Persuasion of the Undecided: Language vs. the Listener

Liane Longpré
Cornell University
lfl42@cornell.edu

Esin Durmus
Cornell University
ed459@cornell.edu

Claire Cardie
Cornell University
cardie@cs.cornell.edu

Abstract

This paper examines the factors that govern persuasion for *a priori* UNDECIDED versus DECIDED audience members in the context of on-line debates. We separately study two types of influences: *linguistic factors* — features of the language of the debate itself; and *audience factors* — features of an audience member encoding demographic information, prior beliefs, and debate platform behavior. In a study of users of a popular debate platform, we find first that different combinations of linguistic features are critical for predicting persuasion outcomes for UNDECIDED versus DECIDED members of the audience. We additionally find that audience factors have more influence on predicting the side (PRO/CON) that persuaded UNDECIDED users than for DECIDED users that flip their stance to the opposing side. Our results emphasize the importance of considering the undecided and decided audiences separately when studying linguistic factors of persuasion.

1 Introduction

Understanding the factors that influence persuasion in the context of argumentation (e.g. debates) has been an important focus in a variety of research areas. Natural language processing (NLP) research on persuasion has focused for the most part on uncovering the *linguistic factors* that determine and define persuasive arguments — features of the language of the argument itself. For example, Tan et al. (2016) and Zhang et al. (2016) have found that the language used in arguments and the patterns of interaction between debaters are important predictors of persuasiveness. Recently, however, studies have emerged that begin to study the effects of *audience characteristics* on persuasion, e.g. features that encode demographic information, the prior beliefs, and debate platform behavior of individual listeners of a debate or readers of an argument. Lukin et al. (2017), for example, find that different types of people are persuaded by different types of arguments. And Durmus and Cardie (2018) show that the prior beliefs of the audience have a significant impact on predicting whether or not a particular audience member will be persuaded to flip their stance on a debated topic.

Research in psychology and political science moreover suggests that there are key differences in the persuasion of undecided versus decided voters/audience members. For example, Petty and Cacioppo (1996) find that prior experiences and beliefs can lead to the re-framing of a message perceived by a person to maintain consistency between their prior beliefs and their attitudes towards the topic of the message. In particular, studies show that *a priori* decided voters simply ignore certain information in order to maintain this consistency (Sweeney and Gruber, 1984; Vecchione et al., 2013; Kosmidis, 2014). In contrast, an undecided voter is asked to make a decision on an issue for which previously received information was somehow unconvincing; and Kosmidis (2014), Kosmidis and Xezonakis (2010), and Schill and Kirk (2014) show that, as a result, these voters are likely to rely heavily on information conveyed in a new message.

The undecided voter group furthermore holds the highest potential for persuasion (Kosmidis and Xezonakis, 2010; Shehryar et al., 2017). Public support for social and political causes often critically depends on the undecided decision makers. To the best of our knowledge, computational studies of persuasion in NLP have not yet studied this important subset of the audience separately.

This paper studies argumentation in the context of online debate to better understand the factors that govern persuasion for *a priori* UNDECIDED

Proceedings of the 6th Workshop on Argument Mining, pages 167–176
Florence, Italy, August 1, 2019.

versus DECIDED members of the audience. We study persuasion at the individual (i.e. audience member) level, and find that the linguistic features most important for persuasion differ for the UNDECIDED and DECIDED audience subgroups. Consistent with results of social and political psychology research, the linguistic feature differences correspond to rhetorical styles found to be effective on undecided and decided audiences. Additionally, we find that certain audience features are more important for predicting undecided cases of persuasion than for predicting decided cases of persuasion.

The remainder of this paper is organized as follows. Related work is described in Section 2. We describe the dataset in Section 3 and experiment methodology in Section 4. Results and analysis is in Section 5, and conclusions are in Section 6.

2 Related Work

Language and persuasion. Extensive work has been done in cognitive and social psychology on the linguistic influence on persuasion. Some of the most critical elements of persuasive text include lexical complexity, language intensity, and power of speech style (Dillard and Pfau, 2002). Studies on linguistic factors effecting the persuasion of the listener have shown that language is a key factor in predicting the outcome of debates (Paxton and Dale, 2014; Jorgensen et al., 1998). These studies find the importance of various language features: lexical qualities such as personal pronoun use, word sentiment, and hedging (Paxton and Dale, 2014), and rhetoric qualities such as precision, firmness, energy, and commitment (Jorgensen et al., 1998). These works in psychology highlight the importance of studying linguistic features in arguments and persuasion.

