TOWARDS UNDERSTANDING PERSUASION IN COMPUTATIONAL ARGUMENTATION

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Esin Durmus

May 2021

TOWARDS UNDERSTANDING PERSUASION IN COMPUTATIONAL ARGUMENTATION

Esin Durmus, Ph.D.

Cornell University 2021

Opinion formation and persuasion in argumentation are affected by three major factors: the argument itself, the source of the argument, and the properties of the audience. Understanding the role of each and the interplay between them is crucial for obtaining insights regarding argument interpretation and generation. It is particularly important for building effective argument generation systems that can take both the discourse and the audience characteristics into account. Having such personalized argument generation systems would be helpful to expose individuals to different viewpoints and help them make a more fair and informed decision on an issue.

Even though studies in Social Sciences and Psychology have shown that source and audience effects are essential components of the persuasion process, most research in computational persuasion has focused solely on understanding the characteristics of persuasive language. In this thesis, we make several contributions to understand the relative effect of the source, audience, and language in computational persuasion. We first introduce a large-scale dataset with extensive user information to study these factors' effects simultaneously. Then, we propose models to understand the role of the audience's prior beliefs on their perception of arguments. We also investigate the role of social interactions and engagement in understanding users' success in online debating over time. We find that the users' prior beliefs and social interactions play an essential role

in predicting their success in persuasion. Finally, we explore the importance of incorporating contextual information to predict argument impact and show improvements compared to encoding only the text of the arguments.

BIOGRAPHICAL SKETCH

Esin was born in Bursa, Turkey. She obtained a Bachelor's degree in Industrial Engineering and Computer Engineering at Koç University, where she was the valedictorian of her class. She was introduced to computers at an early age since her father is a Computer Engineer, and he always motivated her to follow his path. During her undergrad, she explored several research areas such as Computer Verification and Intelligent User Interfaces. However, she was not aware of Natural Language Processing (NLP) research until she started her Ph.D. at Cornell.

During her Ph.D., she took several courses in Natural Language Processing to better understand the state of research. She was very impressed that NLP provided an opportunity to integrate machine learning methods to understand social phenomena such as persuasion. She spent a lot of time thinking about computationally modeling persuasion during her Ph.D. She has always valued collaborations since she learned a lot from other researchers in this field. She did internships at Amazon and Google, where she met a fantastic group of people and worked on text generation and evaluation which motivated her to do a Postdoc at Stanford University to explore these areas further. Apart from research, Esin enjoys watching movies, traveling, trying food from different cuisines, and playing table tennis.

To my parents and Faisal for being there through the journey.

ACKNOWLEDGEMENTS

First and foremost, I am incredibly grateful to my advisor, Claire Cardie, for her support, kindness, patience, and guidance. I was not familiar with Machine Learning (ML) and Natural Language Processing (NLP) until I started my Ph.D. During my first year, I took the "Introduction to Natural Language Processing" course taught by Claire. That course was instrumental in my decision to pursue a career in the field. Claire has always been very patient and understanding. She always gave me the time and opportunity to learn and explore new research directions. I am forever grateful to have had her support during the beginning of my career. I am also very thankful to my committee members John Hopcroft and Cristian Danescu-Niculescu-Mizil for their feedback on my research.

I was fortunate to do great internships at Amazon and Google during my Ph.D. I want to thank my collaborators and mentors, He He, Mona Diab, Kathleen McKeown, Smaranda Muresan at Amazon AWS, and Gaurav Kumar, Ankur Parikh, Diyi Yang, Sebastian Gehrmann, Dipanjan Das at Google Research, who helped me gain a more diverse perspective about research in NLP. I am incredibly grateful to He He and Mona Diab for supporting and guiding me in my career decisions beyond the internship. I am fortunate to have their mentorship.

I want to thank Ana Smith, Arzoo Katiyar, Ashudeep Singh, Jialu Li, Kai Sun, Liane Longpre, Maria Antoniak, Ozan Irsoy, Rishi Bommasani, Tianze Shi, Vlad Niculae, Xilun Chen, Xinran Zhao, Xinya Du, Yao Cheng, and all the members of the NLP seminar for fruitful discussions and providing feedback on my research.

I have been fortunate to be surrounded by amazing friends. Just to name a few: Deniz Altınbüken, Burcu Çanakçı, Connie Ong Blukis, Valts Blukis, Kim

Bomin, Milla Vastavuo, Parsifal Islas, Ilse Alejo, Pablo Lujambio, Faisal Alkaabneh, Nouha Dziri, Amr Sharaf, Tuhin Chakrabarty, Yiqing Hua, Andreas Veit, Eugune Bagdasaryan, Nadia Bagdasaryan, Julia Tolkacheva, Ilia Ilmer and Nidhi Baid. I am forever grateful to my friends who have been instrumental in keeping me motivated and pulling me out of stressful moments of life. I would not have been able to do any of this without them.

I want to thank my parents: Eşref Durmuş and Tülay Durmuş. They have always motivated me to pursue a career in Computer Science and given me tremendous support throughout this journey. I can never pay back their support.

Lastly, I would like to thank Faisal Ladhak for his endless love and support. I have been very fortunate to have him by my side throughout this journey. Through these years, we experienced life in four different cities (Ithaca, Seattle, NYC, and Mountain View), and I enjoyed every moment of my time with him. Besides being my life partner and best friend, he has also been a fantastic collaborator during these years. He always inspires me and brings the best out of me. I am forever grateful to him.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

Argumentation is a discussion in which reasons are provided for and against some proposition or proposal (Toulmin, 1958). It is a crucial activity in decision-making since it encourages critical thinking and motivates people to make fair and informed decisions. The emergence of social media and online argumentation platforms has made it easier for people to express their opinions and debate with other individuals on controversial topics. The reliance on social media and online argumentation platforms as key venues for opinionated discussions (Shearer and Matsa, 2020) has motivated increased research in computational argumentation. One area of focus in computational argumentation has been to explore techniques to automatically analyze the characteristics of arguments, such as their persuasiveness (Habernal and Gurevych, 2016a; Hidey et al., 2017; Tan et al., 2016; Wachsmuth et al., 2017b). Furthermore, researchers have started building systems to automatically generate arguments to present people with diverse viewpoints in order to help them make more informed decisions (Hidey and McKeown, 2019; Hua et al., 2019; Hua and Wang, 2018; Wang and Ling, 2016).

This thesis focuses on understanding the factors of persuasion in computational argumentation. Persuasion is an act of presenting arguments to change people's opinions, values, and behaviors on a controversial topic or an event (Matilda White, 1954). Theories of persuasive communication are applied to various fields such as marketing, advertising, social psychology, and politics

(Shrum, 2012). Researchers in these areas are interested in understanding the factors that influence the success of persuasive communication. The emergence of social media and online argumentation platforms has made it more accessible for people to engage in argumentative discussions with others who may hold differing views. Interpretation of the underlying dynamics of argumentative communication online can help develop methods to improve the effectiveness of arguments. For example, it could be used to provide feedback to users in order to help them improve the structure of their arguments. Moreover, analyzing these interactions can help understand the factors that influence people's behavior in the argumentative process. We specifically study persuasion on online debating platforms to get insights into the factors that govern people's decision-making in persuasion.

Language is the primary tool that is used to convey the content of an argument. Therefore it is a crucial component in persuasion (van Eemeren and Eemeren, 2009; Perelman and Olbrechts-Tyteca, 1971; Petty and Cacioppo, 1984; Richard E. et al., 1981). It is only natural then that the majority of the work in computational studies of persuasion has focused mainly on understanding the characteristics of persuasive text, e.g., what distinguishes persuasive from non-persuasive text (Fang et al., 2016; Habernal and Gurevych, 2016a,b; Hidey et al., 2017; Tan et al., 2016; Wachsmuth et al., 2016; Zhang et al., 2016). However, language is not the only factor in persuading people. Prior research in Social Sciences and Psychology has shown that the recipients of an argument may form their opinion on an issue based on non-content cues such as the characteristics of the speaker (i.e., the source) and their own predispositions (Cialdini, 2001; Petty and Cacioppo, 1984; Richard E. et al., 1981). For example, the credibility and trustworthiness of a speaker (Shelly, 1980; Smith and Shaffer,

1995) and the prior beliefs of the audience (Chambliss and Garner, 1996) have been shown to have a substantial effect on persuasive communication. Furthermore, people with strong prior beliefs on controversial issues have been shown to have biased stances even when presented with empirical evidence: i.e., they tend to find empirical evidence that confirms their prior beliefs more convincing (Charles G. et al., 1979). Given the evidence from Social Sciences and Psychology, we believe that accounting for the impact of these factors, in addition to the language, in computational studies of persuasion is crucially important. This thesis introduces several contributions to make progress towards this goal by exploring the following questions:

- What is the role of people's prior beliefs and initial stance on persuasion?
- How can we disentangle the effect of source and audience factors in order to understand the characteristics of persuasive language?
- What is the effect of social interaction on people's success at persuasion over time?
- Does pragmatic context play an important role in predicting the impact of arguments?

1.2 Contributions

In order to understand the role of speaker and audience effects in persuasion, we primarily look at the following factors of interaction on online argumentation:

1. Prior beliefs and initial stances of the speaker and the audience.

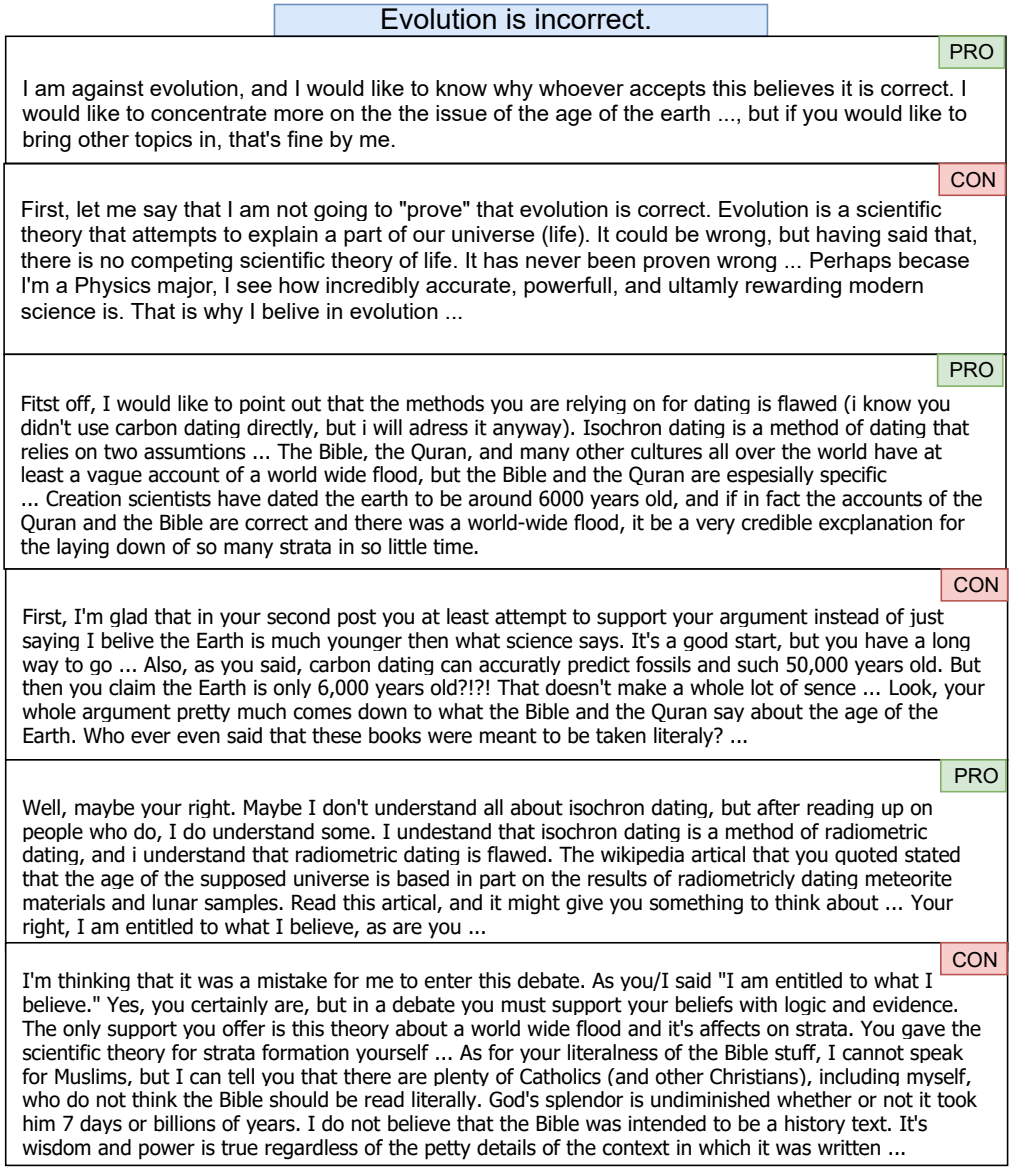


Figure 1.1: A debate on ”Evolution”.

2. Social interactions.

3. Language and pragmatic context.

Specifically, we make the following contributions:

A Dataset to Model Source and Audience Factors in Persuasion. One of the main bottlenecks in studying the effect of source and audience factors is the lack of large-scale datasets that contain information about the characteristics of the users. In order to bridge this gap and enable further studies in this area, we present a large-scale dataset (DDO) with a wide range of user information collected from an online debating website.[1] This dataset contains debates on a wide range of controversial topics. Each debate consists of two debaters with opposing views on a controversial topic, who take turns to provide their arguments. Figure 1.1 shows an example debate on *"Evolution"* contained in this dataset. Along with the debate, the dataset also contains votes from the audience evaluating various aspects of the debaters, such as the persuasiveness of their arguments and their overall debating skills. Besides the debates, the dataset also includes information about the debaters and the audience, such as their stance on controversial topics, political and religious ideologies, education level, etc. We obtain this from the self-identified information that the users provide on their profiles.

The Role of Prior Beliefs in Persuasion. The majority of work in computational persuasion has focused on understanding the characteristics of persuasive language. In this thesis, we mainly focus on understanding the effect of user factors on persuasion. We use the debates, votes, and user information available on the DDO dataset to study the effect of prior beliefs in predicting which debater an individual voter will find more persuasive for a given debate. We find that user factors play a critical role in this prediction task. Furthermore, controlling for the effect of user-level factors allows us to investigate charac-

[1] `https://www.debate.org`

teristics of persuasive language without any influence from these potentially confounding factors.

Effect of Social Interaction on Persuasion Success. Inspired by prior work that shows a strong relationship between a user's social interaction and their influence on social media (Cha et al., 2010; Romero et al., 2011), we study whether success in persuasion might also depend on an individual's social interaction and engagement. In particular, we study whether users can improve the persuasiveness of their arguments as they gain more experience using the debating platform. We show that a user's social interaction is an essential factor in predicting their overall success at debating.

Representing Pragmatic Context in Modeling Argument Impact. We present a new dataset to study the effect of the pragmatic and discourse context when determining an argument's impact. We further propose predictive models that can incorporate pragmatic and discourse context. We find that these models outperform models that rely only on claim-specific linguistic features for predicting the perceived impact of individual claims within a particular line of argument.

1.3 Organization of Thesis

In Chapter 2, we first give an overview of recent developments in computational argumentation, describing recent work on argument analysis and argument generation. In Chapter 3, we discuss the details and statistics of the DDO dataset. With the dataset in hand, in Chapter 4, we present methods that can account for the effects of the speaker and the audience in predicting the persua-

siveness of an argument. In Chapter 5, we describe our contributions on understanding the impact of social interaction on persuasion on online debating platforms. In Chapter 6, we propose a new dataset that allows us to study the effect of pragmatic context (i.e., *kairos*) on assessing the impact of an argument. Finally, in Chapter 7, we summarize our contributions and provide directions for future work.

CHAPTER 2

AN OVERVIEW OF COMPUTATIONAL ARGUMENTATION

This chapter describes the recent developments in various sub-fields of computational argumentation, including computational argumentation mining, computational persuasion, and automated argument generation.

2.1 Computational Argument Mining

Computational argumentation mining aims to extract argument components and the relationships between them from unstructured text, building on theoretical models of argument (Toulmin, 1958; Walton et al., 2008). The main goal is to understand the points in an argument and get insights into how these points support or oppose each other. Having a deeper understanding of the structure of the arguments is important for various applications such as debating technologies (Slonim et al., 2021), legal decision-making (Moens et al., 2007), automated essay scoring (Ong et al., 2014), and computer-assisted writing (Stab and Gurevych, 2017). The identification of argument structure involves several sub-tasks:

1. Determining the "argumentative" vs. "non-argumentative" parts of the text (Moens et al., 2007).
2. Classifying argumentative components into categories such as "Claim" or "Premise" (Chakrabarty et al., 2019a; Mochales and Moens, 2011; Stab and Gurevych, 2017).
3. Identifying relations between argument components (Cabrio and Villata, 2013; Carstens and Toni, 2015; Feng and Hirst, 2011; Hua and Wang, 2017;

Niculae et al., 2017; Palau and Moens, 2009; Park and Cardie, 2014; Stab and Gurevych, 2017).

Most research in computational argumentation mining has proposed methods for a subset of the subtasks mentioned above. Persing and Ng (2016) was among the first to present an end-to-end pipeline approach to determine argumentative components and their relationship using an Integer Linear Programming (ILP) framework. Similarly, Stab and Gurevych (2017) has proposed a joint model that globally optimizes argument component types and relations using ILP. Eger et al. (2017) has presented the first end-to-end neural argumentation mining model obviating the need for designing hand-crafted features and constraints.

Argumentation mining has been applied to various domains such as persuasive essays, legal documents, political debates, and social media data (Dusmanu et al., 2017). For instance, Stab and Gurevych (2017) has built an annotated dataset of persuasive essays with corresponding argument components and relations. Using this corpus, Eger et al. (2017) developed an end-to-end neural method for argument structure identification. Nguyen and Litman (2018) has further proposed an end-to-end method to parse argument structure and used the argument structure features to improve automated persuasive essay scoring. Furthermore, Levy et al. (2014) has studied context-dependent claim detection by collecting annotations for Wikipedia articles. Using this corpus, Rinott et al. (2015) has investigated the task of automatically identifying the corresponding pieces of evidence given a claim. Bar-Haim et al. (2017a) has further proposed the task of claim-stance detection (i.e., given a topic and claims, identifying for each claim whether it supports or opposes the topic.) by further annotating

Wikipedia articles with stance information. Walker et al. (2012) has collected posts from *4forums.com*, a debating forum, and have further annotated part of this corpus for various aspects of arguments such as topic, stance, agreement, and sarcasm. Park and Cardie (2018) has proposed an argument mining corpus from Consumer Debt Collection Practices (CDCP) rule by the Consumer Financial Protection Bureau (CFPB) posted on *regulationroom.org*. Using this corpus, Niculae et al. (2017) proposed a structured prediction model for argumentation mining.

