Exploiting Personal Characteristics of Debaters for Predicting Persuasiveness

Khalid Al-Khatib[1] **Michael Völske**[1] **Shahbaz Syed**[2] **Nikolay Kolyada**[1] **Benno Stein**[1]
[1] Bauhaus-Universität Weimar, Germany, <first>.<last>@uni-weimar.de
[2] Leipzig University, Germany, shahbaz.syed@uni-leipzig.de

Abstract

Predicting the persuasiveness of arguments has applications as diverse as writing assistance, essay scoring, and advertising. While clearly relevant to the task, the personal characteristics of an argument's source and audience have not yet been fully exploited toward automated persuasiveness prediction. In this paper, we model debaters' *prior beliefs*, *interests*, and *personality traits* based on their previous activity, without dependence on explicit user profiles or questionnaires. Using a dataset of over 60,000 argumentative discussions, comprising more than three million individual posts collected from the subreddit *r/ChangeMyView*, we demonstrate that our modeling of debater's characteristics enhances the prediction of argument persuasiveness as well as of debaters' resistance to persuasion.

1 Introduction

Persuasion is a primary goal of argumentation (O'Keefe, 2006). It is often carried out in the form of a debate or discussion, where debaters argue to *persuade* others to take certain stances on controversial topics. Several studies have examined persuasiveness in debates by probing the main factors for establishing persuasion, particularly regarding the role of linguistic features of debaters' arguments (Zhang et al., 2016), the interaction between debaters (Tan et al., 2016), and the personal characteristics of debaters (Durmus and Cardie, 2018).

While the impact of debaters' characteristics on persuasiveness has been observed in online debates, the exploitation of these characteristics for predicting persuasiveness has been done based on *explicit* characteristics-related information in users' profiles or on questionnaires. For example, Lukin et al. (2017a) performed a personality trait test for selected people and asked them for their stances on specific topics to estimate their beliefs. Also, Durmus and Cardie (2018) used the information in users' profiles in an online forum, where their stances on controversial topics are explicitly stated, as a proxy of their beliefs. Such a means of exploitation limits the applicability of predicting persuasiveness, as the characteristics of debaters are usually not explicitly available in online debates, and it is not practicable to survey every debater.

The paper at hand studies how the characteristics of debaters can be modeled *automatically* and utilized successfully for predicting persuasiveness. To this end, we propose a new approach of various features that capture the beliefs, interests, and personality traits of debaters on the subreddit "ChangeMyView" based on the debaters' previous activity on the Reddit.com platform.

We apply this approach to the tasks of predicting argument persuasiveness and predicting debater's resistance to persuasion. Our experiments show that incorporating debater characteristics improves the prediction effectiveness of the two tasks over previous approaches which rely primarily on linguistic features. Interestingly, personality traits alone were the most predictive feature for resistance to persuasion, outperforming the linguistic features of the post itself.

The contribution of this paper is three-fold:

1. A large-scale corpus of argumentative and general discussions mined from Reddit.com.[1]
2. Features that capture the beliefs, interests, and personality traits of debaters based on their posting history.
3. A characteristics-based approach that tackles two persuasiveness tasks with improved effectiveness over previous approaches.[2]

[1] The corpus can be found at webis.de/data and https://zenodo.org/record/3778298

[2] To reproduce our experiments, the code is found here: https://github.com/webis-de/ACL-20

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7067–7072
July 5 - 10, 2020.

2 Related Work

The prediction of argument persuasiveness has been investigated in several studies (e.g., (Tan et al., 2016), (Zhang et al., 2016), (Persing and Ng, 2017), and (Hidey and McKeown, 2018). To mitigate the lack of annotated data, Persing and Ng (2017) proposed a light supervision model for persuasiveness scoring by explicitly modeling errors that negatively impact the persuasiveness of an argument. Musi et al. (2018) built an annotated corpus of concessions in CMV discussions using expert annotations and automatic classification. They observed that concessions are equally distributed among persuasive and non-persuasive threads and that they do not play any significant role as a means of persuasion. Studying the effect of argument sequencing, Hidey and McKeown (2018) provided evidence that the order in which arguments are presented plays a crucial role in persuasion. Considering the importance of linguistic features, Luu et al. (2019) studied debater skill as it improves over time due to prolonged interaction with other debaters. Combining linguistic features such as length of turns, and co-occurrence of hedges and fighting words, they developed a strong estimator of debaters' persuasive skill over time.

