COMPUTATIONAL ANALYSIS OF ARGUMENTS AND PERSUASIVE STRATEGIES IN POLITICAL DISCOURSE

by

Nona Naderi
A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto

Abstract

Computational analysis of arguments
and persuasive strategies in political discourse

Nona Naderi

Doctor of Philosophy

Graduate Department of Computer Science

University of Toronto

2019

Various persuasive strategies are employed in advancing argumentation. This dissertation presents the first computational work in analyzing persuasive strategies in monological and dialogical argumentation in natural language. I begin with reputation defence strategies and show to what extent human annotators agree on these strategies. I present the first manually annotated corpus of parliamentary debates annotated with the most agreed upon face-saving strategies and show that linguistic features automatically extracted from the text of debates can differentiate between these strategies. Having shown the effectiveness of discourse parsing features in the classification of reputation defence strategies, I hypothesize that by directly using the effective features for discourse parsing, the classification results can be improved. My experiments validate this hypothesis and show that the developed methods can automatically label speeches with these strategies. I then explore whether we can automatically predict the language of face-saving in speeches and show that by leveraging the contextual information of the speeches, we can reliably distinguish between reputation defence from non-defence. I further investigate whether we can automatically classify statements in face-threatening and face-saving speeches based on truthfulness using the effective linguistic features introduced in the prior literature and show that while some of these features help identify the expression of dodge, they are not very effective in identifying the truthfulness of the statements.

I further operationalize framing analysis as a classification task and show that neural language models can capture the abstract representations of frames more effectively. My experiments also show that frames are transferable across genres.

Finally, in collaboration with several researchers, we examine to what extent expert and lay annotators can evaluate argumentation aspects, and show that the agreement of both groups is limited.

Acknowledgements

First of all, I would like to thank my advisor, Professor Graeme Hirst, for giving me the opportunity to be his student, for his enormous patience, unfailing support, and generous guidance to my research, career, writings, presentations, applications, and all things, and for all the opportunities that he provided me over the years. He is an excellent supervisor and an influential mentor. I am inspired by his academic integrity, energy, and dedication to his work every day. Thank you, Graeme, for all the time and energy you invested in making me a better researcher and person.

I would like to thank my other advisory committee members, Professor Suzanne Stevenson and Professor Gerald Penn for their insightful questions and comments, encouragements, and valuable advice on my research and career. Suzanne always reminded me of the ”big picture” perspective and provided useful feedback on my presentations. Gerald helped me to express my contributions clearly and supported me in my applications. I would like to thank my external examiner, Professor Vincent Ng, for reading through my thesis and for his insightful comments, questions, and helpful feedback.

I would also like to thank my internal committee members, Professor Yang Xu and Professor Sheila McIlraith for their helpful comments, discussions, and encouragements.

I am grateful to my collaborators in LiPaD and argumentation quality projects: Professor Christopher Cochrane, Kaspar Beelan, Tanya Whyte, Ludovic Rheault, Henning Wachsmuth, Yufang Hou, Ivan Habernal, Yonatan Bilu, Vinodkumar Prabhakaran, Professor Iryna Gurevych, Professor Benno Stein. It’s been a pleasure to learn from them, and I hope we can continue our collaborations into the future. I admire Henning’s approach to collaboration and driving the research forward.

I am grateful to Tong Wang and Jamie Ryan Kiros for providing me with their representation models that I used in my analysis of issue-specific framing, for patiently answering my questions, and for inspiring discussions.

Thank you to Patricia Araujo Thaine, Simon Emond, Krish Perumal, Sara Scharf for annotating my data. I'd like to thank Professor Frank Rudzicz for reading parts of my thesis and providing useful feedback. I would also like to thank Afsaneh Fazly for reaching out and providing helpful advice.

I had the pleasure to work with two of our undergraduate students, Tim Alberdingk Thijm and Michael Kimmins. Thanks to them for working hard during summer.

I am grateful to have been funded by the Natural Sciences and Engineering Research Council of Canada, by the Ontario Graduate Scholarship, and by University of Toronto.

I would also like to thank the staff of the Computer Science department for their administrative support. In particular, Relu Patrascu, Luna Boodram, Marina Haloulos, Lynda Barnes, Celeste Francis Esteves, Neil Reilly, Simon Weber-Brown, Chris Siebenmann, Joseph Raghubar, Sara Burns, Lisa DeCaro, Sarah Lavoie, Vinita Krishnan, Margaret Meaney, Jankie Ramsook.

Thanks to all past and present members of the Computational Linguistics group at UofT for their valuable comments and suggestions on my work and talks.

I have been fortunate to be surrounded by brilliant and smart friends and colleagues at UofT. I am thankful to Abdel-rahman Mohamed, Aditya Bhargava, Aida Nematzadeh, Amin Tootoonchian, Arvid Frydenlund, Alexander Schwing, Bai Li, Barend Beekhuizen, Chloe Pou-Prom, Ella Rabinovich, Eric Corlett, Fartash Faghri, Fernando Flores-Mangas, Gagandeep Singh, Gellert Mattyus, Jackie C. K. Cheung, Jake Snell, Julian Brooke, Katie Fraser, Krish Perumal, Krishnapriya Vishnubhotla, Libby Barak, Mohammad Norouzi, Muuo Wambua, Patricia Araujo Thaine, Sean Robertson, Serena Jeblee, Siavash Kazemian, Tong Wang, Toryn Qwyllyn Klassen, Vanessa Queiroz Marinho, Vanessa Wei Feng, Varada Kolhatkar, Yevgen Matusevych, Zahra Shekarchi, for their support, for their positive energy, for stimulating research environment, for interesting conversations, cookie breaks, dinners, parties, trips, board games, hikes, and tennis and squash games. I owe special thanks to Aida Nematzadeh, Amin Tootoonchian, Patricia Araujo Thaine, and Simon Emond for their continuous support, care, and kindness. They have been reliable and wonderful friends since I moved to Toronto.

I would also like to thank my MSc advisor, Professor René Witte, for introducing me to Natural Language Processing.

I would like to thank my wonderful friends who supported me morally from distance, Louisa Harutyunyan and Mohsen Eftekhari.

I am forever grateful to my family for their endless love and unconditional support in all my pursuits, hopes, dreams, successes and failures, and for never hesitating to help me succeed. Thanks to my parents, Homa and Hossein, for all their sacrifices and efforts that allowed me to follow my own path. Thanks to my sister, Nassim, and my brother, Koorosh, for their constant encouragement and the joy they bring to my life. Thanks to Nassim for being so kind, caring, and generous. I dedicate this thesis to my family.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Argumentation

Arguments constitute a great deal of quotidian discourse. They are used to advance a point of view or to convince others that one's standpoint is correct. Various theories have been proposed for modeling argumentation in the literature. Bentahar et al. (2010) categorize argumentation models into (1) *monological models* (i.e., they address the internal structure of the argument, see Section 1.5.1), (2) *Dialogical models* (i.e., they address the interactions between arguments of two or more parties, see Section 1.5.3), and (3) *Rhetorical models* (i.e., they focus on the persuasive and rhetorical aspects of arguments and how arguments are evaluated by the audience). The persuasive effects of arguments are believed to be achieved through various means, such as ethos (arguer's credibility), pathos (successful emotional appeal to the target audience), and logos (reasons) as in the Aristotelian view of rhetoric. Pragma-dialecticians refer to these means as "strategic manoeuvring" (van Eemeren, 2010), which manifests itself in topical (topoi or argumentation scheme) selectivity, audience adaptation, and presentational devices. Presentational devices include (1) *"syntactic" devices*, e.g., paratactic (short and simple sentence constructs) and hypotactic constructions (more-complex sentence constructs with syntactic subordination), (2) *"semantic" devices,* e.g., metonymy (a figure of speech where

a thing or concept is referred to by the name of something closely associated with that thing or concept), and (3) *"pragmatic" devices,* e.g., rhetorical questions.

Understanding the interaction between these various means of persuasion and understanding their persuasive effects helps us design better decision-making, educational systems (van Eemeren, 2010; Macagno and Konstantinidou, 2013; Modgil et al., 2013), and debating technologies (Levy et al., 2014; Rinott et al., 2015a). This understanding can be achieved through advances in machine learning and computational linguistics methods. Computational analysis of arguments in natural language texts, sometimes called argumentation mining, is an emerging area of research in natural language processing and computational linguistics. Computational studies of arguments have so far focused on the automatic detection of arguments, i.e., determining whether a piece of text is argumentative or not, (Moens et al., 2007; Florou et al., 2013) and their components (see Section 1.5.1), i.e., segmenting a piece of text into argument units and classifying the type of each unit (Mochales and Moens, 2008; Levy et al., 2014; Goudas et al., 2014; Stab and Gurevych, 2014b; Rinott et al., 2015b; Persing and Ng, 2016; Ajjour et al., 2017; Stab and Gurevych, 2017). Computational methods have also been proposed to automatically identify argument types (see Section 1.5.4) (Feng and Hirst, 2011) and relations between argumentative components (Stab and Gurevych, 2014a; Nguyen and Litman, 2016; Stab and Gurevych, 2017); both rely on the previous step of identifying the arguments and their structures (see Section 1.5.2). More recently, researchers have devised methods to evaluate the persuasive effect of arguments (Wei et al., 2016; Habernal and Gurevych, 2016a; Ghosh et al., 2016). However, computational approaches have rarely addressed the analysis of persuasive strategies in argumentation.

