Analyzing Persuasion Strategies of Debaters on Social Media

Matti Wiegmann[1*] **Khalid Al-Khatib**[2*] **Vishal Khanna**[1] **Benno Stein**[1]
[1] Bauhaus-Universität Weimar <first>.<last>@uni-weimar.de
[2] University of Groningen khalid.alkhatib@rug.nl

Abstract

Existing studies on the analysis of persuasion in online discussions focus on investigating the effectiveness of *comments* in discussions and ignore the analysis of the effectiveness of *debaters* over multiple discussions. In this paper, we propose to quantify debaters' effectiveness in the online discussion platform: "ChangeMyView" in order to explore diverse insights into their persuasion strategies. In particular, targeting debaters with different levels of effectiveness (e.g., good vs. poor), various behavioral characteristics (e.g., engagement), and text stylistic features (e.g., used frames) of debaters are carefully examined, leading to several outcomes that can be the backbone of writing assistants and persuasive text generation.

1 Introduction

Persuasion, a primary goal of argumentation, is the ability to convince people to do a certain action or form a particular belief (O'Keefe, 2006). Persuasion has always influenced the dynamics of communication and social interactions, either positively by raising awareness on critical issues like climate change or negatively by influencing the behavior of voters in elections or disseminating propaganda and fake news.

Social media, through its growing role in the formation of beliefs, has gained notable interest as a means to gather a deeper understanding of persuasion (Wang et al., 2021). The ChangeMyView (CMV) subreddit in particular has been used in various studies that model text persuasiveness using a variety of linguistic, argumentative, and behavioral features (e.g., (Hidey and McKeown, 2018), (Longpre et al., 2019), and (Guo et al., 2020)).

However, scholarly work on online persuasion centers around studying *persuasive comments* in individual discussions without considering the importance of analyzing *persuasive debaters* (Luu et al., 2019). Hence, debater strategies and their effectiveness have not been adequately studied. Understanding effective debating strategies and debater persuasiveness can be highly beneficial for media analysis, rhetorical review, and for learning debating skills. Besides, it can advance the development of several applications, where effective strategies can be recommended in writing assistants and dialog management systems or encoded in the backbone of text generation tools.

This paper focuses on analyzing the *debaters' persuasion strategies*, seeking to uncover the behavior, language style, and argumentative techniques that distinguish good from poor debaters. To do so, we categorize CMV debaters based on their effectiveness in persuasion and examine key differences in their behaviors and skills (i.e., engagement and experience), as well as their argument's style (at the semantic, syntactical, lexical, and pragmatic levels). We propose the task of identifying effective debaters and present an approach to tackle it.

Our analysis of debater strategies yields several insights. For example, we find that the effectiveness of persuasion improves over time for average debaters, the distribution of 'frames' in the debaters' arguments can play a major role in persuasion, and argumentative features based on the presence of certain types of arguments in the debaters' text do not seem sufficient to indicate the effectiveness of persuasion.

The contribution of this paper is threefold:

1. An extensive analysis of debater strategies across multiple discussions, addressing their behavior and stylistic aspects of their texts.
2. Insights about several techniques used by successful compared to unsuccessful debaters.
3. A new task of distinguishing good from poor debaters and an approach to address the task with multiple linguistic features.

All the resources developed in this paper can be found at https://doi.org/10.5281/zenodo.7034173.

[*] Equal contribution

Proceedings of the 29th International Conference on Computational Linguistics, pages 6897–6905
October 12–17, 2022.

2 Background and Related Work

In this section, we introduce CMV's core concepts as required for our study. Afterward, we overview the primary studies on modeling persuasiveness on CMV and related platforms.

2.1 Background

CMV is an open platform for users to engage in civilized discussions using sound arguments. CMV discussions are actively moderated to maintain the quality of argumentation. All comments and original posts must abide by the community rules.[1] These rules dictate a predictable structure for CMV debates, making them easy to process.