Although most of the research in Argumentation Mining has focused on English monologues, Peldszus (2015) has collected a corpus of microtexts in German and used this corpus for argument component detection. Furthermore, Basile et al. (2016) has studied relation prediction task in Italian news blogs. Similarly, there has been some recent work investigating argumentation mining beyond monologues, i.e., looking at the process of argumentation in dialogues. For example, Chakrabarty et al. (2019b) has proposed a method to identify the argument structure in persuasive dialogues that can model the micro-level (i.e., the structure of a single argument) and the macro-level (i.e., the interplay between the arguments) characteristics of arguments.

Stance detection and Argumentation Mining are closely related tasks, given that they both aim to understand standpoints from the text on a controversial topic. Contrary to stance detection, argumentation mining aims also to extract a more fine-grained structure of arguments, identifying claims, premises, and the relationship between them. There has been a lot of research on identifying the stance of arguments on a controversial topic (Bar-Haim et al., 2017b; Hasan and Ng, 2013; Sobhani et al., 2015; Sun et al., 2018). For example, Sobhani et al. (2015)

has shown that using argument structure features improves the performance of stance detection models. Wachsmuth et al. (2018) has further studied retrieval of the best counter-arguments, using arguments opposing the same aspect of the controversial topic. In our work (Durmus et al., 2019a), we have found that encoding contextual information using the argument structure tree is crucial to achieving state-of-the-art performance for argument stance detection. Kobbe et al. (2020) has proposed an unsupervised method to assess the stance of the arguments inferring whether the outcome is good vs. bad.

For a more detailed discussion of the argumentation mining literature, refer to comprehensive surveys by Peldszus and Stede (2013), Cabrio and Villata (2018), and Lawrence and Reed (2019).

2.2 Computational Studies of Persuasion

Understanding the characteristics of persuasive language has been a great interest of computational studies of persuasion. Most of the work in this domain has focused solely on language (Atkinson et al., 2019; El Baff et al., 2020; Guerini et al., 2015; Habernal and Gurevych, 2016a; Hidey et al., 2017; Li et al., 2020; Morio et al., 2019; Persing and Ng, 2017; Tan et al., 2016). For instance, Habernal and Gurevych (2016b) has collected a new corpus to study the task of predicting which argument from an argument pair is the more convincing. Zhang et al. (2016) has studied the role of conversational flow and interplay between debaters on persuasion in Oxford-style debates. Hidey and McKeown (2018) has further investigated the role of larger context on persuasion, modeling the sequence of arguments in a discussion thread on "Change My View (CMV)",

a discussion forum on Reddit. Wang et al. (2019) has investigated which types of persuasion strategies have a more significant impact in convincing people to donate to a specific charity. This work is the first step in building personalized persuasive dialogue systems. Furthermore, to study whether particular types of people find particular argument styles more convincing, Lukin et al. (2017) has collected a new corpus of personality information and belief change in socio-political arguments. They have shown that belief change is affected by personality factors. For example, conscientious people are more convinced by dialogic, emotional arguments, while agreeable people are more likely to be persuaded by dialogic, factual arguments. Inspired by this line of research, this dissertation further investigates the effect of source and audience factors in persuasion by asking the following questions:

1. How do the prior beliefs of the audience affect the process of persuasion?
2. Do social interactions play an essential role in people's success in online argumentation?
3. How can we measure the relative impact of source and audience factors, language, and the pragmatic context in computational studies of persuasion and argument impact prediction?

Prior work has also investigated the related tasks of argument quality assessment and argument impact prediction (El Baff et al., 2018; Persing and Ng, 2015; Wachsmuth et al., 2017a). For example, Persing and Ng (2015) has introduced a corpus of argumentative student essays annotated with argument strength scores. They have further proposed a supervised, feature-based model to score the essays based on argument strength automatically. Wachsmuth et al.

(2017b) has studied logical, rhetorical, and dialectical quality dimensions and proposed a taxonomy of argumentation quality from these dimensions. El Baff et al. (2018) has explored argument quality in news editorials, collecting annotations for the perceived effect of editorials from the New York Times. We have further explored the role of pragmatic context in predicting the perceived impact of arguments on online argumentation platforms (Durmus et al., 2019b).

2.3 Argument Generation

Argumentation is a significant part of a wide range of human activities. Humans are constantly confronted by situations where they are trying to persuade or are being persuaded. A major goal of computational argumentation is to build systems that can have meaningful debates and argumentative interactions with humans. Recent work in the area has made progress towards this goal through the automated generation of argumentative text (Bar-Haim et al., 2020; Hua and Wang, 2018; Sato et al., 2015; Zukerman et al., 2000). Zukerman et al. (2000) and Alshomary et al. (2020) have proposed a Bayesian argument generation system to generate arguments given the corresponding argumentation strategies. Sato et al. (2015) has presented a sentence-retrieval-based end-to-end argument generation system that can participate in English debating games. Hua et al. (2019) has explored a neural counter-argument generation method that consists of a text planning decoder and a content realization decoder to select the main talking points and generate an argument given the talking points. Hidey and McKeown (2019) has further proposed a neural model that edits the original claim semantically to produce a claim with an opposing stance. Similarly, Hua and Wang (2018) has studied the task of generating arguments of a

different stance for a given argument. They have further incorporated external knowledge into the encoder-decoder architecture and have shown that their model can generate arguments that are more likely to be on topic. Wang and Ling (2016) and Bar-Haim et al. (2020) have investigated the problem of summarizing the key points of an argument. Most recently, Slonim et al. (2021) has proposed an autonomous debating system (Project Debater) that can engage in a competitive debate with humans by generating a pipeline of four main modules: argument mining, an argument knowledge base (AKB), argument rebuttal, and debate construction. They have shown that their debating system can engage in a competitive debate with humans. However, they highlight the difficulty of achieving this end-goal due to the following reasons:

1. The outcome of the debates (i.e., selection of the winner) is highly subjective and open to interpretation since it may dependend on the characteristics of the audience.
2. Unlike other games such as chess (Campbell et al., 2002) or backgammon (Tesauro, 1995), humans would expect to be able to interpret every move of the system since they vote to decide the winner of the debate.
3. There are a limited number of structured debate datasets to train such systems.

CHAPTER 3

A DATASET FOR MODELING USER CHARACTERISTICS IN PERSUASION

Previous work in Natural Language Processing (NLP) and Computational Social Science (CSS) that studies persuasion has mainly focused on identifying the content and structure of an argument (Feng and Hirst, 2011) along with the linguistic features that are indicative of effective argumentation strategies (Tan et al., 2016). However, the effectiveness of an argument cannot be determined solely by its textual content; instead, it is essential to consider the reader's or participant's characteristics in the debate or discussion. Does the reader already agree with the argument's stance? Is she predisposed to changing her mind on a particular topic? Is the style of the argument appropriate for the individual? To date, existing argumentation datasets have permitted only a limited assessment of such "user" traits because information on the background of users is generally unavailable. This chapter introduces a new dataset with a wide range of user information to make progress towards this goal. We view this new dataset as a resource that allows the NLP and CSS communities to understand the effect of audience characteristics on the efficacy of different persuasion strategies. In the following subsection, we describe our dataset in the context of existing argumentation datasets. We then provide a description and statistics for the key aspects of the dataset.

3.1 Related Work and Datasets

There has been a tremendous amount of research effort to understand the important linguistic features for identifying argument structure and determining

effective argumentation strategies in a monologic text (Feng and Hirst, 2011; Guerini et al., 2015; Mochales and Moens, 2011; Stab and Gurevych, 2014). For example, Habernal and Gurevych (2016a) has experimented with different machine learning models to predict which of the two given arguments is more convincing. To understand what kind of persuasive strategies are effective, Hidey et al. (2017) has further annotated different modes of persuasion (i.e., ethos, logos, pathos) and looked at which combinations appear most often in more persuasive arguments.

Understanding argumentation strategies in conversations and the effect of the interplay between the participants' language has also been an important avenue of research. Tan et al. (2016), for example, has examined the effectiveness of arguments on ChangeMyView[1], a debate forum in which people invite others to challenge their opinions. They found that the interplay between the language of the opinion holder and that of the counterargument provides highly predictive cues of persuasiveness. Zhang et al. (2016) has examined the effect of conversational style in Oxford-style debates and found that the side that can best adapt in response to opponents' discussion points throughout the debate is more likely to be more persuasive.

Although research on computational argumentation has mainly focused on identifying important linguistic features of the argument, there is also evidence that it is important to model and account for the information about the debaters and the people who are judging the quality of the arguments: multiple studies in Social Sciences and Psychology show that people perceive arguments from different perspectives depending on their backgrounds and experiences (Correll et al., 2004; Hullett, 2005; Lord et al., 1979; Petty et al., 1981; Vallone et al.,

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

1985). Lukin et al. (2017) is one of the first to computationally study the impact of the audience, looking at the effect of their OCEAN personality traits (Roccas et al., 2002; T. Norman, 1963) on how they judge the persuasiveness of monologic arguments. This work is the most similar to ours since the effect of users' personalities is explored in the persuasion process. Our dataset does not have explicit information about users' personality traits; however, we have extensive information about their demographics, social interactions, beliefs, and language use. Durmus and Cardie (2019a) describes the details of this dataset.

3.2 DDO Dataset

We collected 78,376 debates from debate.org (DDO)[2] from 23 different topic categories, including *Politics*, *Religion*, *Health*, *Science*, and *Music*.[3] DDO is an online argumentation platform where people can engage in debates, participate in forums and polls, and post their opinions on controversial topics. Participating in debates provides users an opportunity to challenge other users to change their opinions. After participating in debates, they receive feedback from the audience on the platform. This feedback mechanism is helpful for users to develop strategies to improve their debating skills over time. In addition to the text of the debates, we collected votes from the readers of these debates. Votes evaluate different dimensions of the debate, and they are important to determine which debaters are more successful in persuading other users.

Each user creates a profile on this platform to share information about their background and preferences. To study the characteristics of users on persua-

[2]www.debate.org

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

	Debate Rounds (3)	Comments (39)	Votes
PRO	[+14 words...], the year is better spent with a full time parent. Most children will learn very little at the preschool. [... +39 words]		
CON	[+11 words...], the right school can be an excellent resource for a parent. A child need to have a place to meet other children. [... +44 words]		

Figure 3.1: ROUND 1 for the debate claim "PRESCHOOL IS A WASTE OF TIME.".

sion, we collected user information for 45,348 different users. In the next section, we share more details about the debates and the user information on this dataset.

3.2.1 Debates

Debate rounds. Each debate consists of a sequence of ROUNDS in which two debaters from opposing sides (one is supportive of the claim (i.e., PRO) and the other is against the claim (i.e., CON)) provide their arguments. Each debater has a single chance in a ROUND to make their points. Figure 3.1 shows an example ROUND 1 for the debate claim "PRESCHOOL IS A WASTE OF TIME". The number of ROUNDS in a debate ranges from 1 to 5, and most debates contain 3 or more ROUNDS. The goal of the debaters in each ROUND is to provide arguments that would refute the opponent's points and convince readers to side with their stance.

Votes. All users in the *debate.org* community can vote on debates. As shown in Figure 3.2, voters share their stances on the debate topic before and after the debate and evaluate the debaters' conduct, spelling and grammar, persuasive-

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

Figure 3.2: An example post-debate vote.

Online:	1 Day Ago	**Name:**	- Private -
Updated:	5 Months Ago	**Gender**	Male
Joined:	5 Months Ago	**Birthday:**	- Private -
President:	Barack Obama	**Email:**	- Private -
Ideology:	Liberal	**Education:**	Some College
Party:	Democratic Party	**Ethnicity:**	White
Relationship:	Single	**Income:**	Not Saying
Interested:	in Women	**Occupation:**	Student
Looking:	No Answer	**Religion:**	Atheist

Figure 3.3: Demographic and private state information for an example user profile.

ness, and reliability of the sources they refer to. For each such dimension, voters can choose one of the debaters as better or indicate a tie. The audience scores the debaters on these different aspects, and a winner is declared accordingly. [4] This fine-grained voting system gives a glimpse into the reasoning behind the voters' decisions.

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

Debate Statistics	
Debates:	296
Lost:	48
Tied:	23
Won:	225
Win Ratio:	82.42%
Percentile:	99.99%
Elo Ranking:	5,612

Activity Statistics	
Forum Posts:	25,465
Votes Cast:	1,124
Opinion Arguments:	8
Opinion Questions:	1
Poll Votes:	973
Poll Topics:	47

Figure 3.4: Information about the activities of an example user profile.

3.2.2 User information

The dataset includes extensive information about the users' demographics and private state, their activity on this platform, and their stance on various controversial topics. In this section, we describe the user information that is available in this dataset.

Demographic and Private State Information

On *debate.org*, each user has the option to share demographic and private state information such as their age, gender, ethnicity, political ideology, religious ideology, income level, education level, and the political party they support. Figure 3.3 provides an example for the demographic and state information included in a user profile. We can see that these users select their political ideology, ethnicity, education, religious ideology, etc. However, they prefer not to share some of the information about themselves, such as their birthday, email, and income level, since sharing the demographic and state information is optional.

The BIG Issues

Abortion:	Con
Affirmative Action:	Con
Animal Rights:	Con
Barack Obama:	Con
Border Fence:	Pro
Capitalism:	Pro
Civil Unions:	N/O
Death Penalty:	Pro

Drug Legalization:	Pro
Electoral College:	Pro
Estate Tax:	Con
Flat Tax:	Con
Free Trade:	Pro
Gay Marriage:	Pro
Globalization:	Pro
Gun Rights:	Pro

Figure 3.5: Opinions on the **Big Issues** of an example user profile.

User Activity Information

Beyond the demographic and private state information, we have access to information about their activities on the website, such as their debating success rate, their participation both as debaters and voters, their votes, their forum posts, opinion arguments, opinion questions, poll votes, and poll topics that they created. The activities of an example user is shown in Figure 3.4. The availability of this information provides an opportunity to study users' interactions and success on this platform over time.

User Opinions on the *Big Issues*

The editors of the platform determine a list of the most controversial debate topics. These are referred to as *big issues*[5]. Each user has the option to share their stance on each *big issue* on their profile (see Figure 3.5): either PRO (in favor), CON (against), N/O (no opinion), N/S (not saying), or UND (undecided). This gives a glimpse into the prior stance of users on a wide range of controversial topics. Moreover, this information can be used to determine opinion similarity

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

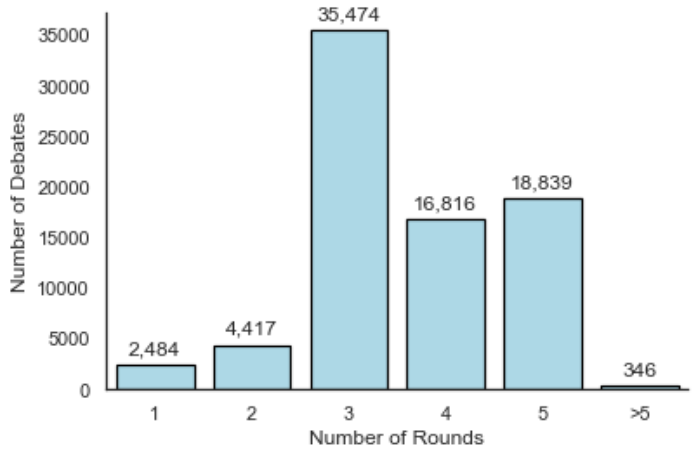


Figure 3.6: The number of debates with the given number of rounds.

between a pair of users.

3.3 Data Statistics

The dataset consists of 78,376 debates from October of 2007 until November of 2017 with comprehensive user profile information for 45,348 users. Statistics on the number of debates with their corresponding number of rounds and votes are shown in Figure 3.6 and Figure 3.7, respectively. The majority of debates have 3 to 5 rounds. There are some debates with only one round; however, most debates have two or more rounds since the debates are highly interactive.

Although there are many debates with no votes, around 21k debates have three or more votes. We disregard the debates with ≤ 3 votes in our studies in order to have enough feedback to model the factors of success in persuasion.

Figure 3.8 shows the number of debates that users participated in. The majority have participated in only a single debate. However, some users actively participate in many debates. For example, around 2k debaters have participated

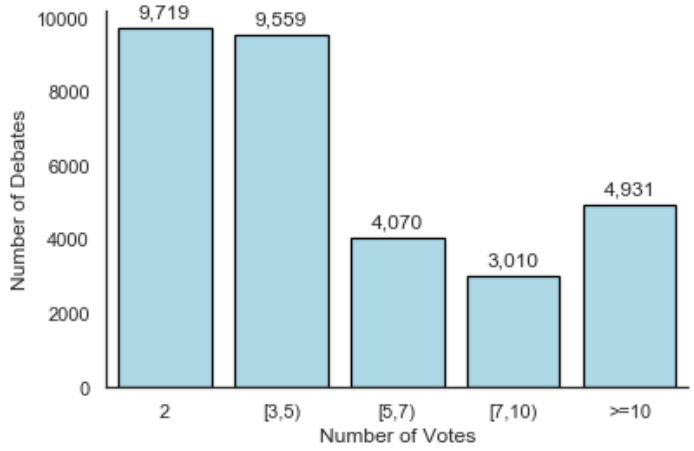


Figure 3.7: The number of debates for a given range of votes.

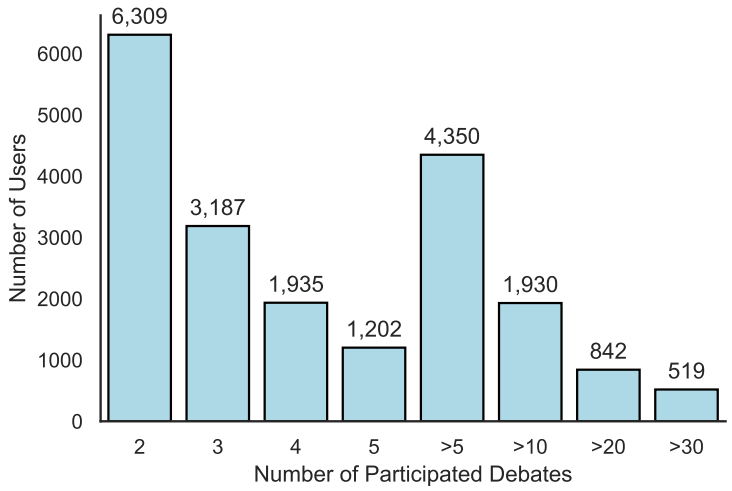


Figure 3.8: The number of users that have participated in a given number of debates.

in more than ten debates during the period included in the dataset. We study these debaters to understand the factors of debating success over time.