Most approaches in analyzing arguments in text rely on supervised machine learning and manually annotated corpora. These approaches achieve promising results; however, their performance is limited due to the challenges involved in argument analysis in general. One major challenge originates from argument components (premises and claims) that are ambiguous or left implicit due to common knowledge, or that are presented far apart from each other in

text. Additionally, various strategies for presenting arguments, e.g., adaptation to the audience, can affect how an argument is perceived (easily recognized by some, but not the others). These challenges make it difficult even for humans to analyze arguments and they further impede corpus development. Current computational models consequently cannot be easily applied to new data or to another genre of discourse. Arguments are primarily generated for persuasion, therefore, in order to better understand and evaluate them, we need to analyze persuasive strategies.

In this dissertation, I examine *whether we can analyze persuasive strategies in natural language using computational approaches based on the methods of classification, and determine what linguistic features can help in automatically identifying these strategies.*

This dissertation brings together methods from communication studies, linguistics, psycholinguistics and computer science to achieve a better understanding of how individuals use persuasive strategies in their arguments to reach their goals.

Political discourse is primarily argumentative and arguments are particularly used to establish political beliefs and appeal to values, or plan policies; hence, it constitutes a rich corpus for the analysis of persuasive strategies.

1.2 Persuasive strategies in argumentation

Argumentation, as Tindale (2006) stated, is always generated with some expected audience (one or more groups) that is addressed through the argumentation. Based on this expected audience demand, the situation, and the arguer, some effective devices are chosen to maximize the likelihood of a successful outcome. For example, one may use metaphors (van Eemeren and Eemeren, 2009) or irony (van Eemeren, 2010) to express his argument. Since persuasive means are developed in communicative activities, such as negotiation and deliberation depending on the institutional needs, it is important to take these institutional needs or goals into account when examining argumentative exchanges (Mohammed, 2008; van Eemeren, 2010).

Previous computational work on persuasive effects of arguments has focused on arguments in isolation and analyzed only a few means of persuasion. However, some arguments are generated in the context of other arguments, so the evaluation of these arguments needs to consider the interactions between them. In this dissertation, I develop methods to automatically classify the persuasive and rhetorical strategies used in monological and dialogical argumentation, and show their importance in argumentation understanding and assessment, and their impact on the target audience.

In the following subsections, we introduce the various persuasive strategies that we examine in this thesis.

1.2.1 Face-saving

Ethos i.e., one's credibility, has been considered as one of the important components of persuasion in Aristotle's rhetoric (Aristotle, 2007). When in danger of losing credibility, one may prepare apologia–that is, a self-defence speech– in response to the criticism or attack. According to Downey (1993), apologia has taken various functions and styles over time; for example, early contemporary apologia, similar to classical apologia, used causal reasoning and detailed evidence; however, after 1960, apologia has been altering into "misleading narratives and dishonest apologies", replete with discrepancies. Similar to Downey's study, most previous work on the analysis of face-saving and persuasive reputation defence language has focused on a few case studies (Brinson and Benoit, 1999; Benoit and Henson, 2009; Zhang and Benoit, 2009; Harlow et al., 2011). These cases focused on manually analyzing reputation defence strategies and only considered the text of the apologia. However, manually analyzing each of these cases is laborious and time-consuming. Furthermore, as Ryan (1982) stated, a complete understanding of apologia and accusatory speech can be achieved through treating them both.

Research questions:

- How and to what extents do human annotators agree on reputation defence strategies in

speeches? Section 2.1.3 presents the first manually annotated corpus of parliamentary debates annotated with the most agreed upon *reputation defence strategies* drawn from communication studies.

- Can we automatically identify reputation defence strategies? Section 2.1.4 proposes a computational model of face-saving strategies in dialogical texts. I propose a set of features that can distinguish between various reputation defence strategies being used in different contexts and for various issues. The results of the classification using these features are represented in Sections 2.1.5 and 2.1.6.

- Can we improve the results of the classification? Can we automatically generate labeled data with reputation defence strategies? In order to address the challenges associated with manually annotating the data with reputation defence strategies, Section 2.2 proposes two approaches to automatically assign labels to speeches with reputation defence strategies and improve the classification results.

- Can we automatically identify a face-saving speech? Does the contextual information help in the classification of the language of face-saving? To answer this question, I propose a dataset of reputation defence, in which the annotations are based on the structure of parliamentary debates. I further explore various models to distinguish reputation defence from non–defence and show that leveraging contextual information in this classification task results in better performance. Section 2.3 presents the proposed dataset and the approach to classify reputation defence from non–defence.

- Can we automatically detect true, false, stretch, and dodge statements in face-threatening and face-saving speeches? Section 2.4 presents our analysis using the Toronto Star dataset that is recently developed by reporters. This section studies whether the effective features in the prior research can help us classify truths, falsehoods, dodges, and stretches in the Canadian debates and shows that while some of these features help us identify dodge

> statements with an F_1 measure as high as 82.57%, they are not very effective in identifying false and stretch statements.

1.2.2 Framing

Theoretical perspectives on framing and frames are diverse, but these theories converge in their conceptualization of framing as a communication process to present an object or an issue. One of the widely accepted and most influential definitions was proposed by Entman:

> Framing involves selection and salience. To frame is to select some aspects of a perceived reality and make them more salient in a communicating text, in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation for the item described. (Entman, 1993, p. 53)

For Entman, frames are schemata with which, "problems are defined, causes are diagnosed, moral judgments are made," and "remedies are suggested." He also notes that some of these functions might not exist in the text, suggesting that frames are independent of the stance that is taken in the text. By contrast, Chong and Druckman (2007) believe that frames define attitudes and thus they are inherently associated with stances of advocacy or opposition. Wohlrapp (2014) views frames as means of success or failure of argumentation.

Yet others instead interpret frames as topics, such as CRIME AND PUNISHMENT or HEALTH AND SAFETY (Card et al., 2015). However, these topics are just the categories of the subject matter in news articles. Obviously, multiple frames can occur in any of the categories, explaining why some theorists such as Boydstun (2014) claim that framing should be perceived as non-issue-specific and be analyzed with a fixed set of framing dimensions, rather than being associated with a specific issue (Entman, 1993; Chong and Druckman, 2007). Nonetheless, we agree with de Vreese (2005) that frames can be classified as either generic or issue-specific; for example, ECONOMIC BENEFITS can be used as a generic frame for various issues. However, a

frame such as MARRIAGE IS ABOUT MORE THAN PROCREATION is specific to *gay marriage* issue.

Although framing manifests through language, the existing definitions pay little attention to its linguistic aspect and provide no guidelines on what constitutes a frame or how to find them. In Entman's view, frames are located in several places, including the "communicator", the "text", the "receiver", and the "culture" (in line with philosophical views on meaning as Hirst (2007) explains), and can be manifested by the presence or absence of "certain keywords, stock phrases, stereotyped images, sources of information, and sentences that provide thematically reinforcing clusters of facts or judgments." (Entman, 1993)

Along these lines, Gamson and Modigliani (1989) suggest that frames can be identified using "particular signature elements for a given frame," such as "metaphors, catch-phrases, or exemplars, depictions, and visual images." In contrast, Chong and Druckman (2007) suggest that in order to find frames, we need to find different attitudes towards a certain issue, because frames underlie these attitudes. Semetko and Valkenburg (2000) believes that there are two approaches to analyze frames: "inductive" and "deductive". In inductive approach, content or a news story is analyzed to extract as many frames as possible. This approach is difficult and cannot be easily replicated. In deductive approach, the content is examined to find occurrences of a set of pre-existing frames; however, some frames can be left unidentified (Semetko and Valkenburg, 2000).

Here, we adapt the definition of *frame* as *a device to highlight an aspect of a given issue, which might or might not also specify the presenter's stance.* We take a deductive approach as it can be replicated.

