion*. USER* represents the best performing combination of user-based features. LANGUAGE* represents the best performing combination of linguistic features. Since using linguistic features only would give the same prediction for all voters in a debate, the maximum accuracy that can be achieved using language features only is 92.86%.

	Accuracy
Baseline	
Majority	57.31%
User-based Features	
Matching religious ideology	62.79 %
Matching religious ideology+ Opinion similarity	62.97%
Linguistic features	
Length [8]	61.01 %
User-based+linguistic features	
USER* + Length	64.56 %
USER*+ Length + Exclamation marks	65.74%

Table 4: Results for Task 1 for debates in all categories. The maximum accuracy that can be achieved using language features only is 95.77%.

Task 1 for debates in category *Religion*. As shown in Table 3, the majority baseline (predicting the winner side of the majority of training examples out of PRO or CON) gets 56.10% accuracy. User features alone perform significantly better than the majority baseline. The most important user-based feature is *matching religious ideology*. This means it is very likely that people change their views in favor of a debater with the same religious ideology. In a linguistic-only features analysis, combination of the *personal pronouns* and *connotation* features emerge as most important and also perform significantly better than the majority baseline at 65.37% accuracy. When we use both user-based and linguistic features to predict, the accuracy improves to 66.42% with *connotation* features. An interesting observation is that including the user-based features along with the linguistic features changes the set of important linguistic features for persuasion removing the *personal pronouns* from the important linguistic features set. This shows the importance of studying potentially confounding user-level factors.

Task 1 for debates in all categories. As shown in Table 4, for the experiments with user-based features only, *matching religious ideology* and *opinion similarity* features are the most important. For this task, *length* is the most predictive linguistic feature and can achieve significant improvement over the baseline (61.01%). When we combine the language features with user-based features, we see that with *exclamation mark* the accuracy improves to (65.74%).

Task 2 for debates in category *Politics*. As shown in Table 5, using user-based features only, the *matching political ideology* feature performs the best (80.40%). Linguistic features (refer to Table 5 for the full list) alone, however, can still obtain significantly better accuracy than the baseline (59.60%). The most important linguistic features include *approval*, *politeness*, *modal verbs*, *punctuation* and *argument lexicon features* such as *rhetorical questions* and *emphasizing*. When combining this linguistic feature set with the *matching political ideology* feature, we see that with the accuracy improves to (81.81%). *Length* feature does not give any improvement when it is combined with the user features.

Task 2 for debates in all categories. As shown in Table 6, when we include all categories, we see that the best performing user-based feature is the *opinion similarity* feature (73.96%). When using language features only, *length* feature (56.88%) is the most important. For this setting, the best accuracy is achieved when we combine user features with *length* and *Tf-idf* features. We see that the set of language features that improve the performance of user-based features do not include some of that perform significantly better than the baseline when used alone (*modal verbs* and *politeness* features).

	Accuracy
Baseline	
Majority	50.91%
User-based Features	
Opinion similarity	80.00 %
Matching political ideology	80.40 %
Linguistic features	
Length	57.37 %
linguistic feature set	59.60 %
User-based+linguistic features	
USER*+ *linguistic feature set*	81.81%

Table 5: Results for Task 2 for debates in category *Politics*. The maximum accuracy that can be achieved using linguistic features only is 75.35%. The *linguistic feature set* includes *rhetorical questions, emphasizing, approval, exclamation mark, questions, politeness, referring to opponent, showing evidence, modals, links,* and *numbers* as features.

	Accuracy
Baseline	
Majority	51.75%
User-based Features	
Opinion similarity	73.96%
Linguistic features	
Length	56.88%
Politeness	55.00%
Modal verbs	52.32%
Tf-idf features	52.89 %
User-based+linguistic features	
USER*+ Length	74.53%
USER*+ Tf-idf	74.13%
USER*+ Length + Tf-idf	75.20%

Table 6: Results for Task 2 for debates in all categories. The maximum accuracy that can be achieved using linguistic features only is 74.53%.

5 Related Work

Below we provide an overview of related work from the multiple disciplines that study persuasion.