3.4 Limitations

The dataset includes comprehensive information about users on the platform, which allows us to model user factors in persuasion. However, we acknowledge that we are unable to represent all demographics due to a lack of data. Participation on the platform tends to be highly skewed towards an American audience. Moreover, even within this group, the distribution of user characteristics may not be representative enough. Therefore, some valid opinions may be under-represented, and this should be accounted for while employing models derived from this data. Furthermore, we assume that the information users share on their profiles is accurate, and we use this information to model their characteristics. However, there is no mechanism on this platform to ensure that users provide accurate information.

3.5 Chapter Summary

In this chapter, we present a novel dataset, DDO, of debates collected from *debate.org*. The dataset includes interactive debates along with votes from the audience to evaluate various aspects of each debater. Moreover, the dataset has comprehensive information about the users on the platform. This allows us to study the effect of source and audience factors in persuasion (Chapter 4). We further use this dataset to model the impact of social interactions on long-term success in online debating (Chapter 5).

CHAPTER 4

THE ROLE OF THE SPEAKER AND THE AUDIENCE FACTORS IN COMPUTATIONAL PERSUASION

Using the DDO dataset described in Chapter 3, this chapter studies the effect of factors associated with the speaker and the audience of a debate to assess the more persuasive debater with respect to an individual audience member.

4.1 Background

Most of the recent work in computational persuasion has focused on identifying the characteristics of persuasive language (Habernal and Gurevych, 2016a; Hidey et al., 2017). However, there is evidence from the Social Sciences and Psychology that non-content cues such as the factors of the speaker and the audience play an essential role in persuasion and opinion formation. Instead of carefully processing the content of the arguments, people may rely on simple non-content heuristics in decision making (Charles G. et al., 1979). Understanding the effect of persuasion strategies on people, the biases people have, and the impact of people's prior beliefs on their opinion change has been an active area of research (Correll et al., 2004; Hullett, 2005; Petty et al., 1981).

Prior work has shown that the speaker's credibility is an essential factor for people's perceptions of the arguments (Shelly, 1980; Smith and Shaffer, 1995). For example, there is a significant correlation between the communication speed and the persuasive effect of the arguments. The audience perceived a communicator with a faster communication rate as more credible without really focusing on the content of the arguments (McGuire, 1969; Smith and Shaffer, 1995). Fur-

thermore, Shelly (1980) has studied the effect of a communicator's perceived likability in opinion formation and found that low-involvement subjects perceive the arguments of likable communicators as more persuasive. High involvement subjects (i.e., the subjects who feel their opinion judgments have essential consequences for themselves) have shown to have a more systematic strategy that assigns a higher weight to the message content in opinion formation. A communicator's perceived attractiveness is also positively correlated with their persuasiveness since the audience perceives more attractive communicators as more effective (Chaiken, 1979; Eagly and Chaiken, 1975).

There is further evidence showing that people's prior beliefs significantly affect their opinion formation (Chambliss and Garner, 1996). People with strong prior beliefs on controversial issues have shown to have biased stances even when they are presented with empirical evidence: i.e., they tend to find empirical evidence that is confirming their prior beliefs more convincing (Charles G. et al., 1979). Similarly, people judge the fairness and reliability of source content in a biased way; i.e., they accept evidence that supports their stance at face value while scrutinizing evidence that threatens their initial position (Vallone et al., 1985). Inspired by these findings, we study the impact of prior beliefs in computational persuasion in this chapter. Lukin et al. (2017) is the most relevant work to ours since they investigated the effect of an individual's personality features (open, agreeable, extrovert, neurotic, etc.) on the type of argument (factual vs. emotional) they find more persuasive. Our work differs from this work since we study debates. In addition, we look at different types of user profile information, such as a user's religious and ideological beliefs and prior beliefs and opinions of the audience on various topics (Durmus et al., 2019b; Longpre et al., 2019).

4.2 Role of Prior Beliefs in Computational Persuasion

Using the DDO dataset (described in Chapter 3), we first analyze which dimensions of argument quality are the most important for determining the successful debater. Then, we investigate whether there is any connection between selected user-level factors and users' opinions on the *big issues* to see if we can infer their opinions from these factors. Finally, using our findings from these analyses, we perform the task of predicting which debater will be perceived as more successful by a given voter. In this study, we particularly aim to understand the role of users' prior beliefs (i.e., their self-identified political and religious ideology) in predicting the more successful debater.

4.2.1 Relationships between argument quality dimensions

In Section 3.2.1, we describe the aspects the voters evaluate in order to determine which debater is more successful. There are two alternative criteria for determining the successful debater. We consider both in our experiments.

Criterion 1: Argument quality. Debaters get points for each dimension of the debate. The most important dimension — in that it contributes most to the point total — is making convincing arguments. The debater with the highest point total is declared the winner. *debate.org* uses Criterion 1 to determine the winner of a debate.

Criterion 2: Convinced voters. Alternatively, since voters share their stances before and after the debate, the debater who convinces more voters to change

their stance can be considered the winner.

Figure 4.1 shows the correlation between pairs of voting dimensions (in the first eight rows/columns) along with the correlation of each dimension with (1) getting the highest point total (row/column 9) and (2) convincing more to change their stance (final row/column). The abbreviations in Figure 4.1 stand for (on the CON side): has better conduct (CBC), makes more convincing arguments (CCA), uses more reliable sources (CRS), has better spelling and grammar (CBSG), gets more total points (CMTP) and convinces more voters (CCMV). For the PRO side we use PBC, PCA, and so on.

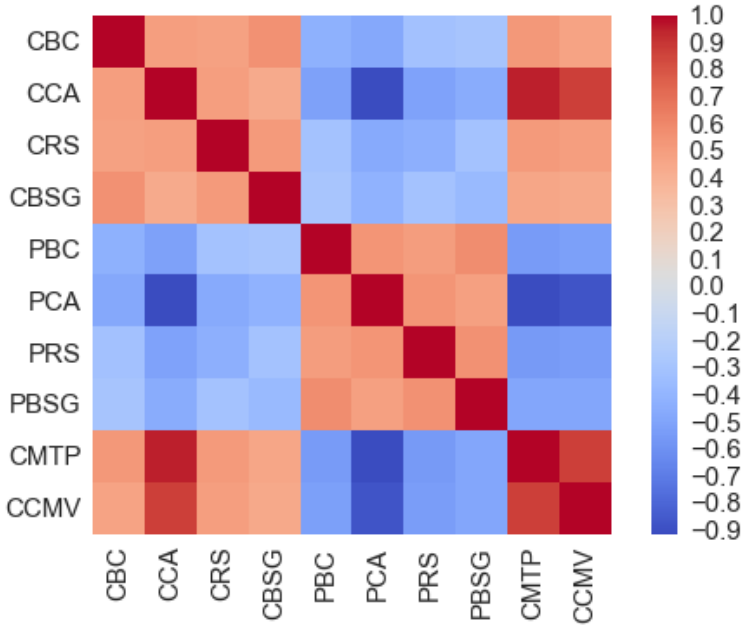


Figure 4.1: The correlations among argument quality dimensions.

From Figure 4.1, we can see that making more convincing arguments (CCA) correlates the most with total points (CMTP) and convincing more voters (CCMV). This suggests that the language of the argument is important in persuading the audience, and it motivates us to identify the linguistic features that are indicative of convincing arguments while taking into account speaker

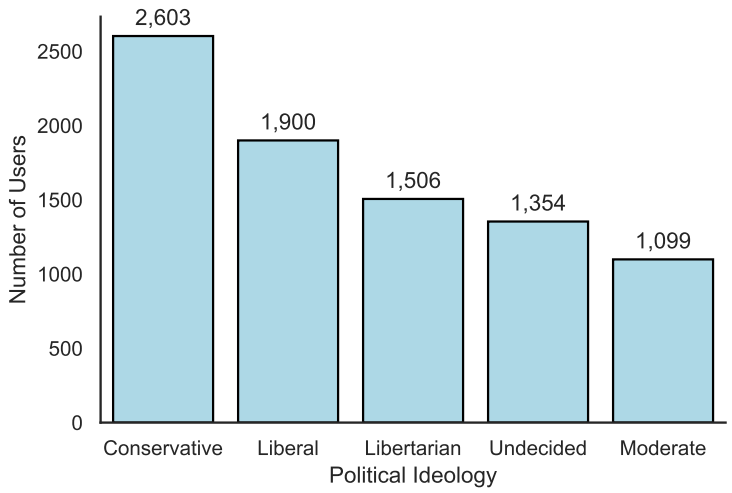


Figure 4.2: The number of users with the given political ideology.

and audience factors.

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

As described in Section 3.2.2, users share their self-identified political and religious ideologies along with their opinions on various controversial issues (i.e., *big issues*). Note that many people prefer not to share their political and religious ideologies. Figures 4.2 and 4.3 show the number of users who self-identify with the given political or religious ideology.

We disentangle different aspects of a person's prior beliefs in order to understand how they correlate with their opinions on the *big issues*. We focus on prior beliefs in the form of self-identified political and religious ideology.

Representing the *big issues*. To represent a user's opinion on a particular

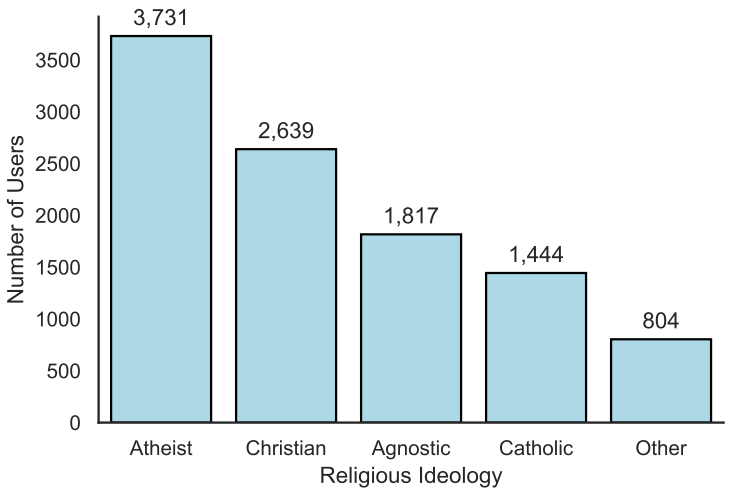


Figure 4.3: The number of users with the given religious ideology.

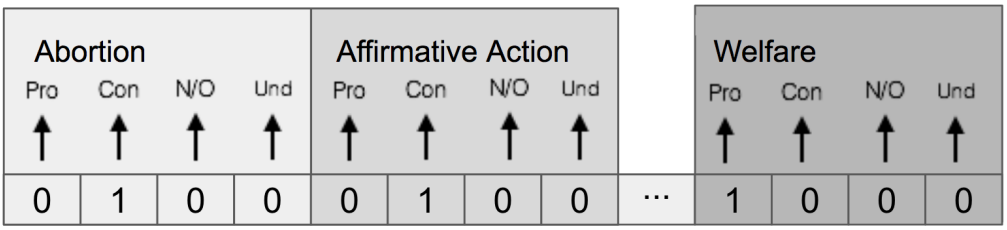
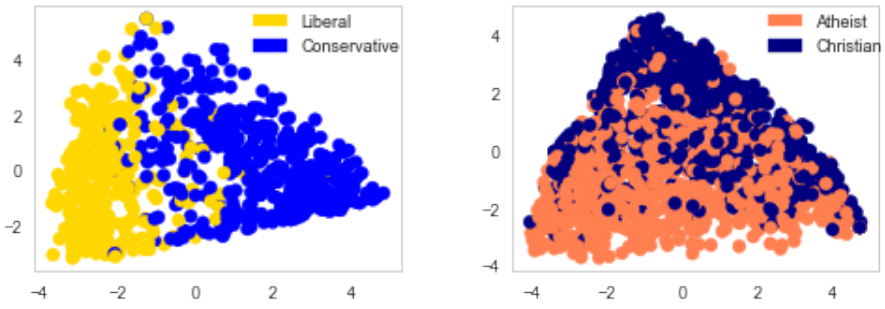


Figure 4.4: The representation of the BIGISSUES vector derived by this user's decisions on Big Issues. Here, the user is CON for ABORTION and AFFIRMATIVE ACTION issues and PRO for the WELFARE issue.

big issue, we use a four-dimensional one-hot encoding where the indices of the vector correspond to PRO, CON, N/O (no opinion), and UND (undecided), consecutively. Note that we do not have a representation for N/S (not saying) since we eliminate users who indicate N/S for any of the *big issues*. We then concatenate the vector for each of the *big issue* to represent a user's stance on all the *big issues* as shown in Figure 4.4. We denote this vector by BIGISSUES.

We test the correlation between an individual's opinion on the *Big Issues* and the selected user-level factors in this study using two approaches: clustering and classification.



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

Figure 4.5: PCA representation of decisions on Big Isues color-coded with political and religious ideology. We see more distinctive clusters for CONSERVATIVE vs. LIBERAL users suggesting that people's opinions are more correlated with their political ideology.

Clustering the users' decisions on the *big issues*. We apply PCA on the BIGISSUES vector of users who identified themselves as CONSERVATIVE vs. LIBERAL (740 users). We do the same for the users who identified themselves as ATHEIST vs. CHRISTIAN (1501 users). In Figure 4.5, we see distinct clusters of CONSERVATIVE vs. LIBERAL users in the two-dimensional representation, while for ATHEIST vs. CHRISTIAN, the separation is not as distinct. This suggests that people's opinions on the *big issues* identified by *debate.org* correlate more with their political ideology than their religious ideology.

Classification approach. We can also treat this as a classification task[1] using the BIGISSUES vector for each user as the input feature and the user's religious and political ideology as the labels to be predicted. Table 4.1 shows the prediction accuracy for religious and political ideology. We see that using the BIGISSUES vector as a feature performs significantly better[2] than the majority

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

[2]We performed the McNemar significance test.

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

Table 4.1: Accuracy using majority baseline vs. BIGISSUES vectors as features.

baseline.[3]

This analysis shows a clear relationship between people's opinions on the *big issues* and the selected user-level factors. It raises the question of whether it is even possible to persuade someone to change their stance on a given issue. It may be the case that people prefer to agree with the individuals with the same (or similar) beliefs regardless of the quality of opposing arguments. Therefore, it is crucial to understand the relative effect of prior beliefs vs. argument strength on persuasion.

4.2.3 Task formulation

Some of the previous work in NLP on persuasion, focuses on predicting the winner of a debate as determined by the change in the number of people supporting each stance before and after the debate (Potash and Rumshisky, 2017; Zhang et al., 2016). However, we believe that studies of the effect of language on persuasion should consider extra-linguistic factors that can affect opinion change. In particular, we propose an experimental framework for studying the effect of language on persuasion by controlling for the prior beliefs of the audience. In order to do this, we formulate a more fine-grained prediction task: for a given voter, predict which side/debater/argument the voter will declare as the

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

winner.

Task 1: Controlling for religious ideology. In the first task, we control for religious ideology by selecting debates where the debaters have differing religious ideologies (e.g., debater 1 is ATHEIST, debater 2 is CHRISTIAN). Also, we only consider voters that (a) self-identify with one of these religious ideologies (e.g., the voter is either ATHEIST or CHRISTIAN) and (b) changed their stance on the topic after the debate. For each such voter, we want to predict which debater did the convincing. Thus, in this task, we use *Criterion 2* to determine the winner of the debate from the voter's point of view. We hypothesize that a voter will be convinced by the debater that espouses the religious ideology of the voter. Given this setting, we can study the factors that govern whether a debater can convince any given voter. It also provides an opportunity to understand how voters who change their minds perceive arguments from a debater with the same vs. opposing prior beliefs.

To study the effect of the debate topic, we perform this study for two cases — debates belonging to the *Religion* category only vs. all categories. The *Religion* category contains debates like "IS THE BIBLE AGAINST WOMEN'S RIGHTS?" and "RELIGIOUS THEORIES SHOULD NOT BE TAUGHT IN SCHOOL". We expect to see a stronger effect due to prior beliefs for debates on *Religion*.

Task 2: Controlling for political ideology. Similar to the setting described above, Task 2 controls for political ideology. In particular, we only select debates where the debaters have differing political ideologies (CONSERVATIVE vs. LIBERAL). In contrast to Task 1, we consider all voters that self-identify with any of the debater's ideologies (regardless of whether the voter's stance changed post-debate). For this task, we predict which debater will get assigned more points

from a given voter. Thus, Task 2 uses *Criterion 1* to determine the winner of the debate from the point of view of a voter. We hypothesize that a voter will assign more points to a debater who shares the same political ideology.

Similar to task 1, we perform the study for two cases — debates from the *Politics* category only and debates from all categories. We expect to see a stronger effect due to prior beliefs for debates on *Politics*.

4.2.4 Features

The features we use in our model are shown in Table 4.2. They can be divided into two groups — features that describe the prior beliefs of the users and linguistic features of the arguments.

User features

We use cosine similarity between a voter and a debater's *big issue* vectors. This feature gives an approximation of the overall similarity of two users' opinions. We also use indicator features to encode whether the religious and political beliefs of a voter match that of a debater.

Linguistic features

We extract linguistic features separately for both the PRO and CON side of the debate (combining all the utterances of each side across the different turns). Table 4.2 contains a list of these features. It includes features that carry information about the style of the language (e.g., usage of modal verbs, length, punctuation),

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

Table 4.2: Feature descriptions.

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

4.3 Results and Analysis

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

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

Task 1 for debates in category *Religion*. As shown in Table 4.3, the majority baseline (predicting the winning side of the majority of training examples out of PRO or CON) gets 56.10% accuracy. User features alone perform significantly better than the majority baseline. The most important user-based feature is *matching religious ideology*. This means it is very likely that people change their

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

Table 4.3: Results for Task 1 for debates in category *Religion*. USER* represents the best performing combination of user-based features. LANGUAGE* represents the best performing combination of linguistic features. Since using linguistic features only would give the same prediction for all voters in a debate, the maximum accuracy that can be achieved using language features only is 92.86%.

views in favor of a debater with the same religious ideology. In a linguistic-only feature analysis, the combination of the *personal pronouns* and *connotation* features emerges as most important and performs significantly better than the majority baseline with 65.37% accuracy. When we use both user-based and linguistic features, the accuracy improves to 66.42% with *connotation* features. An interesting observation is that including the user-based features and the linguistic features changes the set of important linguistic features for persuasion, removing *personal pronouns* from the important linguistic features set. This shows the importance of studying potentially confounding user-level factors.

Task 1 for debates in all categories. As shown in Table 4.4, for experiments with user-based features only, *matching religious ideology* and *opinion similarity* features are the most important. For this task, *length* is the most predictive linguistic feature and can significantly improve the baseline (61.01%). When we

	Accuracy
Baseline	
Majority	57.31%
User-based Features	
Matching religious ideology	62.79 %
Matching religious ideology + Opinion similarity	62.97%
Linguistic features	
Length[4]	61.01 %
User-based + Linguistic features	
USER* + Length	64.56 %
USER* + Length + Exclamation marks	65.74%

Table 4.4: Results for Task 1 for debates in all categories. The maximum accuracy that can be achieved using language features only is 95.77%.