Consider the two statements from marriage debate in the Canadian parliamentary proceedings:

Example 1.2.1 *The right to equal marriage now woven into the very fabric of our nation's identity is not simply a concept to be negotiated between those fortunate enough to sit in this*

chamber.[1]

Example 1.2.2 *Our biggest challenge as human beings is to get along, to learn about each other, to accept differences, to give the same chance to others to live their lives as we would like them to give to us and to allow others to share fully and completely in the world.*[2]

The two examples refer to the same frame IT IS DISCRIMINATORY TO REFUSE GAY COUPLES THE RIGHT TO MARRY or in its generic form FAIRNESS that was put forward by liberals. In order to be able to identify the frame in these statements, we need to deal with the complexity of compositional semantics.

Research questions:

- Is it possible to automatically identify frames in sentences and paragraphs? Section 3.1 explores the use of neural language models in identifying frames at the sentence level in news articles and show that they can capture the abstract representations of frames more effectively than the classifiers trained on topics. Section 3.2 presents an approach to automatically identify frames in parliamentary debates at the sentence and paragraph levels.

- Is it possible to use an already existing set of frames for a specific genre to identify frames in another genre? Section 3.2 presents the first steps in searching for an answer by using the frames annotated for online forums to identify frames in political parliamentary debates. Our model is based on the computed similarity between frames and speeches using a vector-based distance measure.

[1] Mario Silva, 2006-12-6
[2] Ken Dryden, 2006-12-6

1.3 Assessing argumentation quality

As mentioned earlier, there are various theories for modeling argumentation. There have been a few attempts to bring at least some of these models together, e.g., that of Grasso (2002); however, a clear analysis of how these models can contribute to argument understanding and evaluation is missing. In this work, with collaboration with researchers from multiple universities, we perform a literature review of all the perspectives on what contributes to the evaluation of arguments and create a taxonomy of quality dimensions. We further perform an annotation study to examine how subjective and complex each dimension is. My contributions to these studies were primarily (i) conducting the annotation study for the dimensions and (ii) conducting the crowd-sourcing annotation task. In addition, I also (iii) assisted in literature review and in the development of several of the dimensions, (iv) contributed to annotating the arguments, (v) participated in the discussion of the studies, and (vi) assisted in the writing of the papers.

Research questions:

- How can we automatically assess the quality of arguments? As a first step towards automating the assessment of arguments, we survey various existing theories and propose a set of dimensions that can be used to assess the quality of arguments. Appendix A presents the survey and describes the proposed dimensions.

- Can humans evaluate arguments based on this set of dimensions? We created guidelines based on the proposed dimensions and created a benchmark corpus of arguments manually annotated by expert annotators with these dimensions and show which aspects of argumentation are difficult to assess (Appendix A). We further performed a crowdsourcing annotation study to examine how lay annotators' annotations compare to those of the expert annotators (Appendix B).

- How do these proposed dimensions for the assessment of arguments compare with the

proposed approaches in the literature? Section B provides a comparison of our approach with Habernal and Gurevych (2016b)'s approach.

1.4 Contributions

1.4.1 Overview

The contributions of this thesis can be divided into (1) reputation defence analysis, i.e., approaches to identify reputation defence language and reputation defence strategies, (2) framing analysis, i.e., approaches for identifying frames, and (3) assessing argumentation quality, i.e., contributing to the development of a set of dimensions for assessing the quality of argumentation. The following lists provide an overview of these contributions:

Reputation defence analysis:

- We introduce an annotation scheme for modeling reputation defence strategies derived from communication theories.
- We create the first corpus for studying reputation defence strategies. The corpus consists of questions and answers from parliamentary debates. We propose a set of features that can distinguish these strategies.
- We propose two approaches to automatically label data with reputation defence strategies.
- We propose approaches to identify the language of reputation defence in parliamentary debates and show that this language can be reliably identified across different parliaments.
- We provide a comparison of truthfulness annotations in the Canadian parliament and the U.S. fact-checking corpus.

Framing Analysis:

- We provide a systematic summary of the existing approaches in framing analysis and categorize these approaches using a typology of framing by de Vreese (2005).

- We create a corpus of parliamentary debates annotated with a set of predefined frames. We show that frames used in online forums are transferable and can be used to analyze frames in parliamentary proceedings.
- We propose a supervised machine learning approach to classify frames at the sentence and paragraph levels.

Argument quality assessment in collaboration with other researchers:

- We provide a systematic summary of the existing work for assessing the quality of the arguments.
- We propose a set of dimensions for the assessment of arguments.
- We created a corpus of arguments annotated with these dimensions.
- We compare this corpus with the existing corpora.

1.4.2 Publications

Parts of this thesis have been previously published in international peer-reviewed conference and workshop proceedings. We list all the publications below and indicate the chapters and sections of this thesis that were built upon them:

- Reputation defence analysis:
  - Nona Naderi and Graeme Hirst. Recognizing Reputation Defence Strategies in Critical Political Exchanges. *The Eleventh International Conference on Recent Advances in Natural Language Processing (RANLP 2017),* pages 527–535, September 2017. Section 2.1.
  - Nona Naderi and Graeme Hirst. Reputation protection and repair strategies in argumentative political discourse (extended abstract accepted for an oral presentation). European Conference on Argumentation (ECA 17), June 2017. Section 2.1.

  - Nona Naderi and Graeme Hirst. 2018. Automatically labeled data generation for classification of reputation defence strategies. *The Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* Miyazaki, Japan. European Language Resources Association (ELRA). Section 2.2.
  - Nona Naderi and Graeme Hirst. 2018. Using context to identify the language of face-saving. *The 5th Workshop on Argument Mining.* Section 2.3.
  - Nona Naderi and Graeme Hirst. Automated fact-checking of claims in argumentative parliamentary debates. *The First Workshop on Fact Extraction and Verification.* Section 2.4.

- Framing analysis:
  - Nona Naderi and Graeme Hirst. Classifying frames at the sentence level in news articles. *The Eleventh International Conference on Recent Advances in Natural Language Processing (RANLP 2017),* pages 536–542, September 2017. Section 3.1.
  - Nona Naderi and Graeme Hirst. Argumentation mining in parliamentary discourse. 2016. In Baldoni M. et al. (eds), *Principles and Practice of Multi-Agent Systems,* Springer International Publishing, pages 16–25. Section 3.2.
  - Nona Naderi. Argumentation mining in parliamentary discourse. In *the 11th International Conference of the Ontario Society for the Study of Argumentation,* May, 2016, Windsor. Section 1.2.2 and Section 3.2.

- Assessing argumentation quality:
  - Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational Argumentation Quality Assessment in Natural Language. *the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 17),* pages 176–187, April 2017. Appendix A.

– Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. Argumentation Quality Assessment: Theory vs. Practice. *the 55th Annual Meeting of the Association for Computational Linguistics (ACL 17),* pages 250–255, July 2017. Appendix B.

1.5 A quick introduction to argumentation

In this section, we introduce important terminologies and concepts in argumentation. For a more elaborate list, see Stede et al. (2018); Walton (2005).

1.5.1 Definition of an argument

Different models of "argument" have been proposed in the literature. For example, according to Besnard and Hunter (2008), "An argument is a set of assumptions (i.e., information from which conclusions can be drawn), together with a conclusion that can be obtained by one or more reasoning steps (i.e., steps of deduction). The assumptions used are called the support (or, equivalently, the premises) of the argument, and its conclusion (singled out from many possible ones) is called the claim (or, equivalently, the consequent or the conclusion) of the argument." In the same vein, Walton (1992) considers three components to be necessary for an argument: a set of **premises**, a **conclusion**, and an **inference** from premises to the conclusion. In this model, each argument can be **supported** or **attacked** by other arguments or **critical questions**. Toulmin (1958) introduced six roles for the elements of an argument in his model, namely **data**, a **claim**, **qualifiers**, **warrants**, **backing**, and conditions of **rebuttal**. A **claim** is what needs to be justified, and **data** or **facts** provide the basis of the claim. **Warrants** or rules (possibly implicit) connect data and the claim, and together with data, they provide support for the claim. **Qualifiers** indicate the degree of certainty and the strength of the justification, **backing** explains why warrants justify the claim and further supports warrants (e.g., based on previous cases), and **rebuttals** express the circumstances where warrants do not hold. Here, we

use the example presented by Toulmin to show each component of an argument.

Example 1.5.1 *Harry was born in Bermuda (data). A man born in Bermuda will generally be a British subject (warrant), so, presumably (qualifier), Harry is a British subject (claim), on account of the following statutes and other legal provisions (backing)—unless both his parents were aliens or he has become a naturalized American (rebuttal).*

However, real-world arguments do not usually appear in this format and typically lack some of these elements. Summaries of critiques of Toulmin's model can be found in a review by Peldszus and Stede (2013) and Gover (2018)'s work.

What is common to all these models is the concepts of a *claim* that is an assertion thought to be true and *premises* that provide reasons for the claim or conclusion.