Argument mining. Much recent work in argumentation has focused on the automatic detection of argument structures in text (Lippi and Torroni, 2016; Schulz et al., 2018; Stab et al., 2018; Morio and Fujita, 2018). Research has shown promising results on using extracted argument structures as features on tasks that involve predicting convincingness (Ghosh et al., 2016; Yunfan Gu and Huang, 2018; Cano-Basave and He, 2016).

Specific to debates, work has been done on detecting the stance of the speaker. Walker et al. (2012), for example, find that structuring the debates in terms of agreement relations between speakers improves prediction. Lexical and syntactic argument features are shown to improve predictive performance in Somasundaran and Wiebe (2010). More relevant to our work, recent studies have examined the role of language in predicting persuasion outcomes in debates. For example, Tan et al. (2016) find that the linguistic interaction between an opinion holder and opposing debater are highly predictive of persuasiveness. And Zhang et al. (2016) find that debaters who target and address their opponent's points are more likely to win the debate.

While these studies motivate the linguistic features examined in our study, they do not take factors corresponding to audience characteristics into consideration. Our work aims to study the linguistic characteristics of persuasive text, while also considering audience characteristics such as prior beliefs and decidedness.

Prior views of the audience. Persuasion of an audience is not solely dependent on the language used by the speaker. Research in psychology emphasizes the significance of people's prior views on their perception of new information. The effectiveness of a message depends significantly on the prior beliefs and the strengths of beliefs of the message recipient (Johnson et al., 1995; Lau et al., 1991).

Recent work has analyzed the influence of audience characteristics on predicting persuasion (Lukin et al., 2017; Durmus and Cardie, 2018). Lukin et al. (2017) examine the effects of audience factors and argumentation types in belief change. They study dialogs from *4forums.com*[1], which contain argument type annotations. Their results show that information on prior beliefs and personality type improves the ability of the model to predict belief change; more conscientious, open, and agreeable people tend to respond more to emotional argument types.

The importance of considering audience-specific prior belief factors is further illustrated in Durmus and Cardie (2018). Using debate and user data from *debate.org*, they study the effects of prior beliefs on various controversial issues along with linguistic factors on predicting the outcome of debates. Importantly, they find that the linguistic features most important for prediction differ when audience features are considered from when

[1] http://www.4forums.com/political/forum.php/

they are not. To the best of our knowledge, this work is most relevant to ours because it studies debate text and considers prior beliefs of both the audience and the debaters. Our work differs from this study in that we separately consider persuasion of audience members who were undecided before the debate from audience members who switched sides.

The undecided audience. There has been a substantial amount of research effort in the social and political sciences on undecided and decided voters. A study on the 2005 British general election finds that undecided voters are more susceptible to campaign persuasion (Kosmidis and Xezonakis, 2010). This result, elaborated on in Kosmidis (2014), is because decided voters rely more on their prior beliefs while undecided voters place higher weight on information conveyed in campaigns.

Consistent with this account, studies by Schill and Kirk (2014) on 2008 and 2012 U.S. presidential debate outcomes find that the most critical portions of the debate to undecided voters were the content-rich statements, and that the rhetorical strategies shown to be effective to undecideds are strategies that "transcended the personalities of the candidates". In contrast, studies by Adams et al. (2011) on European election campaigns find that in response to policy statements of political parties during elections, voters adjust their Left-Right positions based on their subjective perceptions of the party's campaign and not on the campaign's actual policy statements. Research on selective exposure (favoring information that aligns with an individual's prior beliefs and attitudes) provides insight into the mechanisms behind this tendency. Voters already decided on an issue tend to avoid information that is inconsistent with their attitudes and are receptive to information consistent with their attitudes (Sweeney and Gruber, 1984; Vecchione et al., 2013).

3 Data Description

The debate dataset from Durmus and Cardie (2018) consists of 67,315 debates and user information on 36,294 users obtained from *debate.org*.

3.1 Debates

Debates span over 23 different categories (e.g. 'Politics', 'Education', 'Movies'). Each debate consists of multiple rounds, where a round contains text from the PRO debater and the CON debater. An example debate is shown in Table 1. Other examples of debate titles are: "THE DEATH PENALTY IS A SUITABLE PUNISHMENT" and "ANIMAL TESTING SHOULD BE BANNED".