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

Table 4.5: Results for Task 2 for debates in category *Politics*. The maximum accuracy that can be achieved using linguistic features only is 75.35%. The *linguistic feature set* includes *rhetorical questions*, *emphasizing*, *approval*, *exclamation mark*, *questions*, *politeness*, *referring to opponent*, *showing evidence*, *modals*, *links*, and *numbers* as features.

combine the language features with user-based features, we see that with *exclamation mark*, the accuracy improves to (65.74%).

Task 2 for debates in category *Politics*. As shown in Table 4.5, using user-based features only, the *matching political ideology* feature performs the best (80.40%). Linguistic features (refer to Table 4.5 for the full list) alone can still

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

Table 4.6: Results for Task 2 for debates in all categories. The maximum accuracy that can be achieved using linguistic features only is 74.53%.

obtain significantly better accuracy than the baseline (59.60%). The most important linguistic features include *approval*, *politeness*, *modal verbs*, *punctuation*, and *argument lexicon features* such as *rhetorical questions* and *emphasizing*. When combining this linguistic feature set with the *matching political ideology* feature, we see that accuracy improves (81.81%). The *length* feature does not improve when it is combined with the user features.

Task 2 for debates in all categories. As shown in Table 4.6, when we include all categories, we see that the best performing user-based feature is the *opinion similarity* feature (73.96%). When using language features only, the *length* feature (56.88%) is the most important. For this setting, the best accuracy is achieved when we combine user features with *length* and *Tf-idf* features. We see that the set of language features that improves the performance of user-based features does not include some of the features that performed significantly better than the baseline when used alone (*modal verbs* and *politeness* features).

4.4 Persuasion of the Undecided

Research in psychology and political science suggests that there are critical differences in the persuasion of undecided versus decided voters/audience members. For example, Petty and Cacioppo (1996) has found 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 to maintain this consistency (Kosmidis, 2014; Sweeney and Gruber, 1984; Vecchione et al., 2013). In contrast, an undecided voter is asked to decide on an issue for which previously received information was somehow unconvincing; and prior work has shown that, as a result, these voters are likely to rely heavily on information conveyed in a new message (Kosmidis, 2014; Kosmidis and Xezonakis, 2010; Schill and Kirk, 2014).

Furthermore, the undecided voter group 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. Therefore, in our work, we explicitly study the factors that govern persuasion for *a priori* UNDECIDED versus DECIDED members of the audience (Longpre et al., 2019).

4.4.1 Task Formulation

We aim to study the most important factors in influencing audience members to be persuaded to one side or the other for each case (*a priori* undecided or

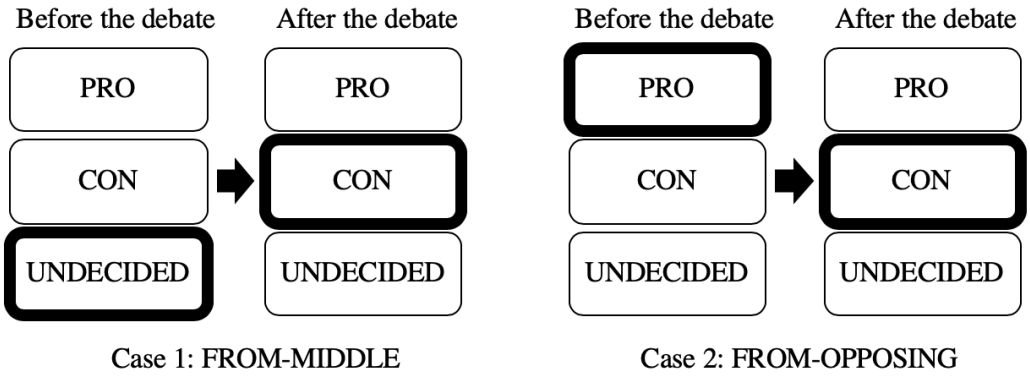


Figure 4.6: Example votes for a debate showing each case of persuasion.

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 experiment with the features described in Section 5.1.

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 (FROM-MIDDLE); and (2) a voter was (seemingly) decided beforehand and then flipped their stance FROM-OPPOSING. 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, we specifically look at which side (PRO vs. CON) did the convincing for a particular voter. Figure 4.6 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 undecided versus decided audience members. Table 4.7 summarizes the dataset statistics relevant to the voter cases.

Persuasion Case	# instances	# debates
FROM-MIDDLE	4,360	3,652
FROM-OPPOSING	2,642	2,183

Table 4.7: Number of voters in FROM-MIDDLE and FROM-OPPOSING categories.

4.4.2 Differences Between Persuasion Groups

We find distinct differences in the important features for predicting the outcome for voter groups FROM-MIDDLE and FROM-OPPOSING. Best-performing set of linguistic features for FROM-MIDDLE includes all features minus the *use of citations*, *referring to the opponent*, and *swear words*, while the best-performing set of linguistic features for FROM-OPPOSING includes all features minus *subjectivity*, *modals*[5], and *bi-/tri-gram TF-IDF*.[6]

The set of linguistic features that are important for each 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 that are the most effective. The use of modals, subjectivity, and general word choice are semantic features of an argument that can affect the perception of an argument's content. Based on our results, these content-based features are more important for undecided voters than for decided voters. In comparison, the use of swear words, citing sources, and referring to the opponent are stylistic features of an argument that can affect the perception of the debater. Our results indicate that these style-based features are not as important for undecided voters as 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

[5]The usage of modal verbs, i.e., *can*, *should*, *will*, and *may*.
[6]Calculated with a maximum of 30 terms.

of Sweeney and Gruber (1984); Vecchione et al. (2013) 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 found that affiliated voters do not adjust their positions in response to a party's actual policy statements but instead adjust their positions based on their subjective perceptions of the party. We have further found that audience-level aspects are comparatively more predictive of outcomes for undecided voters.

4.5 Limitations

In this study, we develop a framework to account for users' prior beliefs in their opinion formation. We mainly focus on users' political and religious ideologies and whether they are undecided vs. decided a priori. However, there are many user aspects such as debating experience, prior interactions, education level, etc., which can impact their opinion formation. We do not propose a method to account for all these factors simultaneously. Moreover, we do not suggest any causal implications since our findings are correlational.

4.6 Chapter Summary

In this chapter, we study the effect of the users' prior beliefs (i.e., political and religious ideology) and their initial stance on persuasion. We formulate the prediction task of determining which debater an individual voter finds persuasive in order to study the effect of these factors. We show that prior beliefs play a crucial role in this task. Furthermore, we explore the factors that govern persua-

sion for an a priori undecided vs. decided audience and find differences in the most predictive features for persuasion.

CHAPTER 5

MODELING THE EFFECT OF SOCIAL INTERACTION IN COMPUTATIONAL PERSUASION

In Chapter 4, we study the impact of prior beliefs on persuasion. In this chapter, using the DDO dataset described in Chapter 3, we explore the effect of a user's social interaction on their debating success, considering all the debates that the user participates in over time.

5.1 Background

There has been a tremendous amount of research on understanding user interactions and behaviour on social media (Backstrom et al., 2011; Benevenuto et al., 2009; Burke et al., 2009; Golder et al., 2007; Kumar et al., 2011; Lim et al., 2015; Macskassy and Michelson, 2011; Maia et al., 2008; Nagarajan et al., 2010; Wilson et al., 2009). For example, Wilson et al. (2009) analyze the interaction graphs of Facebook user traces and show that interaction activity on Facebook is significantly skewed towards a small portion of each user's social links. Lim et al. (2015) investigates how people interact in multiple online social networks. It has been further shown that there is a strong relationship between a user's social interaction and their influence on social media. For example, Romero et al. (2011) and Cha et al. (2010) and have shown that individuals with more activity and personal engagement are more influential on Twitter. Although there is a lot of work on understanding user behavior on social media sites such Facebook and Twitter, understanding the influence of user behavior on their persuasion success on debating platforms has been limited.

Romero et al. (2011) is the most similar to our work, in that the authors study the effect of interaction dynamics, such as participant entry order and degree of back-and-forth exchange in the discussion, on success in changing an opinion holder's stance in a thread. Note that, unlike our study, this work does not consider the effect of social interaction features (such as friendship network or voter network) on users' *success*. Moreover, we study the overall *success* of users over their lifetime, rather than a single debate or discussion thread.

We hypothesize that it is essential to account for the effect of social interactions in computational persuasion. Success in persuasion might also depend on an individual's social interaction and engagement with other users (on the debate platform) over time. For example, being more engaged with others over time may expose an individual to more diverse ideas and people, which could foster argumentation skills that are more applicable to convincing a more diverse audience. Focusing on only individual debates and discussion threads, prior work has not investigated the relative effect of an individual's social interaction, personal traits, and language use on their success in persuasion. In this chapter, we focus on online debates and study success over a user's lifetime by looking at interaction and engagement with the community over time, rather than focusing on individual debates to understand the relative impact of these factors on a user's success in persuasion.

5.2 Methodology

Our study employs the DDO (debate.org) dataset described in Section 3. Its extensive user information and multiple well-structured debates/interactions

per user provides a unique opportunity to study users' success over time while accounting for the effect of individuals' social interactions, personal traits, and language use. Users provide demographic information as well as their stance on controversial topics. They interact with one another in many ways: 1) debating, 2) evaluating the performance of other debaters, 3) commenting on debates, 4) asking/answering opinion questions, 5) voting in polls, 6) creating polls, 7) becoming friends.

5.2.1 Task Description

This section describes the methods used to investigate the underlying dynamics of success in online debate. First, we explain how we measure the users' success, and then we explore the role of personal traits, social interactions, and language in predicting success.

5.2.2 User Success

We compute the overall *success* in debating for a user u as:

$$success_u = \frac{\text{number of debates } u \text{ won}}{\text{number of total debates } u \text{ participated in as a debater}} \tag{5.1}$$

We treat users with $success_u \geq 70\%$ as *successful*, $success_u \leq 30\%$ as *unsuccessful* and $30\% < success_u < 70\%$ as *mediocre*.

5.2.3 Prediction Task

To understand the relative effect of a user's personal traits, social interaction, and language on their *success*, we study the following prediction task: **given a pair of debaters where one of them is *successful*, and the other is *unsuccessful* over the second and third stage of their lifetime, predict the *successful* one**. Note that while determining our label for success, we consider only the debates in the second and third stage of a user's lifetime to be able to study the relative effect of *success* in their first life stage (*success prior*) vs. other factors in a controlled way. We experiment with two settings where we control for the effect of *debate experience* and *success prior* respectively.

SETTING 1. To **control the effect of debate experience** in *success*, we create the pairs by **matching users according to the number of debates** that they participated in (i.e., users within a pair have the same number of debates).[1]

SETTING 2. Given that we're interested in understanding the factors that correlate with *success*, we **control for the *success prior*** in a very specific way – we only consider users that were *unsuccessful* in their initial life stage (**success prior**$\leq 30\%$[2]). This allows us to directly study the factors correlated with users that were initially *unsuccessful*, but later went on to become *successful* debaters.

In the following subsections, we describe each of the factors (i.e., personal traits, social interactions, and language) that we study in our experiments.

[1]There are 2,154 such pairs in our dataset.
[2]There are 957 such pairs in our dataset.

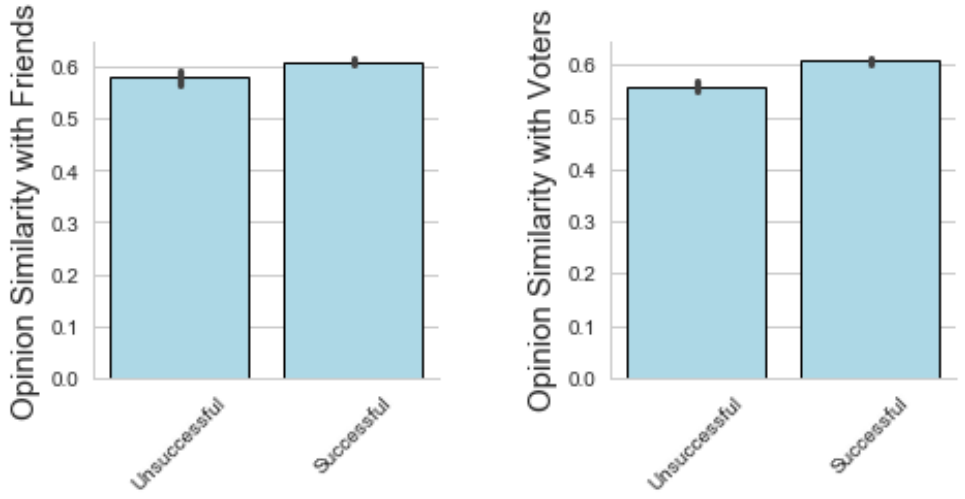
Personal Traits

In Chapter 4, we describe our findings on the role of prior beliefs on users' persuasion success in online argumentation, looking at the individual debates. We further investigate this effect in a debater's *success* over their lifetime. We also extend this study by considering additional personal traits, such as the degree to which a debater's demographic (e.g., gender and ethnicity) matches those of their friends and the voters participating in the debates.

We extract features to encode the similarity for a user's opinion, political ideology, religious ideology, gender, and ethnicity with that of her friends and voters. To compute opinion similarity, we used the information about users' opinions on the *big issues*.[3]

Figure 5.1 shows the similarity of *successful* and *unsuccessful* users' personal traits with that of their friends and voters respectively. We find that *successful* users have significantly higher opinion similarity with their friends than *unsuccessful* users. Moreover, they have significantly higher opinion similarity, religious ideology match, gender match, and ethnicity match with voters than *unsuccessful* users. This implies that having voters with a similar background may be an important factor for *success*, since an audience's decision about the performance of debaters may be influenced by the extent to which their prior beliefs match (Durmus and Cardie, 2018).

[3]We consider issues where users identified their side as either PRO or CON and measure the similarity of their opinion for these issues with their friends and voters.



(a) Opinion Similarity (b) Religious Ideology Match

Figure 5.1: Similarity of Unsuccessful vs. Successsful Users with their Friends and Voters.

Social Interaction

The users interact with each other on the platform in the following ways: 1) debating 2) evaluating the performance of other debaters, 3) commenting on debates, 4) asking/answering opinion questions, 5) voting in polls, 6) creating polls, 7) becoming friends. We present examples for an opinion question, an opinion argument, and a poll topic below:

> **Example Opinion Question.** "Does God exist?" [4]
>
> **Example Opinion Argument.** "He probably does not exist. I don't think that it's possible to say yes or no either way. We can only conclude that there is more logical evidence to say that a God probably does not exist, ..."
>
> **Example Poll Topic.** Do you believe in Evolution or Creationism?

[4]Full discussion on the topic can be found at https://www.debate.org/opinions/does-god-exist.

We hypothesize that modeling these interactions is important to understand the differences between how *successful* and *unsuccessful* users interact on this platform and whether or not these are important factors for success. The ability to interact with others in a myriad of different ways provides users with ample opportunity to learn interesting new strategies and improve their skills over time, as they are exposed to a diverse set of perspectives.

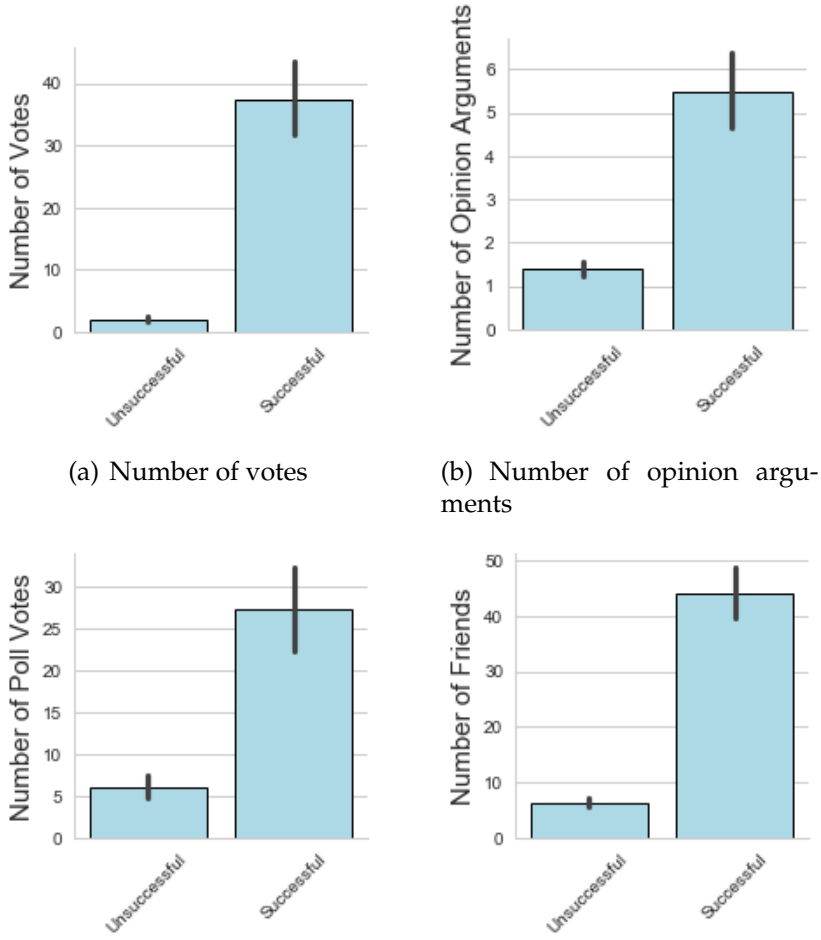
Figure 5.2 shows the interaction statistics for *successful* and *unsuccessful* users.[5] We see that, overall, *successful* users have significantly higher participation on the platform.

Friendship network. We represent the friendship network as an **undirected** graph $G = (V, E)$ where V represents the set of users, and E represents the set of edges where $(x, y) \in E$ if $x \in V$ and $y \in V$ are friends.

Voter network. We represent the voter network as a weighted **directed** graph $G = (V, E)$ where V represents the set of users, and E represents the set of edges where $(x, y) \in E$ if $x \in V$ voted in a debate in which $y \in V$ participated as a debater. The weight of the graph represents how many times x voted in debates y was a debater. Note having (x, y) edge in the graph **does not** imply that x voted for y in a debate.

Hubs and authorities in voter network. Using the HITS algorithm (Kleinberg, 1999), we compute hub and authority scores for each node (user) in the voter network graph. We expect that users that participate in debates as debaters are the authoritative sources of information on the controversial topics on this platform; therefore, they should have higher authority scores. On the

[5]We controlled for number of debates to remove the effect of "being a new user" by pairing *successful* and *unsuccessful* users according the number of debates they participated in.