Enthymeme

Arguments do not always include all the elements (premises and conclusion) that we mentioned in the previous section, and sometimes some of the elements are left implicit due to common knowledge or ambiguity of natural language. An argument with implicitly stated or unstated premises or conclusion is called an **enthymeme** or **enthymematic** argument. Consider the following example[3]:

Example 1.5.2 *Bearing in mind the fact that Axa Insurance has today announced that 210,000 small firms are operating without employer's liability, and therefore illegally, will the Minister ask the Department for Work and Pensions to carry out its review urgently, ensure that it covers issues such as competitive practices and consistency of cover, and ensure that firms are given proper notice when the insurance basis is to be changed?*

In the above example, one annotator[4] annotated the following propositions as implicit and unexpressed:

[3] argument_96 and _255, AraucariaDB.
[4] argument_96, AraucariaDB.

- Implicit conclusion: The Minister ought to ask the Department for Work and Pensions to carry out its review urgently, ensure that it covers issues such as competitive practices and consistency of cover, and ensure that firms are given proper notice when the insurance basis is to be changed.
- Implicit premise: Axa Insurance is a reliable source.
- Implicit premise: If reliable sources have announced that 210,000 small firms are operating without employer's liability, and therefore illegally, the relevant Minister ought to ask the Department for Work and Pensions to carry out a review of related issues such as competitive practies and consistency of cover urgently.

Another annotator[5] considered the following propositions to be implicit:

- Implicit conclusion: Minister should ask the Department for Work to carry out its reviews.
- Implicit premise: Reviews are a good way to compel firms to operate legally.

According to both anotators, the conclusion of the argument is not expressed explicitly and can be inferred from the question. They further recognized different implicit premises.

To correctly understand and interpret an incomplete argument, we need to **reconstruct** it by finding its missing parts.While determining whether an argumentative text expresses a complete argument or not has generally been conceived of as an easy task[6], reconstructing enthymemes and finding the exact and precise missing statements is not completely straightforward, and there are potentially an infinite number of candidates.

1.5.2 The structure of arguments

The structure[7] of an argument indicates how premises are used and related to draw the conclusion in a single argument. Determining the structure of an argument is necessary to understand how

[5] argument_255, AraucariaDB.

[6] Some scholars believe that all arguments are enthymematic Gover (2018).

[7] This term is very ambiguous; in this dissertation, I use it when I talk about how components of an argument are related and will never use any other meanings.

a proponent defends his or her point of view, and to further evaluate the logic of the presented argument. The structure of an argument depends on the model of an argument and its elements that are used to analyze it (see Section 1.5.1). In a simplest case, a **single** premise may support a conclusion. Example 1.5.3;[8] illustrates a claim (conclusion) that is specified in bold face and it is supported by one premise.

Example 1.5.3 ***This bill will probably be challenged in court*** *because it threatens the fundamental right of citizenship.*

However, often multiple premises are presented in an argument. In case all premises are combined together to support the conclusion, the resulting argument structure is referred to as **linked**. Consider the following example (Rowe and Reed, 2008):

Example 1.5.4 *The first year physics course covers Newton's laws of motion. Jon got 90% in the first year physics course. Jon understands Newton's laws of motion.*

In the above argument, the conclusion *Jon understands Newton's laws of motion* can be derived by both premises *The first year physics course covers Newton's laws of motion* and *Jon got 90% in the first year physics course*, but each premise individually is not sufficient to draw the conclusion.

In a **convergent** structure, each premise alone supports the conclusion. In the following example, each premise by itself (*It is friendly* or *It is relatively quiet*) is sufficient to draw the conclusion *A cat makes a good pet.*

Example 1.5.5 *A cat makes a good pet. It is friendly. It is relatively quiet.*

According to Beardsley (1950), "a **serial** argument contains a statement that is both a conclusion and a reason for a further conclusion". Consider the following example:

[8] Philip Toone, "Strengthening Canadian Citizenship Act." *House of Commons Debates (Hansard) of Canada, 41st Parliament,* June 12, 2014.

Example 1.5.6 *He has been playing tennis consistently since twenty years ago. Therefore, he has the experience. He will win the game.*

In the above example, the first proposition *He has been playing tennis consistently since twenty years ago* provides support for the second proposition *Therefore, he has the experience.* Likewise, *Therefore, he has the experience* provides support for the main claim of the argument that is *He will win the game.* In other words, the second proposition has two roles of a premise and a conclusion in this argument; however, this proposition is not the main claim of the argument.

In an argument model in which two or more conclusions are allowed,[9] if one premise supports multiple conclusions, the resulting structure is called **divergent** (see Example 1.5.7); however, since many scholars believe that arguments include only one conclusion, divergent structure has been mainly ignored. In the following example, two claims *Karl must be conservative* and *Mary must be liberal* are supported by the premise *Mary argues with Karl regarding his opposition to abortion.*

Example 1.5.7 *Mary argues with Karl regarding his opposition to abortion. Karl must be conservative and Mary must be liberal.*

It is important to note that some researchers consider the above structure as two separate arguments by replicating the premise for the second claim (Reed and Rowe, 2004).

All the above structures are illustrated schematically in Figure 1.1. In these diagrams, a premise is shown as an encircled P, a claim as an encircled C, and support is represented by an arrow.

Van Eemeren and Grootendorst (2004) also proposed a similar model for the structures of arguments and referred to them as single, multiple (convergent), subordinatively compound (serial), coordinatively compound (linked); however, they believed that the same structures can be applied to the relations among arguments in a dialogue as well. According to van Eemeren

[9]For example, Freeman (1991) defines premises and conclusions as possible argumentative roles in an argument.

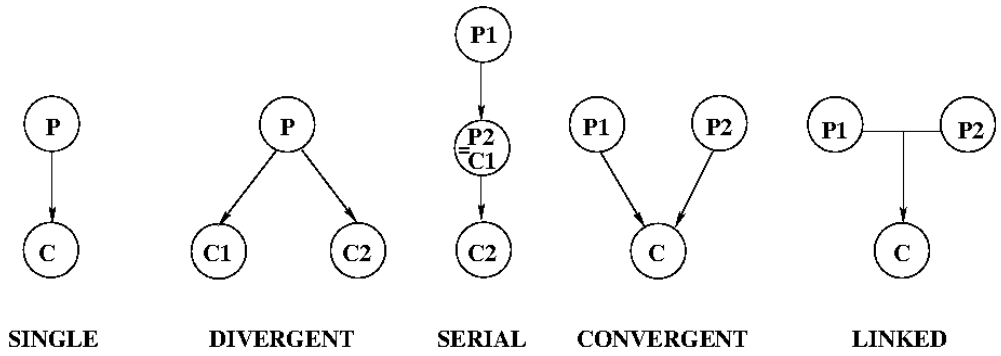

Figure 1.1: The structure of arguments

and Grootendorst's theory, there is a central standpoint (claim) in a dialogue between proponent and opponent, and each agent (arguer) presents his or her arguments to defend or reject that standpoint. Therefore, each agent can apply the same structures to relate the arguments.

1.5.3 Abstract argumentation frameworks

A formal model of abstract argumentation can be considered as a set of arguments with binary **attack** or **defeat** relations. An influential abstract representation for arguments was proposed by Dung (1995). He defined an abstract argumentation framework as a pair of arguments and attack relations (AF = $\langle$ AR, *attacks* $\rangle$, AR is a set of arguments, and *attacks* is a binary relation on AR, i.e., *attacks* $\subseteq$ AR $\times$ AR). As an example, Arg_1attacks Arg_2 if *attacks*(Arg_1, Arg_2) holds. The framework can be presented as a directed graph whose vertices are the arguments in the set and edges are attack relations. Dung's proposed framework is considered abstract as he provided no information regarding the nature of an argument and its elements, nor did he mention how an argument can attack another, e.g., by attacking premises, the conclusion, or the argument itself. Pollock (1987, 1991) focused on the relations between arguments in his model. According to him, arguments can defeat each other in two different ways, namely rebutting defeat and undercutting defeat. An argument is called a **rebutting defeater** if it defeats another argument's conclusion, and an **undercutting defeater** if it presents contradictions to another argument's premises. In a set, where argument B defeats A, and C defeats B, then A is justified and we say that argument C **reinstates** argument A. In addition to attacking or defeating an

argument through attacking or falsifying its premises and/or conclusion, one can also attack or defeat an argument by attacking the argument itself as a whole. This can be done through posing appropriate **critical questions**, which will be discussed in the following section.

1.5.4 Argumentation schemes

Reasoning structures and types of arguments that occur in everyday discourse are called argumentation schemes. While many argumentation schemes have been proposed, Walton et al. (2008) provide the most comprehensive collection, presenting 65 schemes. Each scheme represents the underlying logic of a specific argument and the critical questions that can challenge that argument. As an example, the scheme of *pragmatic inconsistency* expresses an opposition between the facts and stated commitments and is captured as follows:

Example 1.5.8 *Premise: a advocates α, which has proposition A as its conclusion.*

Premise: a has carried out an action, or set of actions, that imply that a is personally committed to $\neg A$ (the opposite, or negation of A).