A CMV discussion begins with a user, called the *original poster (OP)*, submitting a marked request, called *original post*, to the CMV subreddit. The subreddit forbids non-debative submissions. The original post states the OP's stance on a controversial topic, relevant justifications and explanations of this stance, and an (implicit) call to "change my view". All other users of CMV called the *debaters*, can challenge the OP's stance and post opposing argumentative top-level comments. All debaters can respond to other comments to counter, cross-question, or defend their arguments, creating multi-layered and complex threads of conversation.

CMV offers two mechanisms to indicate comment persuasiveness: The delta (Δ) and the comment score. The delta mechanism allows the OP to mark up to one comment as persuasive. The 'awarded' deltas are aggregated and the per debater Δ-count is displayed publicly. Reddit's comment score is the per-comment sum of up and downvotes. The highest scoring comments are shown first. The comment score on CMV serves as an alt-metric indicating the persuasiveness as perceived by the community.

2.2 Related Work

The major work on the analysis of argument persuasiveness on social media (cf. (Tan et al., 2016), (Zhang et al., 2016), (Persing and Ng, 2017), and (Hidey and McKeown, 2018).) tries to determine how persuasive a comment is by solving the task: given two comments with a shared OP, identify the persuasive one. In contrast, our paper provides a higher-level analysis. We try to determine how persuasive a debater is by studying the debaters across multiple discussions, striving to disclose their persuasion strategies.

Employing argumentative features to predict comment persuasiveness is a well-established strategy; Egawa et al. (2019) annotated CMV discussions with elementary argumentative units (EUs) in a token-level five-class scheme: testimony, fact, value, policy, and rhetorical statement. The authors propose a Bi-LSTM-based sequence classifier for EU labeling. They conclude that EUs indicate persuasiveness if used effectively, 'fact' is the most persuasive, that the proportional distribution of types distinguishes CMV comments from original posts, and that persuasiveness is not indicated by the mere presence or absence of certain EUs.

Similarly, Hidey et al. (2017) annotated CMV discussions regarding arguments' claims as interpretation, evaluation, agreement, disagreement, or premises as ethos, logos, and pathos. The authors show that the relative positional distribution of argumentative components in a CMV comment is a signal for its persuasiveness. Additionally, Li et al. (2020) demonstrated the effectiveness of arguments' structural features in persuasiveness prediction. Multiple features were developed based on the usage of the proposition types reference, testimony, fact, value, and policy in the debaters' texts. The feature analysis showed that the presence of 'value' and 'testimony' bi-grams is more prevalent in persuasive argumentative texts, indicating that justifying claims with personal experiences is an effective persuasion strategy.

In this paper, we implemented the previously used argumentative features along with newly utilized ones like syntactic complexity, semantic similarity, and argument framing; the latter is shown to play a role in the debater's persuasiveness.

Different characteristics and behavior patterns of debaters were examined in a few papers. Addressing the debater behavior, Tan et al. (2016) investigated the role of debaters' interaction dynamics with the OP in persuasion and found that the debaters who responded early in the discussion tend to be more successful, that engaging with the original poster improves a debater's odds of success up to a threshold, and that higher debater participation in a discussion improves the odds of persuasion.

Targeting debaters' characteristics, Al-Khatib et al. (2020) modeled debaters' beliefs, personality traits, and interests based on their previous activities on Reddit, utilizing them for tackling the task

[1] CMV rules are stated on their wiki: https://www.reddit.com/r/changemyview/wiki/rules

of persuasiveness prediction. The study found the similarity between the characteristics of the OP and the debaters to be influential for effective persuasion. In comparison, our paper groups debaters based on their persuasiveness so we can probe the diverse strategies used by good vs. poor debaters.

Analyzing the discussion structure, Guo et al. (2020) hypothesized that persuading the OP in a CMV discussion happens gradually throughout a multi-party conversation rather than instantaneously. A prediction task was performed to model the cumulative effect of a sequence of comments in a CMV discussion and detect the position where the persuasion of the OPs occurs. Besides, a user study to evaluate the persuasiveness of debaters' arguments' was conducted, concluding that the perception of persuasiveness differs across individuals and that it is influenced by one's idiosyncrasies i.e. the same argument could be persuasive for one person but not persuasive for another. Likewise, Wei et al. (2016) considered the relevance ranking of CMV comments by their score in a discussion. They found the comment's score to be influenced by its temporal entry order as well as the past credibility of its corresponding debater. The credibility is measured by the number of prior deltas received by a debater. Several feature classes were used for the relevance ranking task, including linguistic features derived from the comment's text, interaction-based features obtained by modeling the CMV discussion as a tree, and argumentative features such as the proportion of argumentative text and argument relevance and originality.