(a) Number of votes

(b) Number of opinion arguments

(c) Number of poll votes

(d) Number of friends

Figure 5.2: Interaction statistics for *unsuccessful* and *successful* users. *Successful* users have significantly higher participation on the platform than *unsuccessful* users.

other hand, users with higher hub scores represent people who may not necessarily be authoritative sources of information on the topic, but they are interested in the topic and; therefore, by providing feedback, they lead other users to these debates. We find that *successful* users have, on average, a significantly higher hub score than *unsuccessful* users ($p < 0.001$). As shown in Figure 5.3, we further observe that *successful* users have, on average, a significantly higher in-degree centrality and out-degree centrality than *unsuccessful* users in the voter

network. Similarly, *successful* users have higher degree centrality and page rank than *unsuccessful* users in their friendship network.

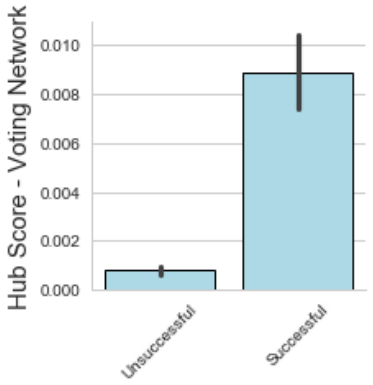
Language

To capture the linguistic style of the debaters' language and its relationship to their *success*, we use textual features that encode 1) users' own language and 2) the interplay between users' and their opponents' language.

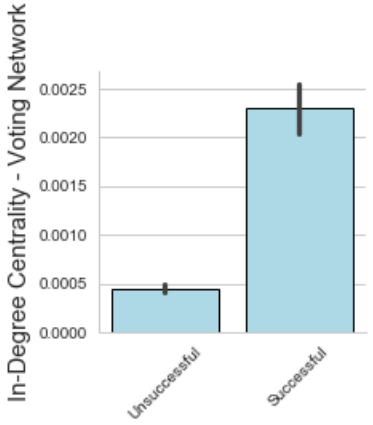
Aspect	Features
Personal Traits	1) **match of the personal traits (e.g., gender, political ideology, religious ideology and ethnicity) with friends and voters**. 2) **opinion similarity with friends and voters**.
Social Interactions	1) **participation features** : # of comments, # of votes, # of friends, # of opinion questions and arguments, # of voted debates, # of poll votes and topics. 2) **friendship network features** : degree, degree centrality, page rank scores. 3) **voter network features**: in-degree, out-degree, in-degree centrality, out-degree centrality, page rank, hub and authority scores.
Language	1) **features of debaters' own language** : # of words, # of definite articles, # of indefinite articles, # of person pronouns, # of positive words, # of negative words, # of hedges, # of swear words, # of punctuation, # of links, average sentiment, type-token ratio, # of quotes, distribution of POS tags, distribution of named entities, BOW. 2) **features to encode the interplay** : exact content word match, exact stop word match, content word match with synonyms.

Table 5.1: Personal Traits, Social Interactions and Language Features.

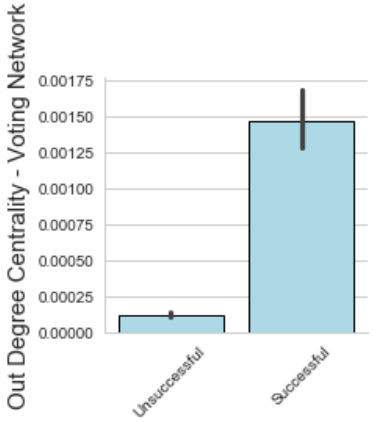
Modeling users' own language. We extract features from the text of users' debates, opinion questions, opinion arguments, poll votes, and poll topics. These features include # of words, word category features (e.g., # of personal



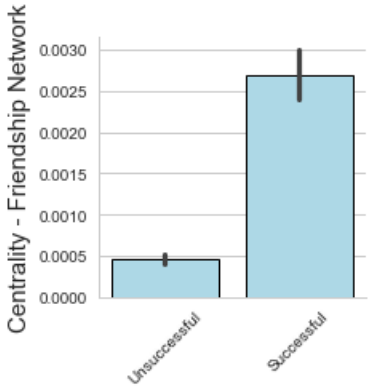
(a) Hub Score - Voter Network



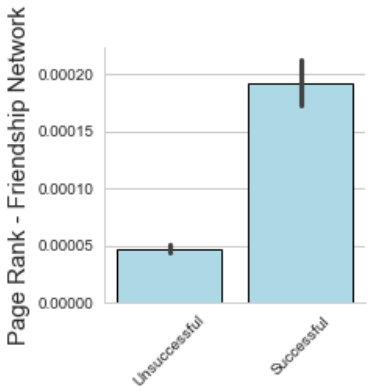
(b) In-degree Centrality - Voter Network



(c) Out-degree Centrality - Voter Network



(d) Centrality - Friendship Network



(e) Page Rank - Friendship Network

Figure 5.3: Characteristics of voter and friendship network for *successful* and *unsuccessful* users.

pronouns, # of positive and negative words), structural features (e.g., distribution of POS tags and named entities), and features to encode the characteristics of the entire language (e.g., type-token ratio)

Modeling interplay between a debater and their opponent. We measure the interplay between debaters and their opponents by measuring how similar a debater's language is to the previous statement made by her opponent. To measure the similarity of a debater's language (D) to that of the opponent's (O) in a round, we look at # of content words that are in both D and O, # of stop words that are in both D and O and # of content words that are in D and have synonyms in O.

The *content word match with synonyms* feature aims to capture the cases where the opponent refers to similar concepts but does not necessarily use the same words as the debater.

The complete list of features modeling the aspects of personal traits, social interactions, and language features is shown in Table 5.1.

5.2.4 Prediction Results

We use weighted logistic regression and choose the amount and type of regularization ($\ell 1$ or $\ell 2$) by grid search over five cross-validation folds. We compute **weighted** precision, recall and F1 scores.

In SETTING 1, we create user pairs (u_1,u_2) where:

- u_1 and u_2 have an equal number of debates they participated in as debaters.

- One of u_1 or u_2 is *successful* and the other one is *unsuccessful* over the second and third stage of their lifetime.[6]

In SETTING 2, in addition to the requirements of SETTING 1, we also require u_1 and u_2 to both have *success prior* ≤ 0.3.

Task. For both SETTING 1 and SETTING 2, we aim to predict whether u_1 or u_2 is *successful* over the second and third stage of her lifetime.

In SETTING 2, by only studying user pairs with low *success priors*, we aim to understand the factors that are important for a user to improve as a debater over time.

Results for SETTING 1

Table 5.2 shows the results for SETTING 1. We compare our model with three simple baselines – majority, debating experience, and success prior. For the majority baseline, we predict the most common label in the training data for each test example. For debating experience baseline, we use # of debates as the only feature to predict the *successful* debater. For success prior baseline, we pick the user with the higher *success prior* as successful.

In SETTING 1, since we do not control for the *success* in the first life stage, we see that the *success prior* information alone can achieve 63.63% F1 score. This implies that there is a correlation between users' *success* in their early life stage and later life stages. This factor may be related to users' prior debating skills.

[6]We consider success only over the second and third stage of users' lifetime in our prediction task, in order to study the effect of *success prior* vs. the other aspects. We use the success in the first life stage as *success prior*.

	Feature	Precision	Recall	F1
	(1) Majority	$26.47_{\pm 1.11}$	$51.44_{\pm 1.08}$	$34.95_{\pm 1.22}$
	(2) Debating experience	$52.70_{\pm 2.91}$	$52.04_{\pm 1.77}$	$41.76_{\pm 2.06}$
	(3) Success prior	$65.20_{\pm 0.77}$	$64.39_{\pm 0.65}$	$63.63_{\pm 0.50}$
Personal Traits	(4) Overall similarity with voters	$61.93_{\pm 1.60}$	$60.86_{\pm 1.70}$	$59.44_{\pm 1.67}$
	(5) Overall similarity with friends	$62.70_{\pm 0.86}$	$59.98_{\pm 1.05}$	$56.94_{\pm 1.14}$
Social Interaction	(6) Participation features	$67.78_{\pm 1.66}$	$66.02_{\pm 2.33}$	$64.82_{\pm 2.70}$
	(7) Friendship network features	$64.23_{\pm 1.40}$	$63.60_{\pm 1.40}$	$62.92_{\pm 1.35}$
	(8) Voter network features	$72.39_{\pm 0.19}$	$70.75_{\pm 0.34}$	$70.20_{\pm 0.70}$
	(6) + (7) + (8)	$72.67_{\pm 0.73}$	$72.29_{\pm 0.93}$	$72.12_{\pm 1.03}$
Language	(9) # of words	$70.37_{\pm 1.41}$	$70.15_{\pm 1.55}$	$69.97_{\pm 1.59}$
	(10) Features of debaters' interplay	$62.11_{\pm 1.09}$	$62.07_{\pm 1.03}$	$61.92_{\pm 1.01}$
	(11) Features of debaters' own language	$72.65_{\pm 2.45}$	$72.66_{\pm 2.45}$	$72.64_{\pm 2.44}$
Combinations	(6) + (7) + (8) + (11)	$78.49_{\pm 1.29}$	$78.46_{\pm 1.32}$	$78.45_{\pm 1.32}$
	(6) + (7) + (8) + (10) + (11)	$\mathbf{81.63_{\pm 1.63}}$	$\mathbf{81.62_{\pm 1.65}}$	$\mathbf{81.61_{\pm 1.65}}$

Table 5.2: Prediction Task Results for SETTING 1. Voter network features are the most predictive social interaction features. Combining interaction and language features achieves the best predictive performance.

We observe that the features that encode debaters' overall similarity with voters and friends achieve significantly better F1 scores than majority and debating experience baselines. However, these features do not have as high a predictive power as the *success prior*. We perform an ablation study for participation features, friendship network features, and voter network features. We find that voter network features are significantly more predictive than the baselines, personal trait features, and other social interaction features. We also perform an ablation study for the language features and find that # of words is a very predictive feature of *success*. When we combine the language features with the interaction features, we get the best predictive performance (81.61% F1 score) for this task which is significantly better than the baselines. This indicates that

	Feature	Precision	Recall	F1
	(1) Majority	$26.67_{\pm 1.61}$	$51.62_{\pm 1.56}$	$35.16_{\pm 1.76}$
	(2) Debating experience	$46.00_{\pm 0.89}$	$50.16_{\pm 1.02}$	$38.98_{\pm 3.92}$
	(3) Success prior	$55.60_{\pm 0.93}$	$55.07_{\pm 0.39}$	$52.10_{\pm 0.47}$
Personal Traits	(4) Overall similarity with voters	$56.55_{\pm 2.43}$	$55.69_{\pm 1.31}$	$52.68_{\pm 1.47}$
	(5) Overall similarity with friends	$55.87_{\pm 3.43}$	$54.23_{\pm 2.35}$	$47.52_{\pm 3.18}$
Social Interactions	(6) Participation features	$59.39_{\pm 4.09}$	$57.68_{\pm 2.34}$	$55.08_{\pm 3.16}$
	(7) Friendship network features	$57.94_{\pm 1.87}$	$57.16_{\pm 1.50}$	$55.41_{\pm 1.67}$
	(8) Voter network features	$70.54_{\pm 1.78}$	$69.91_{\pm 1.79}$	$69.65_{\pm 1.76}$
	(6) + (7) + (8)	$71.66_{\pm 0.71}$	$71.47_{\pm 0.51}$	$71.38_{\pm 0.51}$
Language	(9) # of words	$65.78_{\pm 0.85}$	$64.99_{\pm 1.03}$	$64.41_{\pm 1.16}$
	(10) Features of debaters' interplay	$57.47_{\pm 1.42}$	$57.16_{\pm 1.31}$	$56.41_{\pm 1.29}$
	(11) Features of debaters' own language	$64.48_{\pm 0.74}$	$64.37_{\pm 0.90}$	$64.24_{\pm 0.97}$
Combinations	(6) + (7) + (8) + (11)	$75.44_{\pm 0.90}$	$75.44_{\pm 0.90}$	$75.43_{\pm 0.89}$
	(6) + (7) + (8) + (10) + (11)	$\mathbf{78.06_{\pm 0.88}}$	$\mathbf{78.05_{\pm 0.89}}$	$\mathbf{78.05_{\pm 0.88}}$

Table 5.3: Prediction Task Results for SETTING 2. Similar to SETTING 1, voter network features are the most predictive social interaction features, and combining interaction and language features achieves the best predictive performance.

it is important to account for social interaction and language factors to determine the *successful* debater since these two components encode different kinds of information about the users.

Results for SETTING 2

In this task, by controlling for *prior success*, we aim to understand the factors correlated with *success* by reducing the effect of prior debating skills of the users. As shown in Table 5.3, the F1 score for the *success prior* baseline is not as quite as high as in SETTING 1, since we control for this aspect by ensuring both users in the pair are *unsuccessful* in their initial life stage. However, this does not nec-

essarily mean that the two paired users will have the same success prior, which explains why success prior still performs better than the other baselines. We do not observe any significant difference between the performance of the features encoding personal traits, participation, and the baseline. However, consistent with the SETTING 1, we see that features of the voter network are significantly better (69.65%) in predicting *success*. Although language features achieve a significantly better F1 score than the baseline, they perform significantly worse than the voter network features. Similar to SETTING 1, combining these language features with the social interaction features improves the performance significantly (78.05% F1 score).

Feature Analysis

To understand the important social interaction and language features, we 1) compute the correlation coefficients for the feature values and the labels, 2) analyze the coefficients of the logistic regression classifier, and 3) apply the recursive feature elimination method (Guyon et al., 2002) to rank features according to their importance. In this section, we present the consistently important features for each of these methods.

Analysis of Social Interaction Features. We find that the **most important** social interaction features for SETTING 1 are authority score, hub score, in and out-degree centrality and the page rank of the voter network. Note that all these important features are **positively correlated** with *success*. Although participation and friendship network features (e.g., # of voted debates, degree of the user node in friendship network) are also positively correlated with *success*, the correlation values for these are not as high as the ones of the voter network features.

We also find a high correlation between some of the user activities. For example, users with more # of comments are more active in making friends, voting, providing poll votes, and having higher centrality value in the friendship network. Perhaps surprisingly, we do not observe any correlation between # of voted debates and hub/authority scores in the voter network. However, we see a highly positive correlation between hub scores, authority scores, in-degree centrality, out-degree centrality, and page rank values of the voter network. This implies that *success* is not only about the quantity of voted debates but also about the characteristics of the debaters involved in these debates, since the hub score of a user is influenced by the authority scores of the debaters they vote for. Similarly, the authority score of a user is influenced by the hub scores of the voters that participate in her debates. Therefore, besides the frequency of interaction, the type of the interaction and characteristics of users involved in the interaction are important to consider. Consistent with SETTING 1, in SETTING 2, the most important features (positively correlated with *success*) are authority score, hub score, in and out-degree centrality and the page rank of the voter network. We observe the same patterns of user activities and authority and hub scores as in SETTING 1.

Analysis of Language Features. We find that number of words is positively correlated with *success*. It may be the case that longer text may convey more information and explain the points more explicitly (O'Keefe, 1997, 1998). The bag of words feature is not as predictive as the # of words feature. For both SETTING 1 and SETTING 2, we observe that the value of average sentiment is negatively correlated with *success*. The reason for this may be that negative information is more attention grabbing than positive information (Ditto and F. Lopez, 1992; Homer and Yoon, 1992; Pratto and P. John, 1991) since people are more used

to seeing arguments that are phrased in a more positive way (Meyerowitz and Chaiken, 1987). We also find that *type-token ratio* (diversity of language) is negatively correlated with *success* for both settings. It may be the case that people who talk about a smaller set of topics gain expertise on these topics over time; therefore, they may be more *successful*. We observe that other textual features are positively correlated with *success* for both of these settings. However, the degree of correlation is not as high as it is for type-token ratio and sentiment.

	Feature	Precision	Recall	F1
	(1) Majority	$26.97_{\pm 2.69}$	$51.86_{\pm 2.62}$	$35.46_{\pm 2.95}$
	(2) Debating experience	$53.77_{\pm 2.95}$	$52.43_{\pm 2.91}$	$43.02_{\pm 6.19}$
	(3) Success prior	$39.94_{\pm 7.63}$	$51.00_{\pm 2.23}$	$36.04_{\pm 2.35}$
Personal Traits	(4) Overall similarity with voters	$55.17_{\pm 1.58}$	$55.00_{\pm 2.36}$	$53.94_{\pm 2.99}$
	(5) Overall similarity with friends	$66.38_{\pm 4.11}$	$63.43_{\pm 2.77}$	$60.87_{\pm 3.33}$
Social Interactions	(6) Participation features	$68.88_{\pm 3.57}$	$68.00_{\pm 2.86}$	$67.88_{\pm 2.96}$
	(7) Friendship network features	$65.60_{\pm 4.83}$	$64.00_{\pm 3.81}$	$62.81_{\pm 3.73}$
	(8) voter network features	$64.36_{\pm 1.57}$	$62.72_{\pm 2.37}$	$61.44_{\pm 2.87}$
	(6) + (7) + (8)	$67.80_{\pm 1.86}$	$67.14_{\pm 1.43}$	$66.97_{\pm 1.42}$
Language	(9) # of words	$67.63_{\pm 3.90}$	$66.57_{\pm 2.70}$	$66.29_{\pm 2.39}$
	(10) Features of debaters' interplay	$58.76_{\pm 2.03}$	$57.43_{\pm 0.86}$	$56.60_{\pm 0.93}$
	(11) Features of debaters' own language	$68.47_{\pm 0.21}$	$68.14_{\pm 0.14}$	$68.10_{\pm 0.17}$
Combinations	(6) + (7) + (8) + (11)	$69.32_{\pm 2.48}$	$69.00_{\pm 2.42}$	$69.00_{\pm 2.41}$
	(6) + (7) + (8) + (10) + (11)	$\mathbf{73.60_{\pm 0.80}}$	$\mathbf{73.43_{\pm 0.70}}$	$\mathbf{73.43_{\pm 0.72}}$

Table 5.4: Prediction Task Results for loss of *success*. Participation features are the most important social interaction features. Combining the social interaction features with the language features gives the best prediction performance.

5.3 Understanding the loss of *success*

In the previous section, we show that social interaction and language features are important to predict *successful* debaters. Our findings are consistent for the case when 1) we only control for users' debating experience and 2) we also control for users' *success prior*. Users' participation, the types of interactions they have on the platform, and the characteristics of the users they interact with are predictive of their *success*, regardless of their prior expertise in debating (encoded by the *success prior*).

In SETTING 1, since we did not control for the *success prior*, we studied the factors that are important for a user to become *successful* in their second and third life stages, regardless of their *success* in the beginning. In SETTING 2, we studied the factors that are important for *unsuccessful* users to improve their performance and become *successful* over time. As a natural follow-up, we would also like to understand what factors are correlated with users who are initially *successful*, but later become *unsuccessful* in their lifetime. To do that, in SETTING 3, in addition to the requirements of SETTING 1, we have an additional criterion for all user pairs (u_1,u_2):

- u_1 and u_2 both have *success prior* ≥ 0.7.[7]

[7] We have 700 user pairs with these criteria.