Apart from content-based features, modeling the audience is crucial for predicting persuasiveness. (Lukin et al., 2017a) studied the interaction of social media argument types with audience factors, to compare the belief change that results from social media dialogs to that from professionally curated monologic summaries. Participants were profiled for prior beliefs and personality types—neutral and balanced arguments were successful at changing the beliefs of all participants. In contrast, an entrenched audience was convinced by more emotional dialogs. (Durmus and Cardie, 2018) further explored the role of prior beliefs by predicting the success of debaters with explicitly stated religious and political ideologies, and found that readers were more likely to be convinced by a debater with the same ideology. (Longpre et al., 2019) examined linguistic features of debates together with audience features such as demographic information, prior beliefs, and debate platform behavior. They found that for *a priori* undecided users, audience features were prominent in predicting persuasiveness. For decided users, stylistic features of the argument were more effective.

Closely related to our work, Durmus and Cardie (2019) explored the effects of debaters' language, their prior beliefs and traits, and social interactions with other users on the DDO (debate.org) platform. The social interaction features were crucial in predicting the success of a debater, and combining them with features capturing debaters' language performed best. DDO explicitly provides information on personal traits of debaters, including demographics such as gender, ethnicity, and user's beliefs. Our data source lacks this information, which increases the difficulty of modeling users.

Recently, Guo et al. (2020) modeled the interplay of comments to study their cumulative influence on persuading the audience. They proposed a sequential model that captures the interplay as local and non-local dependencies and outperforms studies focusing only on lexical features.

The "ChangeMyView" subreddit (CMV) has been exploited for argument persuasiveness in many studies. For example, (Tan et al., 2016), (Hidey and McKeown, 2018), and (Habernal et al., 2018) used CMV as a source of real-world persuasive discourse.

3 Persuasiveness Tasks and Data

In this paper, we address the two persuasiveness tasks that have been proposed by Tan et al. (2016):

1. Predicting argument persuasiveness: given a debate topic and an argument regarding it, the task is to predict if the argument is persuasive, in terms of whether it is able to change the stance of an opponent.

2. Predicting resistance to persuasion: given a controversial topic (with a specific stance towards it) written by a debater, the task is to identify whether the debater's stance is resistant.

We use Reddit.com as a source of debates. This platform comprises a variety of user-generated content, organized within communities called "subreddits". The subreddit "*r/ChangeMyView*" (CMV) focuses on organized debates. As shown in Figure 1, contributors to CMV make an original post (OP) stating their stance on a debate topic of their choice. Other Reddit users may post opposing comments in response, to which the submitter of the OP may respond in turn, and award a "delta" to any comment that successfully changed their stance.

The CMV setting allows deriving gold standard labels for the two studied persuasiveness tasks. In

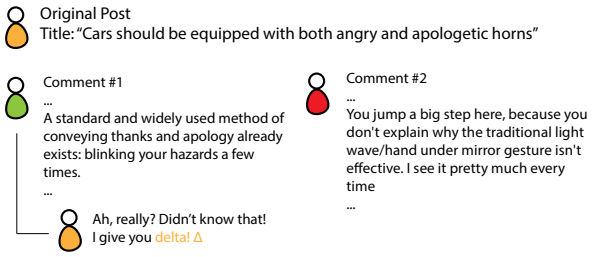

Figure 1: Exemplary excerpt of a CMV original post and two comments (i.e, arguments). Comment #1 is awarded a "delta" by the user who submitted the original post.

specific, we can assume that the comments that receive deltas are persuasive compared to those which do not. Here, an individual comment, made by a Reddit user in response to a CMV post, can be regarded as an argument. This is a simplification, but it is reasonable considering the characteristics of CMV (Tan et al., 2016). Also, If the user who submits the OP gave a delta to any response, we can assume that their stance is malleable.