ROUND 1	
PRO:	... this reason, you are not free to make threats or defamatory statements against another person in ...
CON:	... laws violate the fundamental freedom of speech which democracy is founded upon ...
ROUND 2	
PRO:	... has ignored my point about hate speech breeding an "us vs them" mentality, and how such ...
CON:	... question is, does our government have the right to tell us what our opinions are, and to define ...
ROUND 3	
PRO:	... evidenced by the rise in violence against Hispanics and Muslims I cited in my second round ...
CON:	... courts to be able to decide which opinions are "moral" and which are not? How fascist do ...

Table 1: An example debate titled 'HATE SPEECH LAWS ARE A GOOD IDEA'.

Users can interact with debates by voting on them. Votes include "AGREE WITH BEFORE THE DEBATE" and "AGREE WITH AFTER THE DEBATE" for each debater/side (users can respond with PRO, CON, or TIE). We focus our analysis on two distinct cases of persuasion based on this vote data.

Case 1: voters persuaded from the middle. This category constitutes voters who indicate TIE between PRO and CON for "AGREE WITH BEFORE THE DEBATE" and indicate one side, PRO or CON, for "AGREE WITH AFTER THE DEBATE". We keep instances of persuasion that correspond to this category and refer to this case as FROM-MIDDLE.

Case 2: voters persuaded from the opposite side. This category constitutes voters who indicate one side for "AGREE WITH BEFORE THE DEBATE" and indicate the opposite side (PRO or CON) for "AGREE WITH AFTER THE DEBATE". We keep instances that correspond to this category, referred

Persuasion Case	#instances	#debates
FROM-MIDDLE	4360	3652
FROM-OPPOSING	2642	2183

Table 2: Dataset statistics.

to as FROM-OPPOSING. In our prediction task, the original side of the voter is not given to the model.

Figure 1 illustrates example user votes for each of the two cases. Distinguishing instances of voters being persuaded into these case groupings allows us to examine what makes an argument persuasive to audience members who are undecided versus decided with respect to a particular debate topic. Table 2 summarizes the dataset statistics relevant to the voter cases.

3.2 User Information

User profiles contain self-identified demographic information, such as GENDER and RELIGIOUS IDEOLOGY. Profiles additionally contain users' opinions on current controversial debate topics (denoted by BIG-ISSUES), such as ABORTION, SOCIAL SECURITY, and MINIMUM WAGE[2]. Users can respond with PRO (in favor), CON (against), UND (undecided), N/O (no opinion), or N/S (not saying).

4 Prediction Task

We aim to study what factors are most important in influencing audience members to be persuaded to one side or the other for each of the cases (*a priori* undecided or decided) of persuasion. Encoding audience-level and linguistic factors as features, we structure the prediction task as follows:

> Given an individual voter, predict which debater/side (PRO or CON) the voter will be convinced by after the debate.

We consider only samples from the data where (1) a voter was undecided before the debate and then adopted a stance, i.e. voted for one of the debaters as the winner; and (2) a voter was (seemingly) decided beforehand and then flipped their stance. We do *not* consider samples where (1) a voter declared a "tie" between the debaters after the debate; and (2) a voter was decided beforehand, and voted for the debater with the stance that they agreed with beforehand.

To study the effect of each of the debaters' linguistic and user-based features on persuasion, in this setting, we specifically look at which side (PRO vs. CON) did the convincing for a particular voter. We believe that restricting the samples in the way described above allows us to best study what influences persuasion when voters are successfully convinced.

4.1 Features

Audience features. User profile data is used to generate a number of features for a voter and the PRO and CON debaters for a given debate.

The *gender* of a voter is one-hot encoded to account for the user's option to not include gender in their profile; the elements of the vector correspond to FEMALE, MALE, and OTHER/DID NOT INDICATE. Additionally, information about the debaters' genders are encoded as whether or not the debater's gender is the same as the voter's.

User profile data is also used to capture the prior opinion similarities of the voter and debaters in two ways, as in Durmus and Cardie (2018). First, the political and religious ideologies are encoded as whether or not each of the debaters' ideologies is the same as each of the voter's. We denote this feature by *matching ideology*. Second, the similarity of the voter and debaters' BIG-ISSUES responses are encoded as follows. Each issue in BIG-ISSUES is represented as a one-hot encoding corresponding to PRO, CON, UND, and N/O. The encoding of an example user can be seen in Figure 2. All issue encodings are concatenated to create a BIG-ISSUES vector for each user. The cosine similarity between the voter's BIG-ISSUES vector and each debaters' BIG-ISSUES vector is used as a feature. We denote this feature by *opinion similarity*.