5.3.1 Results

As shown in Table 5.4, features of personality traits, social interactions, and language perform significantly better than the baselines. For this task, the *success prior* baseline performs relatively worse than in the previous two settings. Upon closer examination, we observed that the variance of success priors for this task is an order of magnitude smaller than in SETTING 2. Therefore, as a possible explanation, the *success prior* may not be as predictive for this task.

In social interaction features, similarity with friends is the most predictive feature. However, participation features perform significantly better than the features of personal traits. For this task, contrary to SETTING 1 and SETTING 2, we see that participation features are the most predictive in the set of social interaction features. This implies that a user's participation is important for them to remain *successful*. Lower participation could be a contributing factor for these users to become unsuccessful eventually. Although friendship and voter network features are still significantly more predictive than the baselines, they are not as highly predictive as the participation features. For users with high *success priors*, continued participation may be the most important aspect of their social interaction. We observe that language features alone achieve a similar performance as the social interaction features. Consistent with the SETTING 1 and SETTING 2, combining social interaction and language features gives the best predictive performance (73.43% F1 score).

Analysis of Social Interaction Features. The **most important** social interaction features include # of voted debates, degree of the user node in the friendship network, and hub scores, authority scores, in-degree centrality, out-degree

centrality and page rank values of voter network. All these features indicate higher participation on the platform, and they are positively correlated with staying *successful*. Although the other social interaction features, such as authority and hub scores of the voter network are also positively correlated with *success*, the value of correlation for these is not as high as the previously mentioned features. For users who are initially *unsuccessful*, participation alone may not be enough for them to become *successful* debaters – the types of interactions and the characteristics of people with whom they interact are crucially important for their *success*. On the other hand, users who are initially *successful* may already be experienced debaters, and staying active and participating may be sufficient for them to remain *successful*.

Analysis of Language Features. As in SETTING 1 and SETTING 2, # of words is positively correlated with staying *successful*. We find that the # of first person pronouns is the language feature with the highest positive correlation with staying *successful*. We observe that users who refer to their personal experiences and opinions use first person pronouns more often. It may be the case that debaters may try to appeal to logos by citing personal experience (Cooper and Nothstine, 1992). Consistent with SETTING 1 and SETTING 2, the value of average sentiment is negatively correlated with staying *successful*.

5.4 Limitations

In this study, we investigate the impact of social interaction on debating success. We find that higher participation and engagement improves the success of debaters over time. One potential reason is that users develop strategies to

improve their debating skills. Another factor could be that users only participate in the topics they are comfortable with and do not improve their debating skills overall. Moreover, they may be debating with users that they are confident about defeating to increase their chances of winning. Therefore, in this setup, becoming more successful over time may not necessarily imply developing better argumentative skills. In future work, we would like to explore the effect of debate topics on users' success. Moreover, we aim to understand what characteristics of a user's language change over time and how it affects debating success.

5.5 Chapter Summary

This chapter explores the effect of a user's social interaction on their success in debating over time. We investigate the impact of language, personal traits, and social interaction simultaneously for predicting the successful debater given a pair of debaters where one of them is successful and the other is unsuccessful. We observe that successful debaters are significantly more engaged with others and more active on the platform. We find that a user's social interaction characteristics play a crucial role in determining their success in debates. We achieve the best predictive performance by combining social interaction features with features that encode information on language use.

CHAPTER 6

MODELING PRAGMATIC CONTEXT IN ARGUMENT IMPACT PREDICTION

In the previous chapters, we discuss the impact of prior beliefs (Chapter 4) and social interactions (Chapter 5) on determining the more successful debater and a user's debating success over time, respectively. This chapter introduces a new dataset for argument impact prediction and explores methods to incorporate pragmatic context in determining argument impact.

6.1 Background

Previous work in the social sciences and psychology has shown that the impact and persuasive power of an argument depend not only on the language employed but also on the credibility and character of the communicator (i.e., ethos) (Chaiken, 1979, 1980; Miller et al., 1976), the traits and prior beliefs of the audience (Correll et al., 2004; Davies, 1998; Hullett, 2005; Lord et al., 1979), and the pragmatic context in which the argument is presented (i.e., kairos) (Haugtvedt and Wegener, 1994; Joyce and Harwood, 2014).

Research in Natural Language Processing (NLP) has only partially corroborated these findings. One very influential line of work, for example, develops computational methods to automatically determine the linguistic characteristics of persuasive arguments (Habernal and Gurevych, 2016a; Tan et al., 2016; Zhang et al., 2016), but it does so without controlling for the audience, the communicator, or the pragmatic context. Very recent work, on the other hand, shows that attributes of both the audience and the communicator constitute impor-

tant cues for determining argument strength (Durmus and Cardie, 2018; Lukin et al., 2017). They further show that audience and communicator attributes can influence the relative importance of linguistic features for predicting the persuasiveness of an argument. These results confirm previous findings in the social sciences that show a person's perception of an argument can be influenced by their background and personality traits. To the best of our knowledge, however, no NLP studies explicitly investigate the role of *kairos* — a component of pragmatic context that refers to the context-dependent "timeliness" and "appropriateness" of an argument and its claims within an argumentative discourse — in argument quality prediction.

Among the many social science studies of attitude change, the order in which argumentative claims are shared with the audience has been studied extensively: Haugtvedt and Wegener (1994), for example, summarize studies showing that the argument-related claims a person is exposed to beforehand can affect his perception of an alternative argument in complex ways. Joyce and Harwood (2014) similarly finds that changes in an argument's context can have a big impact on the audience's perception of the argument.

Some recent studies in NLP have investigated the effect of interactions on the overall persuasive power of posts in social media (Hidey and McKeown, 2018; Tan et al., 2016). However, in social media, not all posts have to express arguments or stay on topic (Rakshit et al., 2017), and qualitative evaluation of the posts can be influenced by many other factors such as interaction between the individuals (Durmus and Cardie, 2019b). Therefore, it is difficult to measure the effect of argumentative pragmatic context alone in argument quality prediction without these confounding factors using the datasets and models presented

in prior work.

In this chapter, we study the role of kairos on argument quality prediction by examining the individual claims of an argument for their timeliness and appropriateness in the context of a particular line of argument. We define **kairos** as the sequence of **argumentative** text (e.g., claims) along a particular line of argumentative reasoning. We first present a dataset extracted from *kialo.com* of over 47,000 claims that are part of a diverse collection of arguments on 741 controversial topics. The website's structure dictates that each argument must present a supporting or opposing claim for its parent claim, and stay within the topic of the main thesis. Rather than being posts on a social media platform, these are community-curated claims. Furthermore, for each presented claim, the audience votes on its impact within the given line of reasoning. Critically then, the dataset includes the argument context for each claim, allowing us to investigate the characteristics associated with impactful arguments.

With the dataset in hand, we then propose the task of studying the characteristics of impactful claims by (1) taking the argument context into account, (2) studying the extent to which this context is important, and (3) determining the representation of context that is more effective. To the best of our knowledge, ours is the first dataset that includes claims with both impact votes and the corresponding context of the argument.

6.2 Dataset

Claims and impact votes. We collected claims from *kialo.com*[1][2] for 741 controversial topics and their corresponding impact votes. The users of the platform provide impact votes to evaluate how impactful a particular claim is. Users can pick one of 5 possible impact labels for a particular claim: NO IMPACT, LOW IMPACT, MEDIUM IMPACT, HIGH IMPACT and VERY HIGH IMPACT. While evaluating the impact of a claim, users have access to the full argument context. Therefore, they can assess how impactful a claim is in the given context of an argument. Interestingly, in this dataset, the same claim can have different impact labels depending on the context in which it occurs.

Figure 6.1 shows a partial **argument tree** for the argument **thesis** "PHYSICAL TORTURE OF PRISONERS IS AN ACCEPTABLE INTERROGATION TOOL.". Each node in the argument tree corresponds to a claim, and these argument trees are constructed and edited collaboratively by the users of the platform.

Except for the thesis, every claim in the argument tree either opposes or supports its parent claim. Each path from the root to a leaf node corresponds to an **argument path** which represents a particular line of reasoning on the given controversial topic.

The distribution of argument trees for a given range of claims, and depth is shown in Figures 6.2(a) and 6.2(b) respectively. We see that for the majority of trees, the depth is 4 or higher, and the number of claims is greater than 30.

[1]The data is collected from this website in accordance with the terms and conditions.

[2]There is prior work by Durmus et al. (2019a) which created a dataset of argument trees from *kialo.com*. That dataset, however, does not include any impact labels.

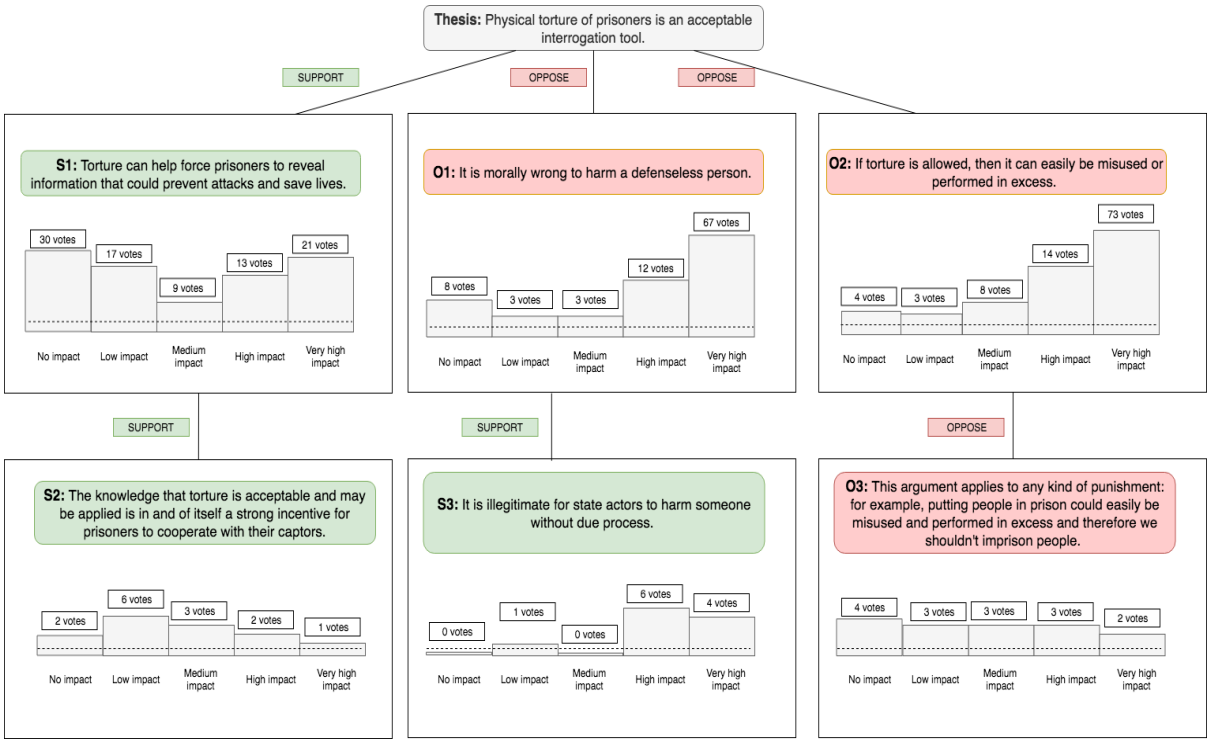
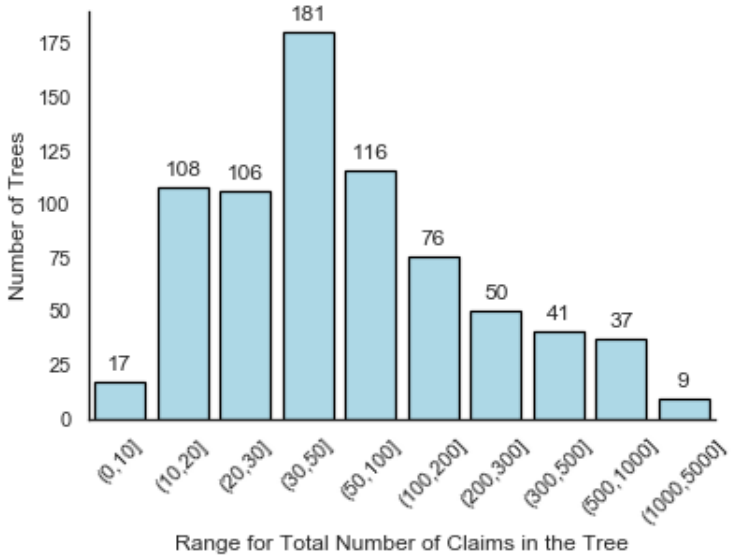


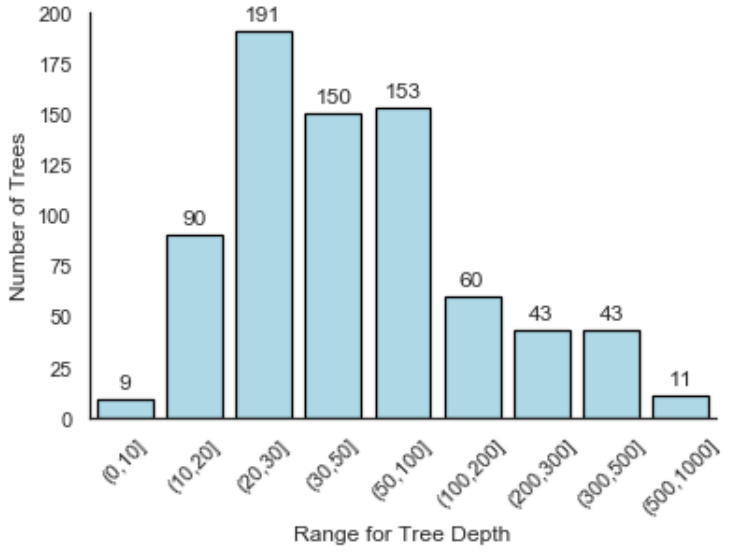
Figure 6.1: Example partial argument tree with claims and corresponding impact votes for the thesis "PHYSICAL TORTURE OF PRISONERS IS AN ACCEPTABLE INTERROGATION TOOL.".

Figure 6.3 shows the total number of claims at a given depth. We see that only $7,618$ out of $95,312$ claims directly support or oppose the theses of the controversial topics. The majority of the claims lie at depth 3 or higher. This shows that the dataset has a rich set of supporting and opposing claims not only for the theses but for claims at different depths of the tree.

Moreover, around 47,000 claims in this dataset have **impact votes** assigned by the users of the platform. The impact vote evaluates how impactful a claim is within its context, which consists of its predecessor claims from the thesis of the tree. For example, claim **O1** "IT IS MORALLY WRONG TO HARM A DEFENSELESS PERSON" is an opposing claim for the thesis, and it is an IMPACTFUL CLAIM since most of its impact votes belong to the category of VERY HIGH IMPACT. However,



(a) Number of trees with given range of total number of claims.



(b) Number of trees with given range of depth.

Figure 6.2: Data statistics: For the majority of trees, the depth of the argument tree is 4 or higher, and the argument tree has more than 30 claims in the tree. Average number of claims and depth per argument tree are 127 and 5 respectively.

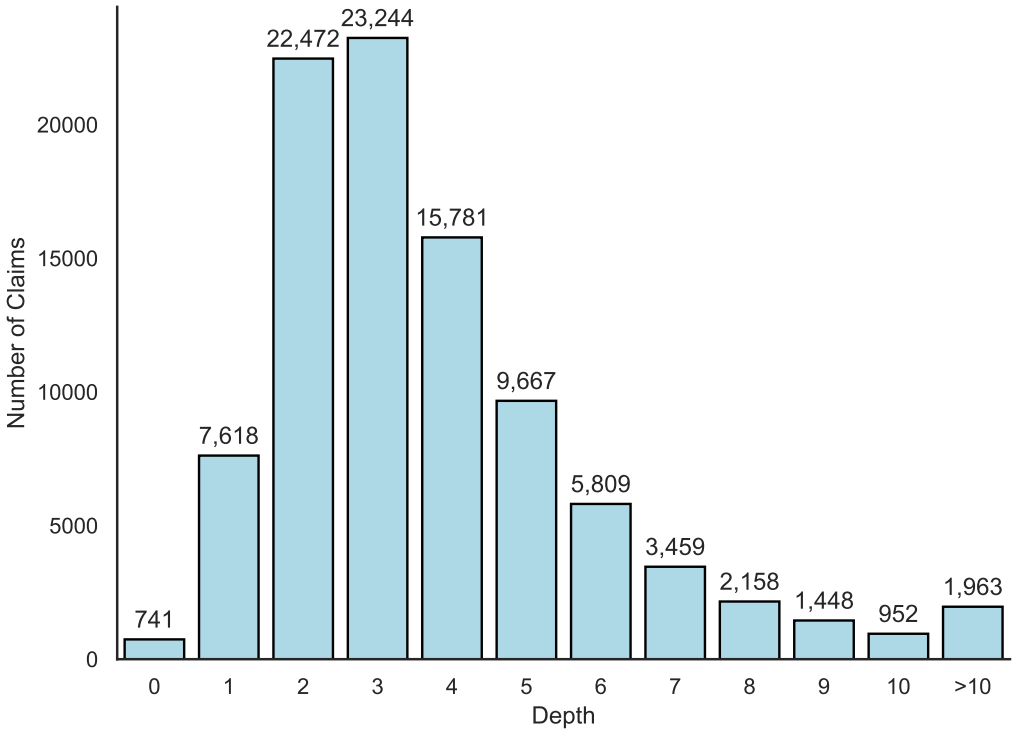


Figure 6.3: Number of claims at given depths.

claim **S3** "IT IS ILLEGITIMATE FOR STATE ACTORS TO HARM SOMEONE WITHOUT THE PROCESS" is a supporting claim for its parent **O1** and it is a less impactful claim since most of the impact votes belong to the NO IMPACT and LOW IMPACT categories.

Impact label statistics. Table 6.3 shows the distribution of the number of votes for each of the impact categories. The claims have 241, 884 total votes. The majority of the impact votes belong to MEDIUM IMPACT category. We observe that users assign more HIGH IMPACT and VERY HIGH IMPACT votes than LOW IMPACT and NO IMPACT votes respectively. When we restrict the claims to the ones with at least 5 impact votes, we have 213, 277 votes in total[3].

[3]26,998 of them NO IMPACT, 33,789 of them LOW IMPACT, 55,616 of them MEDIUM IMPACT,

# impact votes	# claims
[3, 5)	4,495
[5, 10)	5,405
[10, 15)	5,338
[15, 20)	2,093
[20, 25)	934
[25, 50)	992
[50, 333)	255

Table 6.1: Number of claims for the given range of number of votes. There are 19,512 claims in the dataset with 3 or more votes. Out of the claims with 3 or more votes, majority of them have 5 or more votes.