Following the assumptions above, CMV has been crawled and the comments there have been labeled for persuasiveness, resulting in the *CMV corpus* of Tan et al. (2016), which covers the complete set of posts and comments until 2015. To extend this corpus, we collected all available CMV posts and comments from the foundation of the subreddit in 2005 until September 2017. Table 1 shows statistics for both corpora. To acquire debaters' posting history, which we employ in our approach (Section 4), we also collect all posts and comments across all of Reddit for each debater. The resulting extended corpus, *Webis-CMV-20*, is made available to the research community.

4 Modeling Debater Characteristics

We develop features to capture the *interests*, *prior beliefs*, and *personality traits* of debaters, and compute the similarity between two debaters based on these features.

4.1 Debater Interest

We capture debater interests based on their activities across subreddits. We rely on the assumption that the number of posts a debater makes in a subreddit (such as *r/politics* or *r/religion*) indicates their degree of interest in that topic. For instance, if a debater is interested in religious issues, it is likely that she posted to those subreddits which discuss

	CMV corpus	Webis-CMV-20
Discussion trees	20,626	65,169
Discussion Nodes	1,260,266	3,449,917
Posts ("OPs")	14,174	28,722
Unique authors	86,888	155,337

Table 1: Statistics of the corpus collected by Tan et al. (2016) as well as our own corpus.

religion such as 'Christianity' and 'Islam'.

We thus represent each debater by an *interest vector* depicting their interests across all subreddits. To constrain the impact of highly popular subreddits like *r/AskReddit* or *r/announcements*, we adopt a weighting scheme similar to tf-idf, where a subreddit s is represented as the fraction of a debater's total posts made within subreddit s, weighted by the logarithm of the ratio of the number of unique authors that posted in *r/ChangeMyView* to the number of authors that posted in subreddit s. The resulting interest vectors are very sparse (there are around one million subreddits), and thus not well suited for debaters similarity calculation. We apply two compression steps: First, we use data on subreddit topics from Snoopsnoo[3] to group subreddits into 720 categories, each represented as the sum of the interest vector elements for its constituent subreddits. Second, we apply principal component analysis to the result, and retain only the first five principal components, resulting in a 5-dimensional interest vector for each debater.

4.2 Debater Prior Beliefs

We assume that the totality of a debater's stances towards multiple topics is a good proxy for prior beliefs. To operationalize this assumption, we represent each debater by a *belief vector*, with each element representing the stance towards a particular topic. As topics, we consider the titles of Wikipedia articles:[4] across all Reddit posts by a given debater, we identify Wikipedia entities via entity linking,[5] compute the sentiment score[6] of sentences that mention entities, and assign this score as the stance of the debater towards this entity in the belief vector; entities mentioned in multiple contexts receive the median sentiment score.

[3] http://snoopsnoo.com/

[4] Although there is no way to be sure that Wikipedia encodes no bias in its topic coverage, it is by far the best source of important and controversial concepts.

[5] https://github.com/semanticize/semanticizest

[6] https://github.com/cjhutto/vaderSentiment

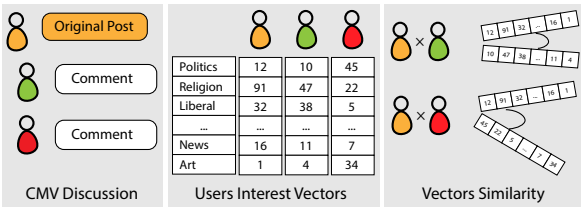

Figure 2: Example of generating debaters' interest features. We create interest vectors of the OP and the opponents, and then compute the cosine similarities between the interest vectors of the OP and each opponent.

4.3 Debater Personality Traits

Previous studies on the role of personality traits in influence (Nguyen et al., 2011) and argument synthesis (El Baff et al., 2019) used a psychometric dictionary-based text analysis. A similar approach for extracting personality traits using an external service (IBM Personality Insights) was done by Shmueli-Scheuer et al. (2019). Overall, those studies showed increased effectiveness on persuasion detection by including personality trait features. Hence, we process debaters' posts to reveal their personality traits, in which we represent each debater by a *traits vector* containing the distribution of the words in the debater's posts across particular classes such as *adventurous*, *genuine*, *self conscious*, to name a few. To this end, we apply the widely used Linguistic Inquiry and Word Count (LIWC) tool (Pennebaker et al., 2015) to the first 1000 words extracted from all Reddit posts made by a debater in temporal order.[7] For factors such as the big five personality traits, LIWC reports both raw scores and percentiles. Based on preliminary experiments, we use the concatenation of both as the final traits vector.