The only work that targets the debater-level analysis is by Luu et al. (2019) and investigates how debaters' skill improves over time as they learn how to interact with other debaters. They present a strong estimator of the development of a debater's persuasive skill over time using several linguistic features, such as length of comments, co-occurrence of hedges, and fighting words.

Our work is distinct in several respects: First, we analyze CMV, as opposed to `Debate.org`, which is more strict and conventional regarding debate structure. Second, we analyze the relationship between the debaters' engagement, experience, and writing style across linguistic dimensions, accounting for the argumentative nature of debate texts. Finally, we address different levels of debater persuasiveness and scrutinize the differences in their argumentation strategies.

3 Data Acquisition and Preparation

To conduct our study on debater persuasion strategies, we created a dataset with $3,801$ CMV debaters, equally sampled for good, average, and poor debater persuasiveness. Here, we detail our data collection method, quantification of debater persuasiveness, and sampling method to balance the dataset by debater persuasiveness.

Quantifying Debater Persuasiveness We define the persuasiveness of a debater d with comments $c_1, \ldots, c_{i,\Delta 1}, \ldots, c_{j,\Delta k}, \ldots, c_n$ in CMV as ratio of delta comments c_Δ to all comments:

$$\text{Persuasiveness}(d) = \frac{k}{n}.$$

The persuasiveness is, hence, the number of debater's delta comments normalized by her total comment count. As Table 1 shows, this normalization is necessary because the delta-comment count correlates strongly with the total comment count.

Based on the persuasiveness score, we categorize debaters into three groups as follows:

1. **Good debaters** with a persuasiveness of 5% or above.

2. **Average debaters** with a persuasiveness between 0% and 5%.

3. **Poor debaters** with a persuasiveness of 0%; These debaters did not receive any delta during their active period on CMV.

The separation of debaters with a non-zero persuasiveness is based on the observation that obtaining any c_Δ is already challenging. Hence, the highly persuasive tail should be studied as a separate population. The 5%-threshold used in categorization separates the non-poor debaters into two groups of approximately equal size.

Collecting Debater Comments We obtained an initial set of CMV debates from the Webis CMV corpus (Al-Khatib et al., 2020), which comprises all CMV debates from June 2005 to September 2017. We extracted all top-level comments from the Webis CMV corpus and grouped them by debater. We discarded all inactive debaters with less than 10 comments and obtained an unbalanced dataset of $13,254$ CMV debaters along with their top-level comments on various debates. We only considered top-level comments since they serve as

"entries to the debate", while lower-level comments are either rebuttals or non-argumentative content like corrections, clarifications, or thanks.

Sampling Data In the intermediate dataset, 80% of the debaters are of poor persuasiveness and have never been awarded a Δ. Since we aim for a controlled analysis, we resampled the dataset in such a way that the distribution of CMV debaters is balanced by persuasiveness. Overall, we end up with $3,801$ entries, evenly distributed across the three debater categories.

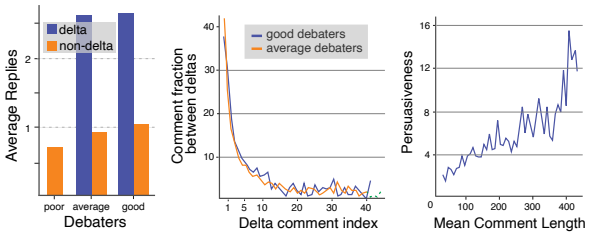
Our resampling strategy first added a "good" debater to the dataset by random and then selected one "average" and one "poor" debater with the same number of comments, or the closest number to that. If multiple candidate debaters existed, we minimized the absolute difference in mean comment length. This resampling strategy minimizes the effects of comment count and comment length in the dataset since both are indicative of persuasiveness (cf. Section 2).