Conclusion: Therefore, a's argument α should not be accepted.

Critical Questions:

CQ_1: Did a advocate α in a strong way indicating her personal commitment to A?

CQ_2: In what words was the action described, and does that description imply that a is personally committed to the opposite of A?

CQ_3: Why is the pragmatic inconsistency indicated by satisfactory answers to CQ_1 and CQ_2 a relevant reason for not accepting argument α?

Chapter 2

Reputation defence analysis

2.1 Reputation defence strategies

Good reputation is one of the most valuable assets one can possess. Criticisms and persuasive attacks pose threats to reputation and they are very common in all social interactions. Every day, we hear and read about allegations regarding organizations (e.g., companies and governments) or individuals (e.g., medical practitioners and politicians), and in response to these allegations various argumentation tactics and persuasive strategies are used to minimize the damage.

A recent prevailing example of reputation threat and defence is various sexual assault allegations and the use of strategies, such as *denial* and *mortification,* i.e., the admission of guilt and apologizing and asking for forgiveness, in response to these allegations.

Here is an example of *denial* in the case of a bribery allegation against Airbus Helicopters in a Greek NH-90 helicopter deal. In a statement, it said:

Example 2.1.1 *These allegations are groundless and damage the reputation of Airbus Helicopters.*[1]

Another example of a reputation defence strategy is the expression of *mortification* in a statement that was issued by the U.S. Secretary of Health regarding the expense of his travel on

[1] *Airbus Helicopters rejects bribery allegations in Greek NH-90 deal*, Reuters, 2015-03-23

private planes:

Example 2.1.2 *I regret the concerns this has raised regarding the use of taxpayer dollars. All of my political career I've fought for the taxpayers. It is clear to me that in this case, I was not sensitive enough to my concern for the taxpayer.*[2]

How effectively an organization or an individual addresses and responds to such crises can have critical consequences.

Maintaining good reputation is especially important in political rhetoric, and is considered as one of its primary goals. When faced with criticism, politicians use various strategies to react to it and defend themselves to others—both to their critic and to their audience. These strategies are a component of political argumentation.

Consider the question-and-answer sessions in Westminister-style parliamentary debates, where the government of the day is held accountable by the opposition. Opposition members ask confrontational questions, and the government ministers respond. In the face of criticism, they may use various reputation defence strategies to try to maintain a positive image. In these question-and-answer sessions, government backbenchers also ask the government ministers questions; however, these questions are generally friendly and promotional questions and the answers given to these questions try to promote government's plans and considered ordinary or reputation-building or reputation-enhancing pairs. This contrast between the questions asked by opposition members and government backbenchers is supported by qualitative studies such as those of de Ayala (2001), Ilie (2006), and Bates et al. (2012), and allows us to study the characteristics of the language of reputation defence.

The motivation behind the computational analysis of face-saving language is its application in analyzing the effectiveness of argumentation. As we mentioned earlier, ethos or one's credibility has been considered as one of the important means of persuasion. Given two pieces of text and their respective face-saving strategy, e.g., denial or concession, one can compare the impact of the argument on the audience. For example, concessions have been perceived more

[2]Health secretary Tom Price apologizes for taking private flights for work, *The Guardian,* 2017-09-28

effective than other reputation defence strategies (Benoit and Drew, 1997). Furthermore, one can examine whether people can detect dodge statements or whether political attitude can have any impact on the detection of dodge statements (Clementson, 2018). Another motivation is the use of face-saving strategies in consistency checking of arguments. Furthermore, computational analysis of face-saving language has potential applications in communication studies (Leichty and Applegate, 1991; Coombs, 1998, 1995), psychology (Juvonen, 2000), language learning studies (Trosborg, 1987), and decision making (Milne and Patten, 2002).

This chapter synthesizes and revises the work originally published Naderi and Hirst (2017a, 2018a,c,b). In the first section of this chapter, we first study whether and to what extent human annotators can agree on reputation defence strategies in speech. Then, we study whether we can automatically classify reputation defence strategies and what linguistic features help us identify them. We examine and evaluate features that have been shown to be effective for the related tasks. In the following section, We examine whether we can improve the classification results. We explore two approaches to automatically label data with reputation defence strategies. In the following section, we examine whether we can predict the language of reputation defence. In the last section, we study whether we can classify true, false, dodge, and stretch statements in the Canadian parliament.

2.1.1 Classification of reputation defence strategies

No annotated data is available for this task, so we examine whether and how reputation defence strategies are used in parliamentary debates to respond to the opposition, and create a new corpus of Canadian parliamentary debates annotated with reputation defence strategies. We focus on the most agreed-upon strategies, namely *denial, excuse, justification,* and *concession* (Benoit, 1995). For example, politicians may deny having caused a bad situation (denial) or try to evade responsibility (excuse), or promise to fix the situation (concession). Table 2.1(a) presents an example from the Canadian parliament, where the government minister makes an excuse for a situation, and Table 2.1(b) presents an example of a concession.

Excuse	Concession
Q. Mr. Speaker, contrary to the Conservatives' claims, we are still short 30,000 jobs to get back to the level we were at before the crisis. For example, the Quebec forestry industry, which has lost 18,000 jobs since 2005, is struggling to get out of this difficult situation. Will the government understand that the crisis is far from over in the forestry industry and that it needs a comprehensive policy to support and modernize the industry, as was the case with the auto industry in Ontario? **A.** Mr. Speaker, all of the forestry experts in the country agree that it is a matter of markets. Unfortunately, the only ones who do not get it are the members opposite. They are playing politics with these people's jobs. The markets are difficult. Our workers are among the best in the world and we will continue to support them. Billions of dollars have been put into improving green practices through the community adjustment fund, and we will continue to support the forestry industry with research and development.	**Q.** Mr. Speaker, on December 9, just a few days from now, the École de médecine vétérinaire de Saint-Hyacinthe will have to report to the American Veterinary Association on the major investments required for its full accreditation to be restored. Does the Prime Minister grasp the urgency of the situation and does he not realize that the Government of Quebec has already put $41 million into the school and that it is now time he and his government did their share? It is urgent, a matter of days. **A.** Mr. Speaker, as has been said many times, this side and the government recognize the importance of the veterinary colleges, not only the one in Quebec but in the other three provinces in this country. We will do all we can to ensure that they maintain and continue their accreditation.

Table 2.1: Question and answer pairs from Canadian parliamentary proceedings annotated with reputation defence strategies: (a) 2011-02-01, Robert Bouchard (Q) and Denis Lebel (A); (b) 2002-12-03, Lyle Vanclief, (Q) and Yvan Loubier (A).

We then investigate what features are good predictors of the reputation defence strategies used in each case. The present work is a step towards a deeper understanding and evaluation of (political) arguments. Natural arguments are generally enthymematic, which means some of their elements are left implicit. Identifying these implicit argument elements is a very difficult task. Knowing what strategy is used in defence arguments may help in reconstruction of these missing elements. Furthermore, extracting defence strategies can facilitate identifying contradictory and inconsistent arguments.

2.1.2 Related work

Most previous studies on reputation defence strategies and their effectiveness are qualitative in nature (Coombs and Holladay, 2008; Sheldon and Sallot, 2008; Burns and Bruner, 2000;

Sheldon and Sallot, 2008; Lyon and Cameron, 2004).

While the task of automatically identifying reputation defence strategies has not been addressed previously, some researchers have focused on classifying the relations between argumentative components (Stab and Gurevych, 2014a; Nguyen and Litman, 2016). Others focused on classifying online discussions as agreement or disagreement with respect to a side of the debate on an issue (Abbott et al., 2011; Wang and Cardie, 2014; Rosenthal and McKeown, 2015). They employed various features, such as thread structure features, lexical (e.g., n-grams, number of words), and syntactic features (e.g., POS tags, dependency relations). Mukherjee and Liu (2013) proposed a semi-supervised generative model to extract agreement and disagreement expression types from discussion forums. Cabrio and Villata (2012) used a textual entailment approach to find pro and con arguments in a set of forum debates selected from Debatepedia. Current approaches, however, have mostly ignored the interaction between the parties involved in the argumentation process, where one party is critical of the other and the other party needs to overcome the doubts. Motivated by this previous work, in this section, we take a traditional feature-based model to study reputation defence strategies in parliamentary debates.