The number of elements in the voter's BIG-ISSUES vector corresponding to PRO and CON, and the number of elements in the vector corresponding to UND and N/O are used to encode the voter's decidedness or undecidedness, respectively. We denote the feature by *decidedness*. An example of the encoding is shown in Figure 2. This feature captures the degree to which the voter's opinions are established on widely discussed topics.

The frequency of a voter being persuaded is encoded as the percentage of other training debates in which the voter changed their stance, out of all

[2] https://www.debate.org/big-issues/

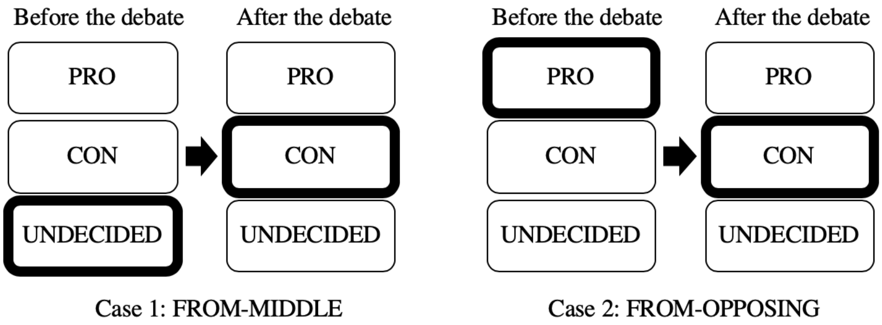
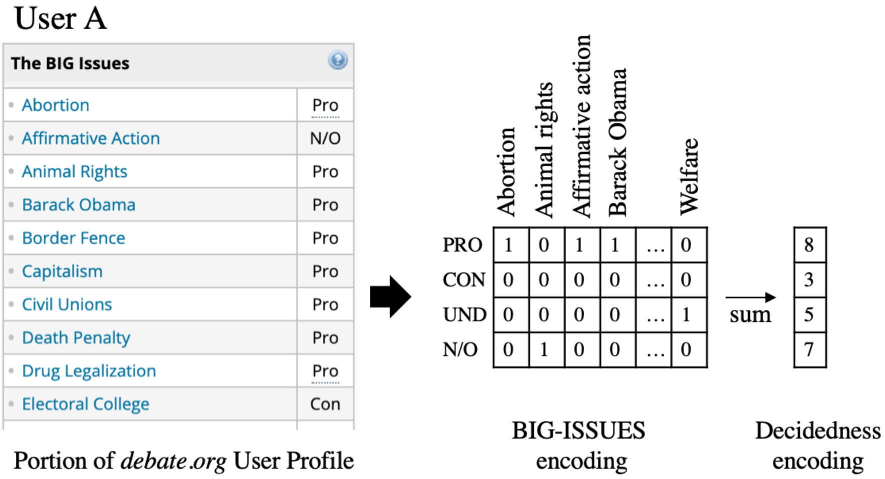


Figure 1: Example votes for a debate showing each case of persuasion.


User A

The BIG Issues	
Abortion	Pro
Affirmative Action	N/O
Animal Rights	Pro
Barack Obama	Pro
Border Fence	Pro
Capitalism	Pro
Civil Unions	Pro
Death Penalty	Pro
Drug Legalization	Pro
Electoral College	Con

Portion of *debate.org* User Profile

	Abortion	Animal rights	Affirmative action	Barack Obama		Welfare
PRO	1	0	1	1	...	0
CON	0	0	0	0	...	0
UND	0	0	0	0	...	1
N/O	0	1	0	0	...	0

BIG-ISSUES encoding

8
3
5
7

Decidedness encoding

Figure 2: Example user profile and corresponding feature encodings.

training debates on which the voter made a vote. We denote the feature by *persuadability*. This feature is an indication of how persuadable a voter is, in general.

Linguistic features. We process debate text and use linguistic features as is done in Durmus and Cardie (2018). The text from all rounds of PRO are concatenated before feature processing. The same is done for the rounds of CON. We use the same set of linguistic features from Durmus and Cardie (2018), described as follows.

Lexical features include *TF-IDF*, *modal verbs*, *swear words*, *spelling errors*, and *punctuation*. A speaker's word choice (i.e. use of hedging, and particular causal connectors and modal particles) are indicative of the mode of argumentation (Gold et al., 2015; Paxton and Dale, 2014).

Style features include *length*, *personal pronouns*, *referring to opponent*, *use of citations*, and *links*. Using citations and addressing an opponent's points are critical components of justification that affect the reception of an argument. Additionally, the length of a speaker's utterance and the language used when referring to self and the opponent exhibit characteristics of respect and participation between the debaters, which are important aspects for communication outcomes (Tan et al., 2016; Gold et al., 2015; Paxton and Dale, 2014).