4 Debater Engagement and Experience

Our first analysis concerns the relationship between the debaters' persuasiveness and their engagement with and experience on the CMV subreddit. We presume that engagement on CMV may correspond to rebuttals in live debates. Our findings suggest that a high engagement is indicative of persuasive debaters. We further inspect the relationship between experience and persuasiveness in both absolute measures such as comment count and active period and relative measures such as changes in style and persuasiveness with experience gain. Our findings suggest that debaters become more persuasive with increased experience, especially average debaters. However, the experience effect is not reflected in absolute experience measures, and hence it is hard to operationalize for classification.

4.1 Engagement

Figure 1a shows that persuasive comments and persuasive debaters are more engaging. We measure debater engagement by the average number of replies to persuasive and non-persuasive comments. Persuasive debaters get ~10% more replies to their total comments compared to average debaters and ~30% more replies compared to poor debaters. Persuasive comments get ~250% as many replies as non-persuasive comments.



(a) Engagement (b) Persuasion Freq. (c) Length

Figure 1: (a) Engagement of debaters by persuasiveness. (b) Evolution of the frequency of persuasive comments. (c) Persuasiveness by debaters' average comment length.

	Persuasiveness	Δ Count	Score
Comments	0.02	0.72	0.03
Active Period	-0.03	0.13	0.15

Table 1: Pearson ρ between three success measures (Persuasiveness; Δ Comment Count; Median Reddit Comment Score) and two absolute experience measures (Active Period: the time between the first and last comment on CMV; Number of Comments).

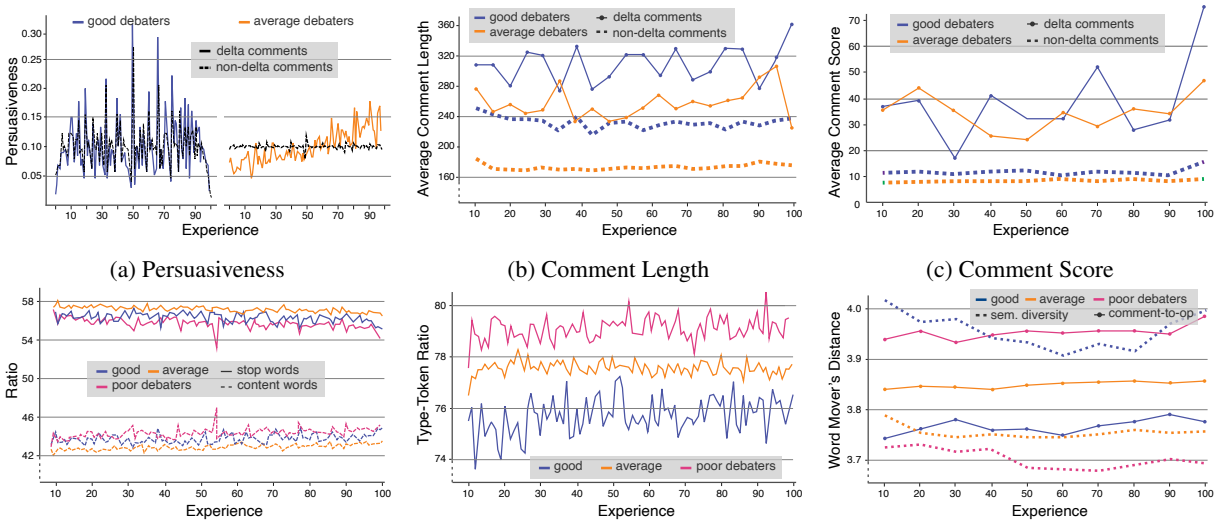
4.2 Absolute Experience

Table 1 shows that the absolute measures of experience are insufficient. We can observe that neither the active period—the time between the first and latest comment—nor the comment count correlates with persuasiveness or Reddit score. We disregard the correlation between the total comment count and the number of persuasive comments as evidence of debater experience without observing a correlation with persuasiveness.