4.4 Debater Characteristics Similarity

Given the debater feature vectors created as described above, we compute the similarity between a pair of debaters as the cosine similarity of the concatenation of their characteristic vectors. Figure 2 shows an example for computing the similarity between users based on their interest vectors.

5 Experiments and Results

We evaluate our approach against the tasks described in Section 3. The tasks are predicting argument persuasiveness as well as predicting resistance to persuasion.

[7]The LIWC API recommends input to be at least 300 words and up to 1000 words long, hence we exclude debaters with less than 300 words of posting activity.

	CMV corpus	Webis-CMV-20
Post-comment pairs for persuasiveness prediction		
Training	3,456	12,496
Holdout	840	3,554
Delta	420	1,838
No delta	420	1,716
Posts for debater's resistance to persuasion prediction		
Training	3,934	6,791
Holdout	780	1,330
Delta	444	896
No delta	336	434

Table 2: Statistics of the Reddit derived task-specific data for the corpus collected by Tan et al. (2016) and our own corpus.

5.1 Experimental Setting

As a basis for our experiments, we use the *CMV corpus* of Tan et al. (2016) and our extended corpus *Webis-CMV-20* (see Section 3).

Since our approach depends on the activity history in previous Reddit.com posts for modeling debaters characteristics, we retain only those original posts and associated discussions where sufficient prior posting history, at least on the author of the original post, is available.

For *predicting argument persuasiveness*, we consider only the discussions where at least one delta was awarded. For each comment that received a delta, we sample another comment of similar length from the same discussion that did not, if exists. This procedure yields a total of 16,050 samples comprising the original post, the comment, the respective author characteristics, and the binary target of whether a delta was awarded, out of which 8,247 are positive (awarded a delta) and 7,803 negative; 3,554 of all samples are held out for testing.

For *predicting resistance to persuasion*, we sample 3,186 submissions whose author awarded at least one delta, and 4,935 submissions where no delta was awarded. Each sample comprises only the original post with its author characteristics, along with the binary target of whether a delta was awarded. We hold out 1,330 samples for testing.

Table 2 shows statistics of the training and holdout datasets for the two studied tasks.

	CMV corpus	Webis-CMV-20
Tan et al. (2016) model		
#words	66%	51.9%
BOW	64%	-
Interplay	**70%**	**57.8%**
Interplay + style	67%	-
All features	68%	57.7%
Our model		
BOW	60.4%	57.7%
Interest	60.5%	58.9%
Beliefs	61.5%	58.6%
Traits	**61.8%**	60.5%
All features	61.6%	**61.1%**

Table 3: Comparison of model effectiveness at the persuasiveness-prediction task. Reported are accuracy numbers for ease of comparison to related work.

5.2 Results

For our experiments, we re-implement the most powerful features proposed by Tan et al. (2016), including BOW, several interplay features (e.g., the number of common words between the original post and the comment), and various style features (e.g., the intensity of emotion and concreteness). We compare these features to the features proposed in Section 4 that model debater characteristics. Following related work, we employ a logistic regression classifier with L1 regularization. We fine-tune the parameters via 5-fold cross validation on the training sets. While incorporating debater characteristics in a persuasiveness prediction model leads to small improvements in our experiments, we find that predicting debater's resistance using only personality traits outperforms all other feature sets.

Predicting argument persuasiveness Features based only on the content of the post pair have proven quite effective at predicting persuasiveness in previous work—Tan et al. (2016) found the comment word count by itself to achieve significantly better than chance accuracy. Due to our sampling strategy which is biased towards negative samples with similar length to the positive ones, this feature performs considerably worse on our new corpus. We further explore how our features for modeling debater characteristics, when combined with linguistic features, improve the classification accuracy. As can be seen in Table 3, personality traits, interests, and beliefs slightly outperform linguistic features. On the *CMV corpus*, a model using only trait features is most effective, achieving 61.8% accuracy (AUC 0.66), while linguistic features only achieve an accuracy of 60.4% (AUC 0.61).