Argumentation mining. Although most recent work on argumentation has focused on identifying the structure of arguments and extracting argument components (Persing and Ng, 2015; Palau and Moens, 2009; Biran and Rambow, 2011; Mochales and Moens, 2011; Feng and Hirst, 2011; Stab and Gurevych, 2014; Lippi and Torroni, 2015; Park and Cardie, 2014; Nguyen and Litman, 2015; Peldszus and Stede, 2015; Niculae et al., 2017; Rosenthal and McKeown, 2015), more relevant is research on identifying the characteristics of persuasive text, e.g., what distinguishes persuasive from non-persuasive text (Tan et al., 2016; Zhang et al., 2016; **?**; Habernal and Gurevych, 2016a,b; Fang et al., 2016; Hidey et al., 2017). Similar to these, our work aims to understand the characteristics of persuasive text but also considers the effect of people's prior beliefs.

Persuasion. There has been a tremendous amount of research effort in the social sciences (including computational social science) to understand the characteristics of persuasive text (Kelman, 1961; Burgoon et al., 1975; Chaiken, 1987; Tykocinskl et al., 1994; Chambliss and Garner, 1996; Dillard and Pfau, 2002; Cialdini, 2007; Durik et al., 2008; Tan et al., 2014; Marquart and Naderer, 2016). Most relevant among these is the research of Tan et al. (2016), Habernal and Gurevych (2016a) and Hidey et al. (2017). Tan et al. (2016) focused on the effect of user interaction dynamics and language features looking at the ChangeMyView[9] (an internet forum) community on Reddit and found that user interaction patterns as well as linguistic features are connected to the success of persuasion. In contrast, Habernal and Gurevych (2016a) created a crowd-sourced corpus consisting of argument pairs and, given a pair of arguments, asked annotators which is more convincing. This allowed them to experiment with different features and machine learning techniques for persuasion prediction. Taking motivation from Aristotle's definition for modes of persuasion, Hidey et al. (2017) annotated claims and premises extracted from the ChangeMyView community with their semantic types to study if certain semantic types or different combinations of semantic types appear in persuasive but not in non-persuasive essays. In contrast to the above, our work focuses on persuasion in debates than monologues and forum datasets and accounts for the user-based features.

Persuasion in debates. Debates are another resource for studying the different aspects of persuasive arguments. Different from monologues where the audience is exposed to only one side of the opinions about an issue, debates allow the audience to see both sides of a particular issue via a

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

controlled discussion. There has been some work on argumentation and persuasion on online debates. Sridhar et al. (2015), Somasundaran and Wiebe (2010) and Hasan and Ng (2014), for example, studied detecting and modeling stance on online debates. Zhang et al. (2016) found that the side that can adapt to their opponents' discussion points over the course of the debate is more likely to be the winner. None of these studies investigated the role of prior beliefs in stance detection or persuasion.

User effects in persuasion. Persuasion is not independent from the characteristics of the people to be persuaded. Research in psychology has shown that people have biases in the ways they interpret the arguments they are exposed to because of their prior beliefs (Lord et al., 1979; Vallone et al., 1985; Chambliss and Garner, 1996). Understanding the effect of persuasion strategies on people, the biases people have and the effect of prior beliefs of people on their opinion change has been an active area of research interest (Correll et al., 2004; Hullett, 2005; Petty et al., 1981). Eagly and Chaiken (1975), for instance, found that the attractiveness of the communicator plays an important role in persuasion. Work in this area could be relevant for the future work on modeling shared characteristics between the user and the debaters. To the best of our knowledge, Lukin et al. (2017) is the most relevant work to ours since they consider features of the audience on persuasion. In particular, they studied the effect of an individual's personality features (open, agreeable, extrovert, neurotic, etc.) on the type of argument (factual vs. emotional) they find more persuasive. Our work differs from this work since we study debates and in our setting the voters can see the debaters' profiles as well as all the interactions between the two sides of the debate rather than only being exposed to a monologue. Finally, we look at different types of user profile information such as a user's religious and ideological beliefs and their opinions on various topics.

6 Conclusion

In this work we provide a new dataset of debates and a more controlled setting to study the effects of prior belief on persuasion. The dataset we provide and the framework we propose open several avenues for future research. One could explore the effect different aspects of people's background (e.g., gender, education level, ethnicity) on persuasion. Furthermore, it would be interesting to study how people's prior beliefs affect their other activities on the website and the language they use while interacting with people with the same and different prior beliefs. Finally, one could also try to understand in what aspects and how the language people with different prior beliefs/backgrounds use is different. These different directions would help people better understand characteristics of persuasive arguments and the effects of prior beliefs in language.