	3-class case	5-class case
Agreement score	Number of claims	Number of claims
> 50%	10,848	7,304
> 60%	7,386	4,329
> 70%	4,412	2,195
> 80%	2,068	840

Table 6.2: Number of claims, with at least five votes, above the given threshold of agreement percentage for 3-class and 5-class cases. When we combine the low impact and high impact classes, there are more claims with high agreement score.

Agreement for the impact votes. To determine the agreement in assigning the impact label for a particular claim, for each claim, we compute the percentage of the votes that are the same as the majority impact vote for that claim. Let c_i denote the count of the claims with the class labels C=[NO IMPACT, LOW IMPACT, MEDIUM IMPACT, HIGH IMPACT, VERY HIGH IMPACT] for the impact label l at index i.

$$\text{Agreement} = 100 * \frac{\max_{0 \leq i \leq 4} c_i}{\sum_{i=0}^{4} c_i}\% \tag{6.1}$$

For example, for claim S1 in Figure 6.1, the agreement score is $100 * \frac{30}{90}\% =$

47,494 of them HIGH IMPACT and 49,380 of them VERY HIGH IMPACT.

33.33% since the majority class (NO IMPACT) has 30 votes and there are 90 impact votes in total for this particular claim. We compute the agreement score for the cases where (1) we treat each impact label separately (5-class case) and (2) we combine the classes HIGH IMPACT and VERY HIGH IMPACT into a one class: IMPACTFUL and NO IMPACT and LOW IMPACT into a one class: NOT IMPACTFUL (3-class case).

Table 6.2 shows the number of claims with the given agreement score thresholds when we include the claims with at least 5 votes. There are more claims with high agreement scores when we combine the low impact and high impact classes. This may imply that distinguishing between no impact-low impact and high impact-very high impact classes is difficult. In our experiments, we use a 3-class representation for the impact labels to decrease the sparsity issue. Moreover, to have a more reliable assignment of impact labels, we consider only the claims with have more than 60% agreement.

Context. In an argument tree, the claims from the thesis node (root) to each leaf node form an argument path. This argument path represents a particular line of reasoning for the given thesis. Similarly, for each claim, all the claims along the path from the thesis to the claim, represent the **context** for the claim. For example, in Figure 6.1, the context for **O1** consists of only the thesis, whereas the context for **S3** consists of both the thesis and **O1** since **S3** is provided to support the claim **O1** which is an opposing claim for the thesis.

Distribution of impact votes. The distribution of claims with the given range of number of impact votes are shown in Table 6.1. There are 19,512 claims in total with 3 or more votes. Out of the claims with 3 or more votes, majority of them have 5 or more votes. We limit our study to the claims with at least 5

Impact label	# votes- all claims
No impact	32,681
Low impact	37,457
Medium impact	60,136
High impact	52,764
Very high impact	58,846
Total # votes	241,884

Table 6.3: Number of votes for the given impact label. There are 241,884 total votes and majority of them belongs to the category MEDIUM IMPACT.

Context length	# claims
1	1,524
2	1,977
3	1,181
[4, 5]	1,436
(5, 10]	1,115
> 10	153

Table 6.4: Number of claims for the given range of context length, for claims with more than 5 votes and an agreement score greater than 60%.

votes to have a more reliable assignment for the accumulated impact label for each claim.

The claims are not constructed independently from their context since they are written in considering the line of reasoning so far. In most cases, each claim elaborates on the point made by its parent and presents cases to support or oppose the parent claim's points. Similarly, when users evaluate the impact of a claim, they consider if the claim is timely and appropriate given its context. There are cases in the dataset where the same claim has different impact labels when presented within a different context. Therefore, we claim that it is not sufficient to study only the linguistic characteristic of a claim to determine its impact, but it is also necessary to consider its context in determining the impact.

Context length (C_l) for a particular claim C is defined by number of claims

included in the argument path starting from the thesis until the claim C. For example, in Figure 6.1, the context length for **O1** and **S3** are 1 and 2 respectively. Table 6.4 shows number of claims with the given range of context length for the claims with more than 5 votes and 60% agreement score. We observe that more than half of these claims have 3 or higher context length.

6.3 Methodology

6.3.1 Hypothesis and Task Description

Similar to prior work, we aim to understand the characteristics of impactful claims in argumentation. However, we **hypothesize** that the qualitative characteristics of arguments are not independent of the context in which they are presented. To understand the relationship between argument context and the impact of a claim, we aim to incorporate the context along with the claim itself in our predictive models.

Prediction task. Given a claim, we want to predict the impact label that is assigned to it by the users: NOT IMPACTFUL, MEDIUM IMPACT, or IMPACTFUL.

Preprocessing. We restrict our study to claims with at least 5 or more votes and greater than 60% agreement to have a reliable impact label assignment. We have 7,386 claims in the dataset satisfying these constraints[4]. We see that the impact class IMPACFUL is the majority class since around 58% of the claims belong to this category.

[4]We have 1,633 NOT IMPACTFUL, 1,445 MEDIUM IMPACT and 4,308 IMPACFUL claims.

For our experiments, we split our data to train (70%), validation (15%), and test (15%) sets.

6.3.2 Baseline Models

Majority

The majority baseline assigns the most common training example label (HIGH IMPACT) to every test example.

SVM with RBF kernel

Similar to Habernal and Gurevych (2016a), we experiment with SVM with RBF kernel, with features that represent (1) the simple characteristics of the argument tree and (2) the linguistic characteristics of the claim.

The features that represent the simple characteristics of the claim's argument tree include the distance and similarity of the claim to the thesis, the similarity of a claim with its parent, and the impact votes of the claim's parent claim. We encode the similarity of a claim to its parent and the thesis claim with the cosine similarity of their tf-idf vectors. The distance and similarity metrics aim to model whether claims which are more similar (i.e., potentially more topically relevant) to their parent claim or the thesis claim are more impactful.

We encode the quality of the parent claim as the number of votes for each impact class and incorporate it as a feature to understand if it is more likely for a claim to be impactful given an impactful parent claim.

Linguistic features. To represent each claim, we extracted the linguistic features proposed by Habernal and Gurevych (2016a) such as tf-idf scores for unigrams and bigrams, ratio of quotation marks, exclamation marks, modal verbs, stop words, type-token ratio, hedging (Hyland, 1998), named entity types, POS n-grams, sentiment (Hutto and Gilbert, 2014) and subjectivity scores (Wilson et al., 2005), spell-checking, readibility features such as *Coleman-Liau* (Coleman and Liau, 1975), *Flesch* (Flesch, 1948), argument lexicon features (Somasundaran et al., 2007) and surface features such as word lengths, sentence lengths, word types, and number of complex words[5].

FastText

Joulin et al. (2017) introduced a simple yet effective baseline for text classification, which they show to be competitive with deep learning classifiers in terms of accuracy. Their method represents a sequence of text as a bag of n-grams, and each n-gram is passed through a look-up table to get its dense vector representation. The overall sequence representation is simply an average over the dense representations of the bag of n-grams, and is fed into a linear classifier to predict the label. We use the code released by Joulin et al. (2017) to train a classifier for argument impact prediction, based on the claim text[6].

[5]We pick the parameters for the SVM model according to the performance validation split, and report the results on the test split.

[6]We used maxNgram length of 2, learning rate of 0.8, num epochs of 15, vector dim of 300. We also used the pre-trained 300-dim wiki-news vectors made available on the fastText website.

BiLSTM with Attention

Another effective baseline (Yang et al., 2016; Zhou et al., 2016) for text classification consists of encoding the text sequence using a bidirectional Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), to get the token representations in context, and then attending (Luong et al., 2015) over the tokens to get the sequence representation. For the query vector for attention, we use a learned context vector, similar to Yang et al. (2016). We picked our hyperparameters based on performance on the validation set and report our results for the best set of hyperparameters[7]. We initialized our word embeddings with glove vectors (Pennington et al., 2014) pre-trained on Wikipedia + Gigaword, and used the Adam optimizer (Kingma and Ba, 2015) with its default settings.

6.3.3 Fine-tuned BERT model

Devlin et al. (2018) fine-tuned a pre-trained deep bi-directional transformer language model (which they call BERT) by adding a simple classification layer on top and achieved the state of the art results across a variety of NLP tasks. We employ their pre-trained language models for our task and compare them to our baseline models. For all the architectures described below, we fine-tune for 10 epochs, with a learning rate of 2e-5. We employ an early stopping procedure based on the model performance on a validation set.

[7]Our final hyperparams were: 100-dim word embedding, 100-dim context vector, 1 layer BiLSTM with 64 units, trained for 40 epochs with early stopping based on validation performance.

	Precision	Recall	F1
Majority	19.43	33.33	24.55
SVM with RBF Kernel			
Distance from the thesis	27.42	33.53	26.05
Parent quality	58.11	47.85	46.61
Linguistic features	**65.67**	38.58	35.42
BiLSTM with Attention	$46.50_{\pm 0.28}$	$46.35_{\pm 0.99}$	$46.22_{\pm 0.58}$
FastText	$51.18_{\pm 0.80}$	$46.09_{\pm 0.64}$	$47.06_{\pm 0.70}$
BERT models			
Claim only	$53.24_{\pm 1.07}$	$50.93_{\pm 2.01}$	$51.53_{\pm 1.53}$
Claim + Parent	$55.79_{\pm 1.72}$	$53.54_{\pm 2.09}$	$54.00_{\pm 1.79}$
Claim + Context$_f$(2)	$56.57_{\pm 0.85}$	$54.76_{\pm 1.71}$	$55.18_{\pm 0.99}$
Claim + Context$_f$(3)	$57.19_{\pm 0.92}$	$\mathbf{55.77_{\pm 1.05}}$	$\mathbf{55.98_{\pm 0.70}}$
Claim + Context$_f$(4)	$57.09_{\pm 1.71}$	$55.31_{\pm 1.09}$	$55.72_{\pm 1.14}$
Claim + Context$_{gru}$(4)	$54.95_{\pm 2.00}$	$51.55_{\pm 1.27}$	$52.37_{\pm 1.26}$
Claim + Context$_a$(4)	$56.60_{\pm 0.52}$	$54.55_{\pm 0.57}$	$54.65_{\pm 0.33}$

Table 6.5: Results for the baselines and the BERT models with and without the context. Best performing model is BERT with the representation of previous 3 claims in the path along with the claim representation itself. We run the models 5 times and we report the mean and standard deviation.

Claim with no context

In this setting, we attempt to classify the impact of the claim based on the text of the claim only. We follow the fine-tuning procedure for sequence classification detailed in Devlin et al. (2018), and input the claim text as a sequence of tokens preceded by the special [CLS] token and followed by the special [SEP] token. We add a classification layer on top of the BERT encoder, to which we pass the representation of the [CLS] token and fine-tune this for argument impact prediction.

Claim with parent representation

In this setting, we use the parent claim's text, in addition to the target claim text, in order to classify the impact of the target claim. We treat this as a sequence pair classification task and combine both the target claim and parent claim as a single sequence of tokens, separated by the special separator [SEP]. We then follow the same procedure above for fine-tuning.

Incorporating larger context

In this setting, we consider incorporating a larger context from the discourse in order to assess the impact of a claim. In particular, we consider up to four previous claims in the discourse (for a total context length of 5). We attempt to incorporate larger context into the BERT model in three different ways.

Flat representation of the path. The first, simple approach is to represent the entire path (claim + context) as a single sequence, where each of the claims is separated by the [SEP] token. BERT was trained on sequence pairs, and therefore the pre-trained encoders only have two segment embeddings (Devlin et al., 2018). So to fit multiple sequences into this framework, we indicate all tokens of the target claim as belonging to segment A and the tokens for all the claims in the discourse context as belonging to segment B. This way of representing the input requires no additional changes to the architecture or retraining, and we can just fine-tune in a similar manner as above. We refer to this representation of the context as a flat representation, and denote the model as $\text{Context}_f(i)$, where i indicates the length of the context that is incorporated into the model.

	$C_l = 1$	$C_l = 2$	$C_l = 3$	$C_l = 4$
BERT models				
Claim only	$48.61_{\pm 3.16}$	$53.15_{\pm 1.95}$	$54.51_{\pm 1.91}$	$50.89_{\pm 2.95}$
Claim + Parent	$51.49_{\pm 2.63}$	$54.78_{\pm 2.95}$	$54.94_{\pm 2.72}$	$51.94_{\pm 2.59}$
Claim + Context$_f$(2)	$52.84_{\pm 2.55}$	$53.77_{\pm 1.00}$	$55.24_{\pm 2.52}$	$57.04_{\pm 1.19}$
Claim + Context$_f$(3)	$\mathbf{54.88_{\pm 2.49}}$	$54.71_{\pm 1.74}$	$52.93_{\pm 2.07}$	$\mathbf{58.17_{\pm 1.89}}$
Claim + Context$_f$(4)	$54.47_{\pm 2.95}$	$\mathbf{54.88_{\pm 1.53}}$	$\mathbf{57.11_{\pm 3.38}}$	$57.02_{\pm 2.22}$

Table 6.6: F1 scores of each model for the claims with various context length values.

Attention over context. Recent work in incorporating argument sequence in predicting persuasiveness (Hidey and McKeown, 2018) has shown that hierarchical representations are effective in representing context. Similarly, we consider hierarchical representations for representing the discourse. We first encode each claim using the pre-trained BERT model as the claim encoder and use the representation of the [CLS] token as claim representation. We then employ dot-product attention (Luong et al., 2015), to get a weighted representation for the context. We use a learned context vector as the query for computing attention scores, similar to Yang et al. (2016). The attention score α_c is computed as shown below:

$$\alpha_c = \frac{exp(V_c^T V_l)}{\sum_{c \in D} exp(V_c^T V_l)} \tag{6.2}$$

Where V_c is the claim representation that was computed with the BERT encoder as described above, V_l is the learned context vector that is used for computing attention scores, and D is the set of claims in the discourse. After computing the attention scores, the final context representation v_d is computed as follows:

$$V_d = \sum_{c \in D} \alpha_c V_c \tag{6.3}$$

We then concatenate the context representation with the target claim representation $[V_d, V_r]$ and pass it to the classification layer to predict the quality. We

denote this model as Context$_a(i)$.

GRU to encode context Similar to the approach above, we consider a hierarchical representation for representing the context. We compute the claim representations, as detailed above, and we then feed the discourse claims' representations (in sequence) into a bidirectional Gated Recurrent Unit (GRU) (Cho et al., 2014), to compute the context representation. We concatenate this with the target claim representation and use this to predict the claim impact. We denote this model as Context$_{gru}(i)$.

6.4 Results and Analysis

Table 6.5 shows the macro precision, recall, and F1 scores for the baselines as well as the BERT models with and without context representations[8].

We see that *parent quality* is a simple yet effective feature, and the SVM model with this feature can achieve significantly higher ($p < 0.001$)[9] F1 score (46.61%) than *distance from the thesis* and *linguistic features*. Claims with higher impact parents are more likely to have a higher impact. *Similarity with the parent and thesis* is not significantly better than the *majority* baseline. Although the BiLSTM model with attention and FastText baselines performs better than the SVM with *distance from the thesis* and *linguistic features*, it has similar performance to the *parent quality* baseline.

We find that the BERT model with *claim only* representation performs sig-

[8]For the models that result in different scores with a different random seed, we run them 5 times and report the mean and standard deviation.

[9]We perform a two-sided t-test for significance analysis.

nificantly better ($p < 0.001$) than the baseline models. Incorporating the *parent representation* only along with the *claim representation* does not give significant improvement over representing the claim only. However, *incorporating the flat representation of the larger context* along with the claim representation consistently achieves significantly better ($p < 0.001$) performance than the claim representation alone. Similarly, *attention representation* over the context with the learned query vector achieves significantly better performance then the *claim representation* only ($p < 0.05$).

We find that the *flat representation* of the context achieves the highest F1 score. It may be more difficult for the models with a larger number of parameters to perform better than the *flat representation* since the dataset is small. We also observe that modeling 3 claims on the argument path before the target claim achieves the best F1 score (55.98%).

To understand for what kinds of claims the best performing contextual model is more effective, we evaluate the BERT model with *flat context representation* for claims with context length values 1, 2, 3 and 4 separately. Table 6.6 shows the F1 score of the BERT model without context and with *flat context representation* with different lengths of context. For the claims with context length 1, adding $\text{Context}_f(3)$ and $\text{Context}_f(4)$ representation along with the claim achieves significantly better ($p < 0.05$) F1 score than modeling the *claim only*. Similarly for the claims with context length 3 and 4, $\text{Context}_f(4)$ and $\text{Context}_f(3)$ perform significantly better than BERT with *claim only* (($p < 0.05$) and ($p < 0.01$) respectively). We see that models with larger context are helpful even for claims which have limited context (e.g., $C_l = 1$). This may suggest that when we train the models with larger context, they learn how to represent the

claims and their context better.

6.5 Limitations

In this study, we find that incorporating pragmatic context is crucial in impact prediction. First, we present a new dataset for this task. We assume that the impact labels in this dataset are provided in good faith by the users. However, we note that the user demographics on the platform may not have a fair representation, and prior beliefs and background could affect which arguments are perceived as more impactful. We should account for this potential bias while using the systems built from this dataset. We further observe that BERT-based models achieve the best predictive performance. However, it is difficult to interpret these systems to understand what aspect of the context plays an important role. In future work, we aim to employ methods such as local surrogate (Ribeiro et al., 2016) or input saliency models (Li et al., 2016) to interpret these systems.

6.6 Chapter Summary

This chapter proposes a new dataset of arguments along with their impact label and the argument path, representing a particular line of reasoning on the given controversial topic. We further propose predictive models that incorporate the pragmatic and discourse context of argumentative claims to predict argument impact. We show that the models representing the pragmatic context outperform models that rely on only claim-specific linguistic features for predicting the perceived impact of individual claims within a particular line of argument.

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this dissertation, we describe our contributions to understand persuasion in computational argumentation. In particular, we show that the characteristics of the people involved highly influence the process of persuasion. We investigate the impact of speaker and audience factors in predicting the more persuasive debater. We also explore whether a user's social interaction on online argumentation platforms affects their success in persuasion over time. We further propose context-aware models to measure the importance of pragmatic context in predicting the impact of the arguments.

7.1 Summary of Contributions

In Chapter 3, we propose a new dataset of debates with extensive user information extracted from an online argumentation platform (i.e., *debate.org*). This is the largest available dataset with such extensive user information, including political ideology, religious ideology, and stance on various controversial topics. Availability of this information has motivated further research in exploring the effect of user factors in persuasion (Durmus and Cardie, 2019b; Luu et al., 2019).