2.1.3 Data

For our analysis, we focus on pairs of questions and answers extracted from the Oral Question period of Canadian parliamentary proceedings. The purpose of questions asked in Oral Question period is to hold the government accountable for its actions[3]. While both government backbenchers and opposition members ask questions during this period, the questions asked by opposition members are more confrontational than the questions asked by the backbenchers. The questions asked by government backbenchers tend to be more clarification questions; therefore, we extracted the pairs where the questions were asked by opposition members.[4]

[3] `http://www.ourcommons.ca/About/Compendium/Questions/c_g_questions-e.htm`

[4] The tradition of question time for government accountability is practiced under different names in various countries; for example, in United Kingdom, it is known as *oral questions*, in Canada as *oral question period*, in Australia and New Zealand as *question time*, and in India as *question hour*.

Reputation defence strategies
Denial: 1. The government denies that the situation in question occurred. 2. The government denies causing the situation in question.
Excuse (evading responsibility): 1. The situation in question occurred in response to some other situations. 2. The situation in question occurred because of lack of information or control over important factors. 3. Some accidents caused the situation. 4. The motives or intentions of the government were good.
Justification (reducing offensiveness): 1. The government tries to increase positive feeling towards it (for example by mentioning positive actions the government performed in the past). 2. The government tries to convince the audience that the situation is not as bad they say. 3. The government tries to distinguish the situation in question from similar but less desirable situations. 4. The government tries to place the situation in a different or broader context. 5. The government attacks the opposition or questions their credibility. 6. The government offers compensation for the situation.
Concession (corrective actions): 1. The government promises to restore the situation to what it was before. 2. The government promises to make changes (for example to prevent the recurrence of the situation).
None of these strategies

Table 2.2: Conditions for each reputation defence strategy.

To study whether reputation defence strategies are used in the parliamentary debates, we first ran a pilot study and asked three expert annotators to annotate 100 random pairs of the extracted questions and answers with one of the reputation strategies or none of the strategies. We prepared detailed guidelines to describe the conditions that need to be satisfied for choosing each reputation defence strategy. Table 2.2 presents the conditions provided to the annotators (all are adapted from Benoit (1995)).

We further conducted a larger annotation study with 1500 random pairs of the extracted questions and answers on the crowd-sourcing platform CrowdFlower[5]. Contributors were shown a question and answer pair from the parliamentary debates on various issues, and were asked to choose which strategies (based on the conditions presented in Table 2.2) had been used

[5]https://www.crowdflower.com/

by the government in response to criticism. We asked for at least three annotations per pair from the English-speaking countries. To maintain the annotation quality, we allowed only the highest-quality contributors to participate, and also included some test pairs. On each page, each participant was presented with one test pair and three other pairs, and had to maintain 70% accuracy throughout the job. In total, we included 56 test questions for 1500 pairs. Each response was paid $0.04. Only 10% of the question and answer pairs were annotated with *none of the strategies* by the annotators, which shows that these strategies can represent the data reasonably well. Almost 70% of the pairs were agreed upon by two or more annotators, but in order to obtain a more reliable corpus, we accepted only the pairs for which at least three annotators agreed on a single answer, and discarded the pairs where fewer than three annotators agreed. For the expert annotations, three annotators achieved full agreement on a single answer for 32 pairs. In total, the reliable crowd and expert annotations resulted in a set of 493 pairs, of which 170 were annotated as *denial,* 36 as *excuse*, 173 as *justification,* 95 as *concession,* and 19 as *none of these strategies.* The average number of tokens in each pair is 171, with the longest pair being 356 words. These pairs of questions and answers are on different topics.

We further examined the discarded pairs of questions that were not agreed upon by at least three annotators to investigate the source of disagreements. Disagreements between the annotators were generally due to the use of multiple strategies or vague answers that do not contribute to the goal of the dialogue; they simply look like relevant answers, but they do not really address the questions. Table 2.3 shows an example of disagreement between three annotators.

2.1.4 Our approach

We consider the presented dataset to be a reasonable starting point for the automatic analysis of reputation defence strategies. We formulate the task as a classification task. Given a question and answer pair, we identify which of the four reputation defence strategies, *denial*, *justification*, *excuse*, and *concession* is used in the answer. In order to capture the characteristics of each

Q. Mr. Speaker, I would like the Minister of Public Safety and Emergency Preparedness to tell that to over 23,000 women who in 2003 were sexually assaulted or raped, and whose lives will never be the same again. Even more, I would like the minister to explain to these women why our prison libraries include pornographic magazines. Will the minister explain why our prison libraries feel it is necessary to provide pornographic material to violent sex offenders?	**A.** Mr. Speaker, as I just said, and maybe the hon. member did not hear me, I want to assure her that strict controls are in place to restrict access to any material that could be considered demeaning, could jeopardize the safety of any individual or the institution, is sexually violent or involves children or could be detrimental to the offender's treatment. We take the safety of our correctional institutions very seriously.

Table 2.3: Disagreement among three annotators, annotated variously as *denial, justification,* and *concession*; 2005-05-30, Lynne Yelich (Q) and Anne McLellan (A).

strategy, we explore two classes of features: features that are based solely on the answers, and features that describe the relation between the question and the answer.

Features from Answers

VerbNet Classes Certain verb classes can indicate defence strategies; for example, *assure* is often used in *justification* or *concession* strategies, e.g., *I want to assure the House that we are taking measures.* To this end, we use the VerbNet lexicon (Schuler, 2005), which groups verbs by their shared semantic meaning and syntactic behavior. Table 2.4 shows the verb classes that we use. We use the count of verb class occurrences as features.

Positive and Negative Sentiments and Emotions Motivated by the conditions for the *justification* strategy (Table 2.2), we examined the positive and negative sentiments and emotions expressed in the answers. Emotion words are extracted using Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010), a tool that counts occurrences of words by their psychological categories, and sentiments are extracted using OpinionFinder (Wilson et al., 2005).

Past and Future Focus Verb tense can reveal the difference between strategies; for example, in *denial*, the focus is more likely to be on the past, e.g., *as I said in French, I never gave advice*

Verb type	Examples
Concealment	conceal
Psych	amuse, admire
Desire	want, long
Judgment	judge, approve
Assessment	estimate
Searching	investigate
Social interaction	correspond, meet
Communication	inquire, advise
Existence	exist, survive
Aspectual	begin, continue
Allow	allow, permit
Admit	admit
Succeed	succeed

Table 2.4: VerbNet classes that we used for classifying reputation defence strategies.

about the privatization of the Toronto airport, whereas in *concession,* the focus tends to be on the future, e.g., *I promise the hon. member and all members of the special forces that I will work with them to ensure they are justly and properly treated*.

Negation *Denials* tend to be expressed using *never*, *not*, *no*, *nobody*, and *none*, e.g., *I never solicited funds*.

Insight and Achievement These categories are mostly associated with *justification* strategies, e.g., *I think when we can help farmers in Canada, it is our duty to do so*, and *We will continue to invest in this fashion. It is a proven success.* To compute these features, we use LIWC. We used 18 LIWC categories, presented in Table 2.5.

Features Describing Relations between a Question and Answer Pair

Discourse Relations Discourse relations have been shown to be effective in identifying support and attack relations in persuasive essays (Nguyen and Litman, 2016). While Nguyen and Litman (2016)'s work focused on only the attack and support relations between argumentative components in a paragraph, nonetheless, we believe that discourse relations can be informative

LIWC category	Examples
Analytic	–
Negations	no, not
Interrogatives	how, what
Affective processes	happy
Positive emotions	nice
Negative emotions	hurt
Cognitive processes	cause
Insight	think
Causation	because
Tentative	perhaps
Certainty	always
Perceptual processes	heard
Achievement	success
Power	superior
Past focus	talked
Present focus	is
Future focus	will
Assent	agree

Table 2.5: LIWC features that we used for classifying reputation defence strategies.

features for identifying reputation defence strategies. Here we use shallow discourse relations (*Class level*), including *Comparison*, *Contingency*, and *Expansion* between the question and answer pairs (extracted using End-to-End PDTB-Styled Discourse Parser (Lin et al., 2014)).[6] For example, consider the question and answer pair in Table 2.6, where the discourse relation (parts in bold) between the question and answer is *Comparison* and indicates the *denial* strategy. While fine-grained discourse relations (*type* level) can be informative for identifying reputation strategies, for our analysis, we focused on only major classes of discourse relations because discourse parsers usually yield less reliable results for fine-grained relations.

Syntactic Production Rules Stab and Gurevych (2014a) used production rules to classify support and non-support argument relations in persuasive essays, and found them to be effective features. Their work also focused on the relations in a paragraph. Here, we explore the impact

[6] *Temporal* relations have not been effective in our classification task, which is also in line with expectations (Biran and Rambow, 2011; Stab and Gurevych, 2014a).