7 Acknowledgements

This work was supported in part by NSF grant SES-1741441 and DARPA DEFT Grant FA8750-13-2-0015. 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, DARPA or the U.S. Government. We thank Yoav Artzi, Faisal Ladhak, Amr Sharaf, Tianze Shi, Ashudeep Singh and the anonymous reviewers for their helpful feedback. We also thank the Cornell NLP group for their insightful comments.

References

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs. In *Semantic Computing (ICSC), 2011 Fifth IEEE International Conference on*. IEEE, pages 162–168.

Michael Burgoon, Stephen B Jones, and Diane Stewart. 1975. Toward a message-centered theory of persuasion: Three empirical investigations of language intensity1. *Human Communication Research* 1(3):240–256.

Shelly Chaiken. 1987. The heuristic model of persuasion. In *Social influence: the ontario symposium*. Hillsdale, NJ: Lawrence Erlbaum, 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. https://doi.org/10.1177/0741088396013003001.

Robert B. Cialdini. 2007. Influence: The psychology of persuasion.

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.

S.M. Croucher and T.M. Harris. 2012. *Religion and Communication: An Anthology of Extensions in Theory, Research, and Method*. Peter Lang. https://books.google.com/books?id=CTfpugAACAAJ.

James Price Dillard and Michael Pfau. 2002. *The persuasion handbook: Developments in theory and practice*. Sage Publications.

Amanda M Durik, M Anne Britt, Rebecca Reynolds, and Jennifer Storey. 2008. The effects of hedges in persuasive arguments: A nuanced analysis of language. *Journal of Language and Social Psychology* 27(3):217–234.

Alice H Eagly and Shelly Chaiken. 1975. An attribution analysis of the effect of communicator characteristics on opinion change: The case of communicator attractiveness. *Journal of personality and social psychology* 32(1):136.

Hao Fang, Hao Cheng, and Mari Ostendorf. 2016. Learning latent local conversation modes for predicting community endorsement in online discussions. *arXiv preprint arXiv:1608.04808* .

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*. Association for Computational Linguistics, pages 987–996.

Paul Goren. 2005. Party identification and core political values. *American Journal of Political Science* 49(4):881–896. https://doi.org/10.1111/j.1540-5907.2005.00161.x.

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

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *ACL (1)*.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *EMNLP*. volume 14, pages 751–762.

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*. Association for Computational Linguistics, Copenhagen, Denmark, pages 11–21. http://www.aclweb.org/anthology/W17-5102.

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

Herbert C Kelman. 1961. Processes of opinion change. *Public opinion quarterly* 25(1):57–78.

Marco Lippi and Paolo Torroni. 2015. Context-independent claim detection for argument mining. http://www.aaai.org/ocs/index.php/IJCAI/IJCAI15/paper/view/10942.

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* .

Franziska Marquart and Brigitte Naderer. 2016. *Communication and Persuasion: Central and Peripheral Routes to Attitude Change*, Springer Fachmedien Wiesbaden, Wiesbaden, pages 231–242. https://doi.org/10.1007/978-3-658-09923-7_20.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law* 19(1):1–22.

Huy Nguyen and Diane Litman. 2015. Extracting argument and domain words for identifying argument components in texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*. Association for Computational Linguistics, Denver, CO, pages 22–28. http://www.aclweb.org/anthology/W15-0503.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured svms and rnns. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pages 985–995. https://doi.org/10.18653/v1/P17-1091.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*. ACM, pages 98–107.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*. Association for Computational Linguistics, Baltimore, Maryland, pages 29–38. http://www.aclweb.org/anthology/W/W14/W14-2105.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 938–948. http://aclweb.org/anthology/D15-1110.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*. pages 543–552. http://aclweb.org/anthology/P/P15/P15-1053.pdf.

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.

Peter Potash and Anna Rumshisky. 2017. Towards debate automation: a recurrent model for predicting debate winners. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. pages 2455–2465.

Sara Rosenthal and Kathy McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *SIGDIAL Conference*. pages 168–177.

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.

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*. Association for Computational Linguistics, pages 116–124.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. volume 1, pages 116–125.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, pages 1501–1510. http://www.aclweb.org/anthology/C14-1142.

D.A. Stout and J.M. Buddenbaum. 1996. *Religion and mass media: audiences and adaptations*. Sage Publications. https://books.google.com/books?id=V4cKAQAAMAAJ.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic- and author-controlled natural experiments on twitter. *arXiv preprint arXiv:1405.1438* .

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*. International World Wide Web Conferences Steering Committee, pages 613–624.

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.

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.

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*. Association for Computational Linguistics, pages 347–354.

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