With the dataset in hand, we study the role of prior beliefs, of both speakers and audience members, on the perceived persuasiveness of arguments. We do this by formulating a new task to determine which debater will be able to persuade a given voter to change their stance. We find that features associated with a user's initial stance are very predictive for this task. This is especially true for debates on political and religious issues, where these features are even

more predictive than linguistic features of the arguments.

In Chapter 5, we further explore whether a user's social interaction impacts their debating success over time on online argumentation platforms. We extract features from a user's friendship and voter networks. We then use these features to explore the role of social interactions as compared to personality traits and language in predicting debating success over time. We find that social interaction features (i.e., primarily features extracted from the voter network) are the most predictive of success. We observe that the best predictive performance is achieved when combining social interaction features with linguistic features. This implies that the characteristics of interactions on online debating platforms are essential to becoming more experienced and successful in persuasion.

Finally, we propose a dataset to study the role of *kairos* (i.e., pragmatic context) in determining argument impact. As described in Chapter 6, the dataset includes the argument context for each claim, along with the impact score within the given line of reasoning. We further explore whether a flat vs. a hierarchical representation of context is more effective for this task. We find that a flat representation of the context achieves the best performance since the dataset may not be large enough to learn the additional parameters needed for a hierarchical model. We observe that models that incorporate context perform significantly better than those that use the claim only. This implies that the context in which an argument is presented is crucial in assessing its impact.

7.2 Ethical Considerations

All the data in our research has been collected and used in accordance with the terms of service of the source. For user studies, we take the utmost care in making sure that the anonymity of the users is preserved. Finally, we make sure that our work does not take a stance on any of the controversial topics, but rather just analyzes the viewpoints of the participants in the datasets we use. One shortcoming we acknowledge is that we are unable to represent all demographics due to a lack of data. The sources we used tend to be highly skewed towards an American audience, and even within this audience, the distribution may not be representative enough.

Given that argumentation is a fundamental part of human communication, the work in this area could be used in both good and ethically less acceptable manners. The driving motivator of this dissertation has always been that argumentation can be used for social good, such as exposing people to diverse viewpoints to help them make more informed decisions or using persuasion to encourage people to contribute to the environment and society. However, even for such use cases, it is vital to be transparent and inform users about the nature of these systems. Moreover, user consent should be required to employ such methods in real-world scenarios.

7.3 Future Directions

Modeling Users in Computational Persuasion. In our study, we explore the role of prior beliefs in persuasion, focusing on political and religious ideologies.

However, there are many aspects of the source and the audience (e.g., education level, prior argumentation skills, credibility, personality traits) that may influence the persuasion process. It is challenging to control for all potential confounding factors to isolate the effect of the linguistic features due to data sparsity. We think it is vital to explore better representations for users to disentangle the impact of user aspects. We further want to explore the following research questions: 1) How do different aspects of users influence their perceptions of the arguments? 2) How do these aspects affect people's language choice while interacting with more similar vs. different people? 3) How does the language use change for different groups of speakers?

Personalized Argument Generation. Understanding the effect of user factors in persuasion could be the first step towards designing personalized argument generation systems capable of conveying relevant and interesting information for a more effective persuasion process (Danilova et al., 2020; Dijkstra, 2008). Personalization is important in increasing engagement and attachment in social interactions on online platforms (Jenny et al., 2018; Kang et al., 2016). Therefore, having personalized systems may increase the quality of persuasive communication and the outcome of this process. For example, Wang et al. (2019) has recently proposed a personalized dialogue system that tries to persuade people to donate to a specific charity. They show that personalized argumentation generation systems can be used for social good. Moreover, such systems could be used to present people with a diverse set of viewpoints to help them make more informed decisions.

Interpretation of Neural Models. Neural networks can model more complex representations that help achieve state-of-the-art performance in various

syntactic and semantic tasks in Natural Language Processing. However, unlike feature-based linear models, it is more challenging to interpret neural models to explain what these models learn and improve them. For example, in Chapter 6, we have found that incorporating context with the argument helps predict its impact. However, it is not straightforward to interpret explicitly which aspects of the context helps to improve the overall performance. Similarly, although neural methods achieve state-of-the-art performance in persuasion prediction tasks, it is difficult to identify the characteristics of persuasive language and effective persuasion strategies. We believe that improved interpretation of neural networks is crucial to draw valuable conclusions in computational persuasion studies and build better models for these tasks.

BIBLIOGRAPHY

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.

Milad Alshomary, Shahbaz Syed, Martin Potthast, and Henning Wachsmuth. 2020. Target inference in argument conclusion generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4345, Online. Association for Computational Linguistics.

David Atkinson, Kumar Bhargav Srinivasan, and Chenhao Tan. 2019. What gets echoed? understanding the "pointers" in explanations of persuasive arguments. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2911–2921, Hong Kong, China. Association for Computational Linguistics.

Lars Backstrom, Eytan Bakshy, Jon M Kleinberg, Thomas M Lento, and Itamar Rosenn. 2011. Center of attention: How facebook users allocate attention across friends.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017a. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 251–261, Valencia, Spain. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Edelstein, Charles Jochim, and Noam Slonim. 2017b. Improving claim stance classification with lexical knowledge expansion and

context utilization. In *Proceedings of the 4th Workshop on Argument Mining*, pages 32–38, Copenhagen, Denmark. Association for Computational Linguistics.

Roy Bar-Haim, Lilach Eden, Roni Friedman, Yoav Kantor, Dan Lahav, and Noam Slonim. 2020. From arguments to key points: Towards automatic argument summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4029–4039, Online. Association for Computational Linguistics.

Pierpaolo Basile, Valerio Basile, Elena Cabrio, and Serena Villata. 2016. Argument Mining on Italian News Blogs. In *Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016)*, Naples, Italy.

Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. 2009. Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM Conference on Internet Measurement*, IMC '09, pages 49–62, New York, NY, USA. ACM.

Moira Burke, Cameron Marlow, and Thomas Lento. 2009. Feed me: motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 945–954. ACM.

Elena Cabrio and S. Villata. 2013. A natural language bipolar argumentation approach to support users in online debate interactions†. *Argument Comput.*, 4:209–230.

Elena Cabrio and Serena Villata. 2018. Five years of argument mining: a data-

driven analysis. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 5427–5433. International Joint Conferences on Artificial Intelligence Organization.

Murray Campbell, A.Joseph Hoane, and Feng hsiung Hsu. 2002. Deep blue. *Artificial Intelligence*, 134(1):57–83.

Lucas Carstens and Francesca Toni. 2015. Towards relation based argumentation mining. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 29–34, Denver, CO. Association for Computational Linguistics.

M. Cha, H. Haddadi, F. Benevenuto, and K.P. Gummadi. 2010. Measuring user influence in twitter: The million follower fallacy. In *4th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Shelly Chaiken. 1979. Communicator physical attractiveness and persuasion. *Journal of Personality and Social Psychology*, 37(8):1387 – 1397.

Shelly Chaiken. 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39:752–766.

Tuhin Chakrabarty, Christopher Hidey, and Kathy McKeown. 2019a. IMHO fine-tuning improves claim detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 558–563, Minneapolis, Minnesota. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019b. AMPERSAND: Argument mining for PERSuAsive oNline discussions. In *Proceedings of the 2019 Conference on Empirical Methods in*

Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Marilyn J. Chambliss and Ruth Garner. 1996. Do adults change their minds after reading persuasive text?. (3):291.

Lord Charles G., Ross Lee, and Lepper Mark R. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, 37(11):2098 – 2109.

Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

R.B. Cialdini. 2001. *Influence: Science and Practice*. Allyn and Bacon.

Meri Coleman and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283 – 284.

M. Cooper and W.L. Nothstine. 1992. *Power Persuasion: Moving an Ancient Art Into the Media Age*. Educational Video Group.

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

Anastasia Danilova, Alena Naiakshina, and Matthew Smith. 2020. One size does not fit all: A grounded theory and online survey study of developer

preferences for security warning types. *2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE), Software Engineering (ICSE), 2020 IEEE/ACM 42nd International Conference on, ICSE*, pages 136 – 148.

M. F Davies. 1998. Dogmatism and belief formation : Output interference in the processing of supporting and contradictory cognitions. *Journal of personality and social psychology*, 75(2):456 – 466.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Arie Dijkstra. 2008. The psychology of tailoring-ingredients in computer-tailored persuasion. *Social and Personality Psychology Compass*, 2(2):765–784.

Peter Ditto and David F. Lopez. 1992. Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions. *Journal of Personality and Social Psychology*, 63:568–584.

Esin Durmus and Claire Cardie. 2018. Exploring the role of prior beliefs for argument persuasion. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2019a. A corpus for modeling user and language effects in argumentation on online debating. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607, Florence, Italy. Association for Computational Linguistics.

Esin Durmus and Claire Cardie. 2019b. Modeling the factors of user success in online debate. In *The World Wide Web Conference*, WWW '19, pages 2701–2707, New York, NY, USA. ACM.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019a. Determining relative argument specificity and stance for complex argumentative structures. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4630–4641, Florence, Italy. Association for Computational Linguistics.

Esin Durmus, Faisal Ladhak, and Claire Cardie. 2019b. The role of pragmatic and discourse context in determining argument impact. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5668–5678, Hong Kong, China. Association for Computational Linguistics.

Mihai Dusmanu, Elena Cabrio, and Serena Villata. 2017. Argument mining on Twitter: Arguments, facts and sources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2317–2322, Copenhagen, Denmark. Association for Computational Linguistics.

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

F.H. van Eemeren and F.H. Eemeren. 2009. *Examining Argumentation in Context: Fifteen Studies on Strategic Maneuvering*. Argumentation in context. John Benjamins Publishing Company.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11–22, Vancouver, Canada. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464, Brussels, Belgium. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the Persuasive Effect of Style in News Editorial Argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Hao Fang, Hao Cheng, and Mari Ostendorf. 2016. Learning latent local conversation modes for predicting comment endorsement in online discussions. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 55–64, Austin, TX, USA. 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.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221 – 233.

Scott A Golder, Dennis M Wilkinson, and Bernardo A Huberman. 2007. Rhythms of social interaction: Messaging within a massive online network. In *Communities and technologies 2007*, pages 41–66. Springer.

Marco Guerini, Gözde Özbal, and Carlo Strapparava. 2015. Echoes of persuasion: The effect of euphony in persuasive communication. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1483–1493, Denver, Colorado. Association for Computational Linguistics.

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan. Asian Federation of Natural Language Processing.

Curtis P. Haugtvedt and Duane T. Wegener. 1994. Message Order Effects in Persuasion: An Attitude Strength Perspective. *Journal of Consumer Research*, 21(1):205–218.

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 and Kathy McKeown. 2019. Fixed that for you: Generating contrastive claims with semantic edits. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1756–1767, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Pamela M. Homer and Sun-Gil Yoon. 1992. Message framing and the interrelationships among ad-based feelings, affect, and cognition. *Journal of Advertising*, 21(1):19–33.

Xinyu Hua, Zhe Hu, and Lu Wang. 2019. Argument generation with retrieval, planning, and realization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2661–2672, Florence, Italy. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2017. Understanding and detecting supporting arguments of diverse types. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 203–208, Vancouver, Canada. Association for Computational Linguistics.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230, Melbourne, Australia. Association for Computational Linguistics.

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

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Ken Hyland. 1998. *Hedging in Scientific Research Articles*. John Benjamins.

Bronstein Jenny, Aharony Noa, and Bar-Ilan Judit. 2018. Politicians' use of facebook during elections : Use of emotionally-based discourse, personalization, social media engagement and vividness. *Aslib Journal of Information Management*, 70(5):551 – 572.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2*,

Short Papers, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Nick Joyce and Jake Harwood. 2014. Context and identification in persuasive mass communication. *Journal of Media Psychology: Theories, Methods, and Applications*, 26:50.

M. (1) Kang, D.-H. (2) Shin, and T. (3) Gong. 2016. The role of personalization, engagement, and trust in online communities. *Information Technology and People*, 29(3):580–596.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jon M Kleinberg. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632.

Jonathan Kobbe, Ioana Hulpuș, and Heiner Stuckenschmidt. 2020. Unsupervised stance detection for arguments from consequences. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 50–60, Online. Association for Computational Linguistics.

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.

Shamanth Kumar, Reza Zafarani, and Huan Liu. 2011. Understanding user migration patterns in social media.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Jialu Li, Esin Durmus, and Claire Cardie. 2020. Exploring the role of argument structure in online debate persuasion. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8905–8912, Online. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Bang Hui Lim, Dongyuan Lu, Tao Chen, and Min-Yen Kan. 2015. # mytweet via instagram: Exploring user behaviour across multiple social networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 113–120. ACM.

Liane Longpre, Esin Durmus, and Claire Cardie. 2019. Persuasion of the undecided: Language vs. the listener. In *Proceedings of the 6th Workshop on Ar-*

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

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

Stephanie 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.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Kelvin Luu, Chenhao Tan, and Noah A. Smith. 2019. Measuring Online Debaters' Persuasive Skill from Text over Time. *Transactions of the Association for Computational Linguistics*, 7:537–550.

Sofus A Macskassy and Matthew Michelson. 2011. Why do people retweet? anti-homophily wins the day!

Marcelo Maia, Jussara Almeida, and Virgílio Almeida. 2008. Identifying user behavior in online social networks. In *Proceedings of the 1st Workshop on Social Network Systems*, SocialNets '08, pages 1–6, New York, NY, USA. ACM.

Riley Matilda White. 1954. Communication and persuasion: Psychological stud-

ies of opinion change. carl i. hovland irving l. janis harold h. kelley. *American Sociological Review*, 19(3):355 – 357.

William J McGuire. 1969. The nature of attitudes and attitude change. (2nd ed., vol.3). *The handbook of social psychology*.

Beth E Meyerowitz and Shelly Chaiken. 1987. The effect of message framing on breast self-examination attitudes, intentions, and behavior. *Journal of personality and social psychology*, 52(3):500.

Norman Miller, Geoffrey Maruyama, Rex J. Beaber, and Keith Valone. 1976. Speed of speech and persuasion. *Journal of Personality and Social Psychology*, 34(4):615–624.

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

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, page 225–230, New York, NY, USA. Association for Computing Machinery.

Gaku Morio, Ryo Egawa, and Katsuhide Fujita. 2019. Revealing and predicting online persuasion strategy with elementary units. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6274–6279, Hong Kong, China. Association for Computational Linguistics.

Meenakshi Nagarajan, Hemant Purohit, and Amit P Sheth. 2010. A qualitative examination of topical tweet and retweet practices.

Huy V. Nguyen and Diane J. Litman. 2018. Argument mining for improving the automated scoring of persuasive essays. 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 5892–5899. AAAI Press.

Vlad Niculae, Joonsuk Park, and Claire Cardie. 2017. Argument mining with structured SVMs and RNNs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 985–995, Vancouver, Canada. Association for Computational Linguistics.

Daniel J. O'Keefe. 1997. Standpoint explicitness and persuasive effect: A meta-analytic review of the effects of varying conclusion articulation in persuasive messages. *Argumentation and Advocacy*, 34(1):1–12.

Daniel J. O'Keefe. 1998. Justification explicitness and persuasive effect: A meta-analytic review of the effects of varying support articulation in persuasive messages. *Argumentation and Advocacy*, 35(2):61–75.

Nathan Ong, Diane Litman, and Alexandra Brusilovsky. 2014. Ontology-based argument mining and automatic essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 24–28, Baltimore, Maryland. Association for Computational Linguistics.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, page 98–107, New York, NY, USA. Association for Computing Machinery.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2018. A corpus of eRulemaking user comments for measuring evaluability of arguments. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A. Peldszus. 2015. An annotated corpus of argumentative microtexts.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. *Int. J. Cogn. Inform. Nat. Intell.*, 7(1):1–31.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

C.P. Perelman and L. Olbrechts-Tyteca. 1971. *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 543–552, Beijing, China. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the 2016 Conference of the North American Chapter*

of the Association for Computational Linguistics: Human Language Technologies, pages 1384–1394, San Diego, California. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2017. Why can't you convince me? modeling weaknesses in unpersuasive arguments. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 4082–4088. AAAI Press.

Richard E. Petty and John T. Cacioppo. 1984. The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology*, 46(1):69 – 81.

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

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

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

Felicia Pratto and Oliver P. John. 1991. Automatic vigilance: The attention-grabbing power of negative social information. *Journal of personality and social psychology*, 61:380–91.

Geetanjali Rakshit, Kevin K. Bowden, Lena Reed, Amita Misra, and Marilyn A. Walker. 2017. Debbie, the debate bot of the future. *CoRR*, abs/1709.03167.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the*

22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Petty Richard E., Cacioppo John T., and Goldman Rachel. 1981. Personal involvement as a determinant of argument-based persuasion. *Journal of Personality and Social Psychology*, 41(5):847 – 855.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 440–450, Lisbon, Portugal. Association for Computational Linguistics.

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

Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. 2011. Influence and passivity in social media. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, page 113–114, New York, NY, USA. Association for Computing Machinery.

Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. End-to-end argument generation system in debating. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 109–114, Beijing, China. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

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.

Elisa Shearer and Katerina Eva Matsa. 2020. News use across social media platforms.

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.

Chaiken Shelly. 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, 39(5):752 – 766.

L. J. Shrum. 2012. Persuasion in the marketplace: How theories of persuasion apply to marketing and advertising.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. An autonomous debating system. *Nature*, 591(7850):379—384.

Stephen M. Smith and David R. Shaffer. 1995. Speed of speech and persuasion: evidence for multiple effects. *Personality & Social Psychology Bulletin*, 21(10):1051.

Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From argumentation mining to stance classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77, Denver, CO. Association for Computational Linguistics.

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

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.

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

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2399–2409, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Paul D. Sweeney and Kathy L. Gruber. 1984. Selective exposure: Voter infor-

mation preferences and the Watergate affair. *Journal of Personality and Social Psychology*, 46(6):1208–1221.

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

Chenhao Tan, 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, page 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Gerald Tesauro. 1995. *TD-Gammon: A Self-Teaching Backgammon Program*, pages 267–285. Springer US, Boston, MA.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

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

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

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *COLING*, pages 1680–1691.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017a. Argumentation quality assessment:

Theory vs. practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 250–255, Vancouver, Canada. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017b. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia. Association for Computational Linguistics.

Marilyn Walker, Jean Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 812–817, Istanbul, Turkey. European Language Resources Association (ELRA).

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lan-*

guage Technologies, pages 47–57, San Diego, California. Association for Computational Linguistics.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Christo Wilson, Bryce Boe, Alessandra Sala, Krishna PN Puttaswamy, and Ben Y Zhao. 2009. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. Acm.

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.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489, San Diego, California. 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 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 136–141, San Diego, California. Association for Computational Linguistics.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 207–212, Berlin, Germany. Association for Computational Linguistics.

Ingrid Zukerman, Richard McConachy, and Sarah George. 2000. Using argumentation strategies in automated argument generation. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 55–62, Mitzpe Ramon, Israel. Association for Computational Linguistics.