4.3 Relative Experience

We model the relative experience of a debater on CMV as seen from the comment: A debater is inexperienced for her first comment and very experienced for her last; that is to say, the experience of the debater d of a comment c_t in a sequence $c_1, \ldots, c_n$ is $\text{Experience}(c_t) = \frac{t}{n}$. We analyze the impact of experience gain of good and average debaters on persuasiveness, persuasion frequency, comment length, as length is the most indicative feature in comment classification, and average comment score, which represents the CMV community's opinion on persuasiveness.

Persuasiveness Figure 2a shows that the overall persuasiveness of good debaters is largely unaffected by experience while the persuasiveness of average debaters almost doubles.



(a) Persuasiveness (b) Comment Length (c) Comment Score

(d) Stop Word Ratio (e) Type-Token Ratio (f) Semantic Similarity

Figure 2: Changes of various debater-level features with increasing relative experience. The color indicates persuasiveness and the line style (dashes/dots) indicates the secondary variable.

Persuasion Frequency Figure 1b shows that the persuasion frequency increases sharply up to the 5th persuasive comment for both good and average debaters and increases slightly up to the 15th persuasive comment. This indicates that debaters learn to replicate persuasive strategies and become more persuasive with experience. We measure persuasion frequency as the number of non-delta comments that occur between two consecutive delta comments, as a fraction of the total comments made. A decreasing delta-to-non-delta rate indicates more frequent persuasions.

Comment Length Figure 2b confirms the established assumption that length is highly indicative of persuasiveness. There is no indication that relative experience has any substantial impact on the length of delta or non-delta comments.

Average Comment Score Figure 2c shows that the mean-comment score, the alt-metric for community persuasiveness, increases with experience but not consistently. On average, however, debaters score higher on persuasive comments with increasing experience. The effect. however, is negligible on non-delta comments.

5 Debater Style Analysis

Stylistic features are frequently used to determine the characteristics of authors. Since stylistic features are indicative of persuasive comments, we consider stylistic features to also be indicative of persuasive debaters. In particular, we study the relationship between a debater's persuasiveness and the lexical, syntactic, semantic, and pragmatic dimensions. We found notable differences in persuasiveness in each dimension. The most substantial feature is again comment length. Additionally, we found that better debaters tend to have lower lexical diversity and syntactic complexity, but a higher semantic diversity. We also found correlations between certain word class patterns and certain patterns of elementary argumentative units, particularly rhetorical statements. Lastly, found that persuasive debaters use political and cultural identity frames more often.

5.1 Lexical Dimensions

Within the lexical dimension of style, we analyze the relation between debater persuasiveness and the (1) comment length and the (2) lexical diversity, in particular the stop-word and type-token ratio.

Comment Length Figure 1c shows that debaters with a higher mean comment length are also, consistently and without apparent bound, more persuasive on average. Figure 2b shows, independently of the debater's experience, that persuasive comments are longer than non-persuasive comments and that good debaters write longer (~20%) comments. These findings are consistent with previous evidence (cf. Section 2) and suggest that the comment length is highly indicative of the persuasiveness of comments and debaters alike.

WC n-gram	ρ	WC n-gram	ρ
IN JJ	0.11	PRP VBP	-0.13
NN IN JJ	0.10	PRP	-0.12
JJ NN IN	0.09	WRB VBP	-0.11
VBG DT JJ	0.08	NN WRB	-0.11

Table 2: Top Pearson ρ between a word class n-gram and persuasiveness.

Lexical Diversity Figure 2d shows that the differences in the stop-word ratio are consistently small (<1%) and have no direction since good debaters are between poor and average ones. However, Figure 2e shows that the type-token ratio has a higher effect size of 2% among the debater groups and has a direction. This suggests that good debaters write comments with lower lexical diversity.

5.2 Syntactic Dimensions

Within the syntactic dimension of style, we analyze the relationship between persuasiveness and syntactic complexity and the word class n-gram distribution.