Q. Mr. Speaker, contrary to what the Prime Minister says, Canada's actions so far lead us to conclude that it is siding with the United States by supporting, through its silence, comments made by U.S. Secretary of Defense, Donald Rumsfeld, who wants to ignore NATO and the UN if it suits his purposes. Is the Prime Minister aware that his silence is contributing to undermining international institutions and that this complacent attitude breaks with Canada's tradition of respecting major international institutions?	**A. Mr. Speaker, I firmly reject the suggestion that the Prime Minister has been silent.** Our position is clear. We have always encouraged and supported an approach that goes through the United Nations and through the Security Council. We have gotten here, in some measure, thanks to the efforts of the Prime Minister. He has never been silent, he has been active on the international scene and we are very proud of what he has done.

Table 2.6: An example *Comparison* relation between two parts of question and answer, specified in bold; 2003-02-12, Francine Lalonde (Q) and Bill Graham (A).

of the production rules in capturing the syntactic characteristics of reputation management strategies. We consider binary features for production rules (e.g., VP $\rightarrow$ VBZ NP SBAR, VP $\rightarrow$ VB NP PP) that appear only in the answer, and both in the question and the answer (Lin et al. (2009) and Feng and Hirst (2012) used these features for identifying shallow discourse relations and RST discourse relations, respectively). We used the Stanford parser (Klein and Manning, 2003) to perform the pre-processing.

Similarity Measures Simple lexical similarity methods have been shown to be robust in recognizing textual entailment, which can help capture strategies such as *denial* and *concession*. We compute the average semantic similarity between the question and the answer sentences from the cosine similarity between their vectors. To represent the questions and answers, we sum their word2vec embeddings (Mikolov et al., 2013).

2.1.5 Evaluation

The classification is performed using a class-weighted Support Vector Machine model with a linear kernel[7]. The classifiers were trained and tested with the crowd-sourced data (described in section 2.1.3) using five-fold cross validation. The baselines that we use are the majority

[7]LibSVM implementation (Pedregosa et al., 2011).

class, where all instances are classified as *justification*, and the bag-of-words representations (weighted using *tf-idf*) of the question and answer pairs and the bag-of-words representations of answers. The bag-of-words representation of answers is the strongest baseline on our dataset and yields an accuracy of 53.35%. To determine the efficacy of the features, we train individual classifiers on the feature classes. The results are reported in terms of accuracy and average F_1-measure.

Multi-class Classification Table 2.7 reports the results for multi-class classification. The best performance was 57.59% accuracy, which was achieved by using discourse relations and cosine similarity between the question and answer, and verb classes, positive and negative sentiments (extracted using OpinionFinder), negations, and the unigrams from the answers. This model yields a 20-point improvement over the majority baseline and at least a 4-point improvement over bag-of-words baselines. Our studies to measure the contributions of different components show that all features are helpful, with verb classes, sentiments, negations, and unigrams (from answers) being the most helpful for distinguishing between strategies. Using the LIWC features also improves the performance over all the baselines. While production rules are informative features, the performance of this classifier is lower than the bag-of-words baseline.

Table 2.8 reports the average F_1-measure of five-fold cross validation for each reputation defence strategy in multi-class classification. The best performance for identifying *denial, justification,* and *concession* is achieved by the best model. LIWC features are most informative for identifying *excuse* strategy.

Pairwise Classification We further experimented with *pairwise* classification (one-versus-one) for the six possible pairings of the four strategies to find the most informative features for each strategy (Table 2.9). For each of the six classifiers, we considered the data for the two strategies against each other. In *pairwise* classification, almost all models improve over the majority baseline, except for *excuse*, for which the training data is very small.

In distinguishing between *denial* and *justification*, the combination of verb classes, senti-

Features	Acc.(%)	F_1 (%)
Majority Class (justification)	36.50	–
Production rules	49.78	46.31
Unigrams (q + a) (tf-idf)	52.53	49.54
Unigrams (a) (tf-idf)	53.35	51.32
Unigrams (a) (tf-idf) + LIWC	53.57	53.07
Unigrams (a) + VerbNet v class	53.78	51.62
Unigrams (a) + VerbNet v class + Sentiments	56.11	54.02
Unigrams (a) + VerbNet v class + Sentiments + Negation	56.33	55.55
Unigrams (a) + Discourse + Similarity	55.26	53.04
Unigrams (a) + VerbNet v class + Sentiments + Negation + Discourse	56.96	56.33
Unigrams (a) + VerbNet v class + Sentiments + Negation + Discourse + Similarity (best model)	**57.59**	**56.92**

Table 2.7: The performance of different models for classification of four reputation defence strategies (five-fold cross-validation).

ments, negations, discourse relations, cosine similarity, and unigrams from the answers yields the best performance. The features that capture the interactions between reputation threat and reputation defence speeches are most effective for distinguishing between *denial* and *justification* strategies. The most informative features in distinguishing *concessions* and *justifications* are VerbNet classes. In distinguishing between *denial* and *concession,* the features extracted from the answers contribute the most.

Reputation Defence Errors Figure 2.1 shows confusion matrices for the best model, the baseline unigram (a) model, LIWC model, and production rule model for the first fold of cross-validation. The most common confusion is misclassifying the *concession* strategy as the *justification* strategy. The best model makes this error less often. Production rules often misclassify the *concession* strategy as the *denial* strategy as well.

2.1.6 Discussion

The results show that the features proposed above are successful in distinguishing *denial* and *justification* strategies, but the small training set for *excuse* and *concession* strategies did not

Features	Denial	Excuse	Justification	Concession
Production rules	59.4	0.0	51.8	30.8
Unigrams (q + a) (tf-idf)	62.6	10.0	55.6	28.2
Unigrams (a) (tf-idf)	62.4	13.6	55.6	36.4
Unigrams (a) (tf-idf) + LIWC	64.0	**19.4**	54.2	41.0
Best model	**65.0**	18.0	**59.8**	**48.0**

Table 2.8: Average F_1 of different models for classification of four reputation defence strategies (five-fold cross-validation).

		Denial		Justification		Concession	
	Features	Acc(%)	F_1(%)	Acc(%)	F_1(%)	Acc(%)	F_1(%)
Justification	Best model	**74.35**	**74.74**			70.51	69.14
	BOW + LIWC	72.59	72.49			66.39	64.51
	BOW + VerbNet	70.85	70.79			**70.87**	**69.28**
	BOW + VerbNet + Sent + Neg	72.89	72.72			69.39	68.01
	BOW + Discourse + Similarity	73.18	73.04			67.15	65.48
	Production rules	67.95	67.80			65.32	63.43
	Majority	50.44	–			64.55	–
Concession	Best model	76.23	76.40	70.51	69.14		
	BOW + LIWC	**77.36**	**76.72**	66.39	64.51		
	BOW + VerbNet	75.09	74.52	**70.87**	**69.28**		
	BOW + VerbNet + Sent + Neg	**76.98**	**76.91**	69.39	68.01		
	BOW + Discourse + Similarity	75.85	75.16	67.15	65.48		
	Production rules	76.98	75.90	65.32	63.43		
	Majority	64.15	–	64.55	–		
Excuse	Best model	83.02	**81.69**	82.31	78.16	66.35	64.74
	BOW + LIWC	**84.43**	80.28	**83.74**	**79.15**	71.68	67.93
	BOW + VerbNet	82.98	78.89	83.28	78.72	68.60	66.15
	BOW + VerbNet + Sent + Neg	81.57	80.25	81.84	77.85	68.60	66.81
	BOW + Discourse + Similarity	**84.43**	79.91	83.26	78.28	71.71	66.88
	Production rules	82.00	75.01	83.29	76.98	71.71	64.80
	Majority	82.52	–	82.78	–	**72.51**	–

Table 2.9: The performance of the models for pairwise classification (five-fold cross-validation). Best model includes discourse relations, cosine similarity, unigrams, verb classes, negations, and positive and negative sentiments in the answers.