Semantic features include *sentiment*, *subjectivity* (Wilson et al., 2005), *connotation* (Feng and Hirst, 2011), and *politeness*. The sentiment and subjectivity of an argument impacts the reception of the message, and are predictive of argument stance (Somasundaran and Wiebe, 2010). In addition to these attributes, connotation and politeness cues contribute to the patterns of interaction of debaters, which are critical in predicting persuasiveness (Tan et al., 2016).

Argumentation features, as in (Somasundaran et al., 2007), have been shown to predict the stance and opinion of a speaker. These include the following: *assessment, authority, conditioning, contrasting, emphasizing, generalizing, empathy, inconsistency, necessity, possibility, priority, rhetorical questions, desire,* and *difficulty*.

4.2 Hypotheses

We hypothesize that there are key differences in the linguistic features important for persuasion of an *a priori* undecided audience member and the persuasion of an *a priori* seemingly decided audience member to change their mind. Drawing from social and political science studies, we hypothesize that the persuasion of undecided audience members will critically depend on content-centric language features, while the persuasion of seemingly decided audience members will be more influenced by stylistic language features. Additionally, we hypothesize that audience features will provide important context, improving predictive performance.

4.3 Methodology

We use Logistic Regression to perform the classification task. Prediction accuracy is evaluated using 5-fold cross validation. We use 3-fold cross validation on the training set to select model parameters. We perform ablation analysis first on audience features only and linguistic features only, then on combinations of the best-performing audience and linguistic features. This analysis is done separately for the subsets of data corresponding to undecided and decided cases of persuasion (FROM-MIDDLE and FROM-OPPOSING, respectively). We use majority classifier as a baseline.

5 Results and Analysis

Results for models and feature ablation experiments are show in Table 3. Majority baseline produces 57.43% and 59.42% accuracy for FROM-MIDDLE and FROM-OPPOSING, respectively. This baseline predicts the majority debater/side between PRO and CON in the training set of examples.

Accuracy of Models

	FROM-MIDDLE	FROM-OPPOSING
Majority Baseline	57.43%	59.42%
All Features	**69.01%**	**67.22%**
Audience Features	**61.47%**	**61.54%**
- persuadability	61.46%	61.51%
- gender	61.44%	61.47%
- matching ideology	61.42%	61.39%
- decidedness	61.33%	61.13%
- opinion similarity	59.04%	59.80%
Linguistic Features	**66.95%**	**66.65%**
- unigram TF-IDF	65.25%	64.54%
- use of citations and referring to opponent	67.20%	66.12%
- subjectivity	66.03%	67.79%

Table 3: **Accuracy results, for majority class baseline, all features, audience features, and linguistic features.** Remaining results are ablation studies, where '- *feature*' denotes the removal of the feature. Underlined results are feature combinations that improve performance over including all features.

Linguistic vs. audience features. As shown in Table 3, the best performance is achieved when both audience and linguistic features are included, obtaining 69.01% and 67.22% accuracy for FROM-MIDDLE and FROM-OPPOSING, respectively. We find that linguistic features are more important for predictive accuracy than audience features. Relying only on audience features obtains accuracies of 61.47% for FROM-MIDDLE and 61.54% for FROM-OPPOSING. Using all linguistic features produces a significant improvement over baseline accuracy, achieving 66.95% and 66.65% for FROM-MIDDLE and FROM-OPPOSING, respectively. This result is surprising and in contrast to results from Durmus and Cardie (2018), who find that audience features improve accuracy more than linguistic features. We suspect that this difference arises because our experiments consider debates from *all* categories, while Durmus and Cardie (2018) restrict analysis to *political* and *religious* debate categories. Political and religious debate topics tend to be more controversial in nature Fichman and Hara (2014), and correspond more closely to the issues encoded in the audience features; the BIG-ISSUES elements consist primarily of political and religious issues [3]. As such, these features will be more informative in *political* and *religious* debate settings.

Audience features. Feature ablation across user-based features shows that all audience features are helpful in predicting vote outcomes for both voter groups. We find that the most important feature is *opinion similarity*[4]; removing this feature decreases prediction accuracy from 61.47% to 59.04% for FROM-MIDDLE, and from 61.54% to 59.80% for FROM-OPPOSING. This result is

[3] https://www.debate.org/big-issues/

[4] For UserA and UserB, the cosine similarity of BIG-ISSUES$_A$ and BIG-ISSUES$_B$.

consistent with research on voter behavior from Arcuri et al. (2008) and Friese et al. (2012), who find that despite reporting uncertainty, undecided voters have implicit attitudes that are predictive of voting behavior.