	CMV corpus	Webis-CMV-20
Tan et al. (2016) model		
BOW only	**0.54**	**0.52**
Our model		
BOW only	0.56	0.52
Traits only	**0.64**	**0.62**
All features	0.55	0.60

Table 4: Comparison of model effectiveness at the debater's resistance to persuasion prediction task. Reported are ROC-AUC numbers for ease of comparison to related work.

Predicting debater resistance to persuasion Table 4 shows the results for the debater's resistance to persuasion task on both corpora. As reported by Tan et al. (2016), predicting debater's resistance to persuasion using only linguistic features is a very challenging task (they showed that human annotators performed at no better than chance level). Our personality trait features vastly outperform merely linguistic features across both corpora. However, the individual traits themselves show a weak association with resistance to persuasion; for instance, the Pearson correlation coefficients are very small for the big five personality traits agreeableness (0.075), conscientiousness (-0.037), extraversion (-0.046), neuroticism (-0.067), and openness (0.019), suggesting that only a complex interplay of these characteristics is predictive of resistance to persuasion.

6 Conclusion

This paper proposes a new approach for modeling the personal characteristics of debaters including interests, prior beliefs, and personality traits for predicting both argument persuasiveness and debaters' resistance to persuasion. We hypothesize that these characteristics can be induced automatically from the history of debaters' activity such as their earlier texts. Based on this hypothesis, we develop a set of various features to capture debaters characteristics using the Reddit.com platform. Applying these features on persuasiveness corpora derived from the subreddit r/ChangeMyView, we accomplish a fair improvement on the effectiveness of tackling the studied persuasiveness tasks, particularly in predicting the debaters' resistance to persuasion. In the future, we plan to consider the ethos mode of persuasion by exploring how debaters strengthen their credibility in debates.

References

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. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2019. Modeling the Factors of User Success in Online Debate. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2701–2707. ACM.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational Argumentation Synthesis as a Language Modeling Task. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 54–64.

Zhen Guo, Zhe Zhang, and Munindar Singh. 2020. In Opinion Holders' Shoes: Modeling Cumulative Influence for View Change in Online Argumentation. In *Proceedings of The Web Conference 2020*, WWW '20, pages 2388–2399, New York, NY, USA. Association for Computing Machinery.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. Before name-calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation. 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 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher Hidey and Kathleen R. McKeown. 2018. Persuasive Influence Detection: The Role of Argument Sequencing. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5173–5180. AAAI Press.

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*, pages 167–176, Florence, Italy. Association for Computational Linguistics.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017a. 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. Association for Computational Linguistics.

Kelvin Luu, Chenhao Tan, and Noah A. Smith. 2019. Measuring Online Debaters' Persuasive Skill from Text over Time. *Trans. Assoc. Comput. Linguistics*, 7:537–550.

Elena Musi, Debanjan Ghosh, and Smaranda Muresan. 2018. ChangeMyView Through Concessions: Do Concessions Increase Persuasion? *arXiv preprint arXiv:1806.03223*.

Thin Nguyen, Dinh Phung, Brett Adams, and Svetha Venkatesh. 2011. Towards Discovery of Influence and Personality Traits through Social Link Prediction. In *Fifth International AAAI Conference on Weblogs and Social Media*.

Daniel J. O'Keefe. 2006. Persuasion. In *Encyclopedia of Rhetoric*. Oxford University Press.

James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. 2015. Linguistic Inquiry and Word Count: LIWC2015. Pennebaker Conglomerates, Austin, TX. www.liwc.net.

Isaac Persing and Vincent Ng. 2017. Lightly-Supervised Modeling of Argument Persuasiveness. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 594–604, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Michal Shmueli-Scheuer, Jonathan Herzig, David Konopnicki, and Tommy Sandbank. 2019. Detecting Persuasive Arguments based on Author-Reader Personality Traits and their Interaction. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*, pages 211–215.

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, pages 613–624, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

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. Association for Computational Linguistics.