Syntactic Complexity The complexity of a debater's text was measured based on the dependency parse trees of all sentences in her top-level comments. We measure the Pearson correlation between debater persuasiveness and three common syntactic complexity measures:[2] Outdegree centrality ($\rho = -0.17$), Closeness centrality ($\rho = -0.16$), and the number of dependents per word ($\rho = 0.17$). Since a high centrality indicates complex syntax, and persuasiveness is negatively correlated with centrality, our results suggest that good debaters use less complex syntax. However, all correlations are weak ($\rho <= 0.25$).

Word class n-grams Table 2 shows the word class 1–3-grams with the strongest correlation with persuasiveness. Here, better debaters use adjectives more and `PRP VBP` (e.g. you did ...) as well as `WRB VBP` (e.g. how did ...) less frequently. Although the correlation is weak and word-class n-grams are difficult to interpret, these results may indicate an impact of certain syntactical structures on debater persuasiveness as for comment persuasiveness (cf. Tan et al. (2016)). We determined the word class n-grams using NLTK and the Penn tagset since all CMV comments are English. We only inspected the 1,000 most frequent n-grams.

[2]We measured the complexity using https://github.com/tsproisl/textcomplexity

5.3 Semantic Dimension

Within the semantic dimension of style, we measure the relation between debater persuasiveness and the (1) semantic similarity between a debater's comment and the original post and the (2) semantic diversity within the comments of a debater. We use Word Movers Distance[3] (WMD, Kusner et al., 2015) to measure the semantic similarity.

Similarity between Comment and Original Post Figure 2f shows that the WMD is lower the more persuasive a debater is. Hence persuasive debater's comments are semantically more similar to the original post.

Semantic Diversity Figure 2f shows the semantic diversity for debaters with different persuasiveness, whereas the semantic diversity is higher for better debaters.

The semantic diversity should indicate if a debater prefers semantic depth (few different concepts discussed) or breadth (many different concepts discussed) within each comment. For lack of a better (lexeme-agnostic) intra-document semantic similarity measure, we use a sentence-based heuristic:

$$\mathrm{SemDiv}(c_k) = \frac{2}{n^2 - n} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathrm{WMD}(s_i, s_j).$$

Here, the semantic diversity of a debater is the average diversity of the comments $c_k = s_1, \ldots, s_n$, and the diversity of the comments is the average WMD between each pair of sentences (s_i, s_j). We assume WMD captures the semantic diversity between two sentences in this way.

5.4 Pragmatic Dimension

Within the pragmatic dimension of style, we measure the relation between debater persuasiveness and (1) the distribution of argumentative units: elementary units, claims, and premises, (2) framing strategies.

Argumentative Units Table 3 shows the argumentative unit n-grams which correlate the strongest with debater persuasiveness, while all other unit n-grams do not correlate with $\rho \leq 0.05$. All correlating units are elementary units, with rhetorical statements being the most persuasive. No claim or premise types correlate in a meaningful way with persuasiveness.

[3]We use Gensim with fastText embeddings

Unit n-gram	ρ	Unit n-gram	ρ
rhetoric	-0.194	policy	-0.110
value	-0.126	rhetoric rhetoric	-0.101
rhetoric value	-0.114	rhetoric rhetoric none	-0.063

Table 3: Argumentative units with largest absolute Pearson ρ with CMV debaters' persuasiveness. All other combinations correlated with $\rho \leq 0.05$

We measure the Pearson correlation between persuasiveness and the relative frequency of elementary unit 1–3-grams, where each sentence of a debater's comment is assigned one unit. We use the five elementary units testimony, fact, value, policy, and rhetorical statement proposed by Egawa et al. (2019) for CMV comments. We determine the elementary unit of a sentence with a BERT-based classifier trained on Egawa et al. (2019)'s annotated dataset of CMV comments and original posts; The classifier reaches a 6-class (including *None*) micro-accuracy of 0.75 on the standard split. Since the dataset annotates units on a token level, we assign each sentence the unit assigned to its tokens, discarding sentences with multiple units annotated.