Linguistic features. The most important linguistic feature for both voter groups is *unigram TF-IDF*[5], whose removal decreases performance to from 66.95% to 65.25% for FROM-MIDDLE, and from 66.65% to 64.54% for FROM-OPPOSING. However, not all linguistic features are helpful in predictive accuracy. For instance, removing *use of citations*[6] and *referring to opponent*[7] features increases accuracy from 66.95% to 67.20% for FROM-MIDDLE. Similarly, removal of the *subjectivity*[8] feature improves accuracy for FROM-OPPOSING from 66.65% to 67.79%.

It should be noted that the linguistic features whose removal improves performance for FROM-MIDDLE and FROM-OPPOSING are different, showing that there are distinctions in the important factors for persuasion between the voter groups. These differences are further explored in the following sections.

5.1 Differences Between Persuasion Groups

5.1.1 Linguistic Feature Differences

We find distinct differences in the important features for predicting the vote outcome for voter groups FROM-MIDDLE and FROM-OPPOSING. Table 4 shows that the best-performing set of linguistic features for FROM-MIDDLE includes all features minus *use of citations*, *referring to opponent*, and *swear words*, while the best-performing set of linguistic features for FROM-OPPOSING includes all features minus *subjectivity*, *modals*[9], and *bi-/tri-gram TF-IDF*[10]. These linguistic feature sets are denoted by MIDDLE* and OPPOSING*, respectively. Using features OPPOSING* increases accuracy for FROM-OPPOSING from 67.22% to 68.39%, while decreasing accuracy for FROM-MIDDLE from 69.01% to 68.51%. Conversely, using features MIDDLE* increases accuracy for FROM-MIDDLE from 69.01% to 69.17%, while decreasing accuracy from 67.22% to 66.92%.

Accuracy of Models	FROM-MIDDLE	FROM-OPPOSING
All Features	**69.01%**	**67.22%**
- persuadability	68.33%	67.52%
- matching ideology	68.99%	67.30%
User+MIDDLE*	**69.17%**	**66.92%**
- persuadability	69.16%	66.84%
- matching ideology	68.60%	66.92%
User+OPPOSING*	**68.51%**	**68.21%**
- persuadability	68.46%	68.32%
- matching ideology	67.96%	68.39%

Table 4: **Accuracy results, for all features and best-performing linguistic feature sets.** Remaining results are ablation studies, where '- *feature*' denotes the removal of the feature. Underlined results are feature combinations that improve performance over including all features. MIDDLE* denotes the best-performing combination of linguistic features for FROM-MIDDLE, which includes all linguistic features minus *use of citations*, *referring to opponent*, and *swear words*. OPPOSING* denotes the best-performing combination of linguistic features for FROM-OPPOSING, which includes all linguistic features minus *subjectivity*, *modals*, and *bi-/tri-gram TF-IDF*.

The linguistic feature differences of the two groups have subtle differences in nature. A possible analysis that distinguishes the groups is that there is a difference in the rhetorical strategies most effective for undecided versus decided audiences. Use of modals, subjectivity, and general word choice are semantic features of an argument that affect the perception of the content of the argument. Based on our results, these content-based features are more important for undecided voters than they are for decided voters. In comparison, use of swear words, citing sources, and referring to the opponent are stylistic features of an argument that affect the perception of the debater producing the argument. Based on our results, these style-based features are not as important for undecided voters as they are for decided voters. This account is consistent with the findings of Schill and Kirk (2014) that undecided voters respond most to content-rich rhetorical strategies, and the findings of Vecchione et al. (2013); Sweeney and Gruber (1984) that decided voters tend to selectively attend to information in a message based on prior attitudes. The account is also in line with experiments conducted by Adams et al. (2011), which

[5] Calculated with a maximum of 50 terms.

[6] The number of explicit source citations.

[7] The usage of phrases like "according to my opponent".

[8] Number of words with negative strong, negative weak, positive strong, and positive weak subjectivity.

[9] The usage of modal verbs, i.e. *can*, *should*, *will*, and *may*.

[10] Calculated with a maximum of 30 terms.

find that affiliated voters do not adjust their positions in response to a party's actual policy statements, but rather do adjust their positions based on their subjective perceptions of the party.