We also measure the Pearson correlation between persuasiveness and the relative frequency of 1–3-grams of claim and premise types, where each sentence of a debater's comment is assigned one type. We use the 2-stage classification scheme proposed by Hidey et al. (2017) for CMV comments. Each sentence is first classified with a BERT model as claim, premise, or neither. Claims are then classified as interpretation, evaluation/rational, evaluation/emotional, or agreements. Premises are classified into eight classes, one for each combination of ethos, logos, and pathos using three binary classifiers. We trained each of the five needed classifiers on Hidey et al. (2017)'s datasets of CMV discussions.

Frames Figure 3 shows how often debaters with different persuasiveness use certain frames in their comments. Most frames are used equally often independently of persuasiveness, except for the *political* and *cultural identity* frames, which are used notably more often by better debaters.

We determined frames by classifying each sentence of each comment of a debater with one of the 15 frames used in Card et al. (2015)'s Media Frames corpus of manually annotated news articles. We trained a BERT classifier to classify the sentences, which reaches a micro accuracy of 0.68 in 5-fold random cross-validation.

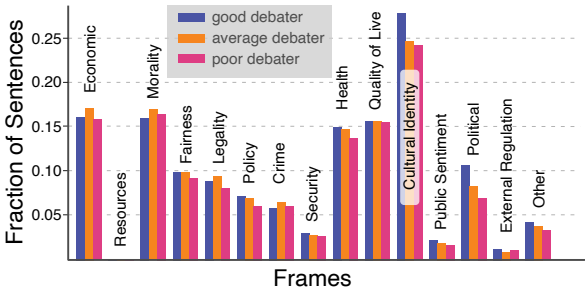


Figure 3: Distribution over the 15 sentence-level frames for good, average, and poor debaters.

6 Debater Persuasiveness Prediction

In addition to the analytical scrutiny of debater persuasiveness, we conduct an experimental validation of our findings by classifying debaters by persuasiveness. We define the general task of debater-level persuasiveness prediction as: Given a debater d with comments $c_1, \ldots, c_n$, classify this debater as persuasive (good) or non-persuasive (average or poor). To conclusively supplement our analysis, we individually inspect the classification performance of the introduced features (cf. Section 5).

We encoded the syntactic, semantic, and pragmatic features of our analysis for each of the 3,801 debaters in our CMV debaters' corpus. Each encoding was chosen to obfuscate comment length as far as reasonable. We encoded the word class and all argumentative unit n-grams tf-idf vectors of the aggregated comments. We encoded the numerical features, outdegree centrality, closeness centrality, and the number of dependents for text complexity and comment-op distance and within-comment distance for WMD, by averaging comment-level counts per debater. We encoded each of the 15 frames with the absolute and relative number of comments that utilize a frame.

As baselines, we selected feature sets previously used for *comment persuasiveness* prediction: Bag-of-Words, vocabulary interplay after (Tan et al., 2016), which covers OP and commenters' vocabularies' absolute and relative overlap and Jaccard similarity, and common stylometrics, which cover counts of words, selected word classes, links, word lists, symbols, type-token ratio, and readability scores. The baseline feature sets were implemented trivially following the related work.

We consider two binary classification settings for our experimental validation: (1) good vs. average and (2) good vs. poor. We maintained a balanced distribution of the classes (1,267 each). The

Features	vs Average	vs Poor
Baseline Features		
Bag of Words	0.60	0.68
Stylometry	0.62	0.67
Vocabulary Interplay	0.58	0.67
Syntactic Features		
Word class n-grams	0.57	0.51
Text Complexity	0.65	0.61
Semantic Features		
Word Mover's Distance	0.59	0.63
Pragmatic Features		
Elementary Units	0.51	0.59
Claim or Premise	0.47	0.55
Claim Type	0.48	0.58
Premise Type	0.48	0.58
Claim and Premise Types	0.48	0.58
Frames	**0.70**	**0.72**

Table 4: Macro F1 score of the two classification settings: Good vs. Average debaters and Good vs. Poor debaters.

classification is done using logistic regression with default parameters on a random 80-20 train-test split. The effectiveness of the classifiers is reported using macro F1-score as shown in Table 4.

The classification results reveal several findings: First, most features distinguish good from poor debaters better than good from average ones. Syntactic features are the only exception to this trend, which can not be explained by our analysis. Second, Bag-of-words is a strong feature for the two settings as it outperforms most of the other features. Besides, the weak effectiveness of the argumentative features is similar to the observations of Egawa et al. (2019); the mere distribution of argumentative units in the text is insufficient to identify its persuasiveness. Third, the distribution of the frames in the debaters' comments results in the best scores across the two experimental settings. The most discriminating frames are 'Quality of Life', 'Morality', and 'Health and Safety', all with negative weights towards the 'good debater' class.

7 Conclusion and Discussion

The persuasion skills of debaters can vary depending on different cultural and social aspects, among others. Understanding how people argue and what makes some debaters more successful than others are interesting research questions that have been neglected in the literature. This paper has contributed in this regard by modeling debater effectiveness in CMV and analyzing their behavior and argumentative stylistic choices, demonstrating several interesting insights that can be utilized for improving the persuasion skills of new debaters and assessing the development of advanced text generation and writing assistant tools.

Still, we think there is room for further analysis. First, we quantified the persuasiveness of CMV debaters based on awarded Δs only; Although it appears to be a standard method in previous work, we believe that a more comprehensive quantification, possibly using human judgments and a more fine-grained scale, would account for a degree of subjectivity to consider the evaluating user's idiosyncrasies. Guo et al. (2020) touches on this briefly, finding that despite general agreement about what is persuasive, there are differences in the assessment of persuasion based on the positions of the evaluating party.

Second, while argumentative features based on the distribution of argumentative units did not perform well in our prediction task, possible improvements can be achieved through modeling features that can capture the effective use of argumentative units. A possible direction is to identify the interdependencies between the different argumentative units in the text (Li et al., 2020) as well as their relative arrangement (Hidey et al., 2017).

Third, other, more in-depth features can disclose useful insights into debater persuasiveness. Conceivable are features that better model behavior like experience and the dynamic of debater interaction or the velocity of experience gain.

Fourth, using more sophisticated models in the prediction task may lead to better results, although our logistic regression is ideal to compare class separability by feature. Guo et al. (2020) proposed using conditional random fields (CRF) to model the cumulative effect of persuasion in CMV discussions, and Li et al. (2020) used bi-LSTM and BERT to model their persuasiveness task.

8 Impact Statement

In a broader context, the findings of this work support the detection, writing, or generation of highly persuasive text, particularly in a social media register. This capability can be abused to generate highly persuasive and misleading, deceptive, fake, or abusive text. Hence, knowledge about debater persuasiveness bears the potential for more persuasive believable social bots. Our work, however, with its focus on the analytical side, bears the same potential to detect generated, hyper-persuasive text.

As with all author-level work on social media, our methods bear the potential to profile users of social media platforms and use the information against them, for example, to automatically block or downvote poor debaters, track them across websites, and possibly reveal their identity. We have taken care to anonymize the user's IDs in our dataset and not release any models that would (re-)generate information about them.

References

Khalid Al-Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7067–7072. Association for Computational Linguistics.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. 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 2: Short Papers*, pages 438–444. The Association for Computer Linguistics.

Ryo Egawa, Gaku Morio, and Katsuhide Fujita. 2019. Annotating and analyzing semantic role of elementary units and relations in online persuasive arguments. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop*, pages 422–428. Association for Computational Linguistics.

Zhen Guo, Zhe Zhang, and Munindar P. Singh. 2020. In opinion holders' shoes: Modeling cumulative influence for view change in online argumentation. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 2388–2399. ACM / IW3C2.

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.

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, ArgMining@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pages 11–21. Association for Computational Linguistics.

Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. 2015. From word embeddings to document distances. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 957–966. JMLR.org.

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. *CoRR*, abs/2010.03538.

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.

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.

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

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.

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.

Yizhi Wang, Yuwan Dai, Hao Li, and Lili Song. 2021. Social media and attitude change: Information booming promote or resist persuasion? *Frontiers in Psychology*, 12.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. The Association for Computer Linguistics.

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.