5.1.2 Audience Feature Differences

The inclusion of certain audience features has different effects on prediction accuracy between FROM-MIDDLE and FROM-OPPOSING voter groups. As shown in Table 4, removing the *persuadability* feature improves the accuracy for FROM-OPPOSING from 67.22% to 67.52% when all linguistic features are included, and improves the accuracy from 68.21% to 68.32% when OPPOSING* linguistic features are used. Similarly, removing the *matching ideology* feature improves the accuracy for FROM-OPPOSING from 67.22% to 67.30% when all linguistic features are included, and improves the accuracy from 68.21% to 68.39% when OPPOSING* linguistic features are used. The reverse is true for FROM-MIDDLE. For this voter group, removing the *persuadability* and *matching ideology* features decreases accuracy from 69.01% to 68.33% and 68.99%, respectively, when all lingusitic features are included, and decreases the accuracy from 69.17% to 69.16% and 68.60%, respectively, when MIDDLE* features are included.

It should be noted that the best-performing overall feature set for FROM-OPPOSING includes neither the *persuadability* feature nor the *matching ideology* feature. In contrast, all audience features are present in the best-performing overall feature set for FROM-MIDDLE. This difference suggests that certain audience-level aspects are comparatively more predictive of vote outcomes for undecided voters. The result emphasizes the importance of considering audience factors for people who are undecided with respect to an issue; in order to understand vote behavior of the undecided audience, it is critical to consider audience factors.

5.2 Influence of Audience Features

We perform ablation across linguistic features separately for when audience features are included and for when they are not. Results in Table 5 show that the linguistic features most important for model performance differ when audience features are present. For instance, experiments on voter group FROM-OPPOSING show that including *argument lexicon* features improves performance from 67.22% to 67.52% when audience features are not included, while performance is decreased from 66.65% to 65.90% when audience features are included. Comparatively, inclusion of the *swear words* feature improves performance for FROM-MIDDLE from 69.01% to 69.08% when audience features are not included, but negatively impacts performance from 66.95% to 66.65% when audience features are included.

Accuracy of Models

	FROM-MIDDLE	FROM-OPPOSING
Linguistic Features	**66.95%**	**66.65%**
- argument lexicon	66.22%	65.90%
- use of citations and referring to opponent	67.20%	66.12%
- swear words	66.65%	66.65%
- subjectivity	66.03%	67.79%
All Features	**69.01%**	**67.22%**
- argument lexicon	68.46%	67.52%
- use of citations and referring to opponent	69.17%	66.99%
- swear words	69.08%	67.20%
- subjectivity	68.76%	67.90%

Table 5: **Accuracy results, for all features and linguistic features.** Remaining results are ablation studies, where '- *feature*' denotes the removal of the feature. Underlined results are feature combinations that improve performance over including all features.

We find that the best-performing sets of linguistic features for FROM-OPPOSING and FROM-MIDDLE differ when audience features are included versus when they are not. The best-performing set of linguistic features for FROM-OPPOSING when audience features are not considered includes *modals* and *bi-/tri-gram TF-IDF*, while these features are not present in the best-performing set of features when all features are considered (denoted by OPPOSING*). Similarly for FROM-MIDDLE, the *swear words* feature is not in MIDDLE*, while it is present in the best-performing set of linguistic features when audience features are not considered.

These results are consistent with findings from Durmus and Cardie (2018) and re-affirm the importance of considering audience features when analyzing linguistic effects of persuasion.

6 Conclusion

In this paper, we separately examine what linguistic and audience-level factors are most important for predicting vote outcomes of previously undecided and decided audiences. We show that different linguistic features are critical for predicting the successful side of persuasion of undecided versus decided voters. We find that some audience features that are important for predicting the side of persuasion of undecided voters are not as helpful in predicting persuasion of decided voters.

This paper examines the differences between the undecided and decided audiences in persuasion, which has been under-studied in a computational framework. The results of our work validate the importance of analyzing the undecided versus decided audience separately.

Acknowledgments

This work was supported in part by NSF grants IIS-1815455 and SES-1741441. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NSF or the U.S. Government.

References

James Adams, Lawrence Ezrow, and Zeynep Somer-Topcu. 2011. Is anybody listening? Evidence that voters do not respond to European parties' policy statements during elections. *American Journal of Political Science*, 55(2):370–382.

Luciano Arcuri, Luigi Castelli, Silvia Galdi, Cristina Zogmaister, and Alessandro Amadori. 2008. Predicting the vote: Implicit attitudes as predictors of the future behavior of decided and undecided voters. *Political Psychology*, 29(3):369–387.

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

James P. Dillard and Michael Pfau. 2002. *The Persuasion Handbook: Developments in Theory and Practice*, pages 371–380. Sage Publications, Inc., Thousand Oaks, CA.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2018*, pages 1035–1045. Association for Computational Linguistics.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 987–996. Association for Computational Linguistics.

Pnina Fichman and Noriko Hara. 2014. *Global Wikipedia: International and Cross-Cultural Issues in Online Collaboration*, pages 25–41. Rowman & Littlefield Publishers, Lanham, Maryland.

Malte Friese, Colin Tucker Smith, Thomas Plischke, Matthias Bluemke, and Brian A. Nosek. 2012. Do implicit attitudes predict actual voting behavior particularly for undecided voters? *Public Library of Science One*, 7:1–14.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 549–554. Association for Computational Linguistics.

Valentin Gold, Mennatallah El-Assady, Annette Hautli-Janisz, Tina Bgel, Christian Rohrdantz, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2015. Visual linguistic analysis of political discussions: Measuring deliberative quality. *Digital Scholarship in the Humanities*, 32(1):141–158.

Blair T. Johnson, Hung-Yu Lin, Cynthia S. Symons, Laura Ann Campbell, and Geoffrey Ekstein. 1995. Initial beliefs and attitudinal latitudes as factors in persuasion. *Personality and Social Psychology Bulletin*, 21(5):502–511.

Charlotte Jorgensen, Christian Kock, and Lone Rorbech. 1998. Rhetoric that shifts votes: An exploratory study of persuasion in issue-oriented public debates. *Political Communication*, 15(3):283–299.

Spyros Kosmidis. 2014. Heterogeneity and the calculus of turnout: Undecided respondents and the campaign dynamics of civic duty. *Electoral Studies*, 33:123 – 136.

Spyros Kosmidis and Georgios Xezonakis. 2010. The undecided voters and the economy: Campaign heterogeneity in the 2005 British general election. *Electoral Studies*, 29(4):604 – 616.

Richard R. Lau, Richard A. Smith, and Susan T. Fiske. 1991. Political beliefs, policy interpretations, and political persuasion. *The Journal of Politics*, 53(3):644–675.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology*, 16(2):10:1–10:25.

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

Gaku Morio and Katsuhide Fujita. 2018. End-to-end argument mining for discussion threads based on parallel constrained pointer architecture. In *Proceedings of the 5th Workshop on Argument Mining*, pages 11–21, Brussels, Belgium. Association for Computational Linguistics.

Alexandra Paxton and Rick Dale. 2014. Leveraging linguistic content and debater traits to predict debate outcomes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, pages 592–596. Cognitive Science Society.

Richard E. Petty and John T. Cacioppo. 1996. *Attitudes and Persuasion: Classic and Contemporary Approaches*, pages 95–160. Westview Press, New York, NY.

Dan Schill and Rita Kirk. 2014. Courting the swing voter: "Real time" insights into the 2008 and 2012 U.S. presidential debates. *American Behavioral Scientist*, 58(4):536–555.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multitask learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41, New Orleans, Louisiana. Association for Computational Linguistics.

Omar Shehryar, Kelly Weidner, and Dan Moshavi. 2017. Persuading the undecided: An interdisciplinary approach to increase public support for the arts. *Journal of Public Affairs*, 18(2):e1652.

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 the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Paul D. Sweeney and Kathy L. Gruber. 1984. Selective exposure: Voter information preferences and the Watergate affair. *Journal of Personality and Social Psychology*, 46(6):1208–1221.

Chenhao Tan, Vlad Niculae, Cristian DanescuNiculescu-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*, pages 613–624. International Conference on World Wide Web.

Michele Vecchione, Gianvittorio Caprara, Francesco Dentale, and Shalom H. Schwartz. 2013. Voting and values: Reciprocal effects over time. *Political Psychology*, 34(4):465–485.

Marilyn A. Walker, Pranav Anand, Robert Abbott, and Ricky Grant. 2012. Stance classification using dialogic properties of persuasion. In *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596. Association for Computational Linguistics.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics.

Maoran Xu Hao Fu Yang Liu Yunfan Gu, Zhongyu Wei and Xuanjing Huang. 2018. Incorporating topic aspects for online comment convincingness evaluation. In *Proceedings of the 5th Workshop on Argument Mining*, pages 97–104. Association for Computational Linguistics.

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