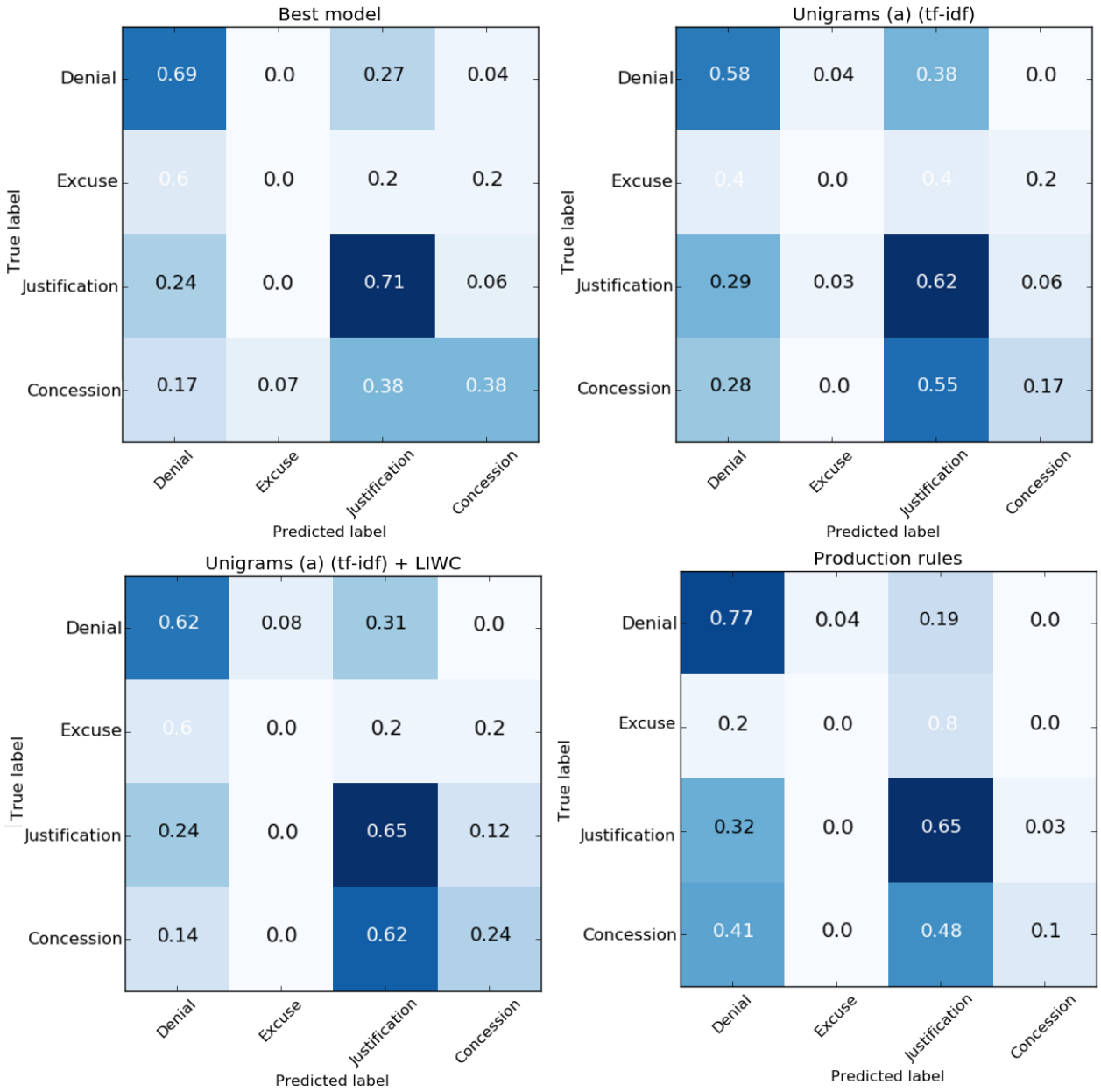

Figure 2.1: Normalized confusion matrices for reputation defence classification.

Q. Mr. Speaker, Canadians are being prevented from obtaining their passports under the guise of increased national security. In the last six months my constituency office has been inundated by hundreds of angry constituents. Some have even been forced to cancel trips, costing them thousands of dollars, due to the incompetence of the government. I have repeatedly raised their concerns with the passport department of foreign affairs to no avail. When the advertised processing time is 45 working days, why are my constituents waiting months for their passports?	**A.** Mr. Speaker, the hon. member was good enough in the introduction to his question to point out there is a problem in terms of new security measures and there is a great deal of increased flow of demands for passports. The passport office is making a serious and concerted effort to respond to these requests. I regret any inconvenience to the hon. member or to Canadian citizens. I want to assure the House that we are taking measures. We have brought in people this weekend and we will be working around the clock to reduce and eliminate the backlog of requests. We have put in measures to enable people to get their passports more quickly and to deal with it more efficiently. I will be circulating to the hon. member, and all members, statements as to how the department is responding to this.

Table 2.10: An example of the *justification* strategy used together with the *concession* strategy; 2003-02-12, Andy Burton (Q) and Bill Graham (A).

allow the model to effectively detect these strategies. While the performance of the model can benefit from more training data, the limited performance could be also due to the labeling task. By limiting the crowd annotators to choose the most prominent strategy, we attempted to study the characteristics of each strategy in isolation, but the results of the annotation process and classification task show that some defence strategies can be employed in combination with each other. Table 2.10 shows an example from our corpus that was misclassified by the model as the *concession* strategy, and when we examined the pair[8], we observed that although the main strategy in the defence is *justification* to reduce the offensiveness, corrective actions are further offered (the *concession* strategy).

Moreover, some questions express multiple reputation threats, which may require multiple defence strategies to address the threats. These cases require further analysis of the reputation threats and allegations. We chose parliamentary debates to study reputation defence strategies

[8]Three annotators marked this relation as *justification* and one annotator marked it as *concession*, we considered agreement by three annotators as gold.

because reputation threat and defence arguments are more numerous in this data, and the data is easily accessible.

2.1.7 Conclusion

We have addressed the task of automatically identifying reputation defence strategies. We showed how the proposed features contribute to the classification of reputation defence strategies and that the models can benefit from the reputation threat information. The analysis is limited in the number of ways. The data set is relatively small and hence not appropriate for approaches that require large training datasets. We limited our analysis to the four most agreed upon strategies and did not take into account the answers that use more than one strategy in our analysis. In the next section, we will explore whether we can improve the classification results and how we can automatically label data with reputation defence strategies.

2.2 Automatically generating labeled data for the classification of reputation defence strategies

Here, we investigate whether we can improve the results of the classification of reputation defence strategies and propose two semi-supervised approaches for identifying these strategies. We rely on our manually annotated corpus to identify lexical information important in recognizing these strategies. One approach uses the observed word pairs from both reputation threat and reputation defence, and the other uses pattern-based representations of reputation defence.

We evaluate a subset of the automatically labeled data against crowd-sourced annotations. We further assess the impact of the extended dataset in a multi-class classification task and show how far we can get with this extended dataset.

2.2.1 Related work

Previous studies on argumentation have shown that manually annotating argument-related information is difficult and results in moderate agreement (Rosenthal and McKeown, 2012; Schneider et al., 2013; Habernal et al., 2014; Aharoni et al., 2014; Wachsmuth et al., 2017). Here, we aim to automatically create a large corpus of reputation defence strategies. We propose two approaches and examine the quality of the extracted data using these approaches.

2.2.2 Data

Here, we used the Lipad[9] (Linked PArliamentary Data) dataset (Beelen et al., 2017). This dataset consists of Canadian Hansards since 1901. We extracted 14,134 pairs of questions and answers from Oral Question period (1994–2014) as our unlabeled data. Since the questions asked by the government backbenchers are generally friendly and intended for clarification, we only focused on the questions asked by the opposition members and their respective answers by the government ministers. Furthermore, we extracted only the first question and answer pairs of each topic of discussion, because the remaining pairs require the context. We made sure that the pairs of questions and answers from the reputation defence dataset were not included in our unlabeled dataset. We extracted two sets of features to assign scores to unlabeled question and answer pairs: (1) observed word pairs, (2) surface patterns. We will discuss these features in the following sections.

2.2.3 Our approach

For our analysis, we used a dataset described in Section 2.1.3. This dataset consists of 493 pairs of questions and answers from Oral Question period from Canadian parliamentary proceedings, manually annotated with four reputation defence strategies (170 pairs of questions and answers are annotated as denial, 36 pairs as excuse, 173 pairs as justification, 95 pairs as concession, and

[9] https://WWW.lipad.ca

19 as none of these strategies). Here, we removed 19 pairs that were annotated as being *none* of these strategies, and focused on the remaining pairs. We refer to this corpus as the reputation defence strategy dataset throughout this section. Given these manually labeled examples, we extracted a set of features to assign scores to unlabeled pairs of questions and answers and automatically expanded the training set.

Pairs of words

Word pairs from a pair of arguments have been shown to be informative features in identifying implicit discourse relations between the two arguments (Marcu and Echihabi, 2002; Pitler et al., 2009; Biran and McKeown, 2013).

Additionally, in the classification of the reputation defence strategies, we have shown that discourse relations between the question and answer sentences can help in capturing the relations between reputation threat and defence instances, and they can be informative features for the detection of reputation defence strategies. Thus, we hypothesized that directly using word pairs in our analysis may improve the identification of reputation defence strategies. Consequently, we considered all the possible word pairs extracted from the cross-product of the question and answer. To represent the relevance of each word pair to each reputation defence strategy, we computed a score using our seed examples. A score is assigned to each question and answer based on simple occurrences:

$$\left(\frac{\text{Count unique word pairs of Label}_i}{\text{Count total unique word pairs}}\right)$$

The raw score was then normalized by dividing by the sum of raw scores of all four strategies.

Pattern extraction

For extracting the surface patterns, we took an approach similar to that of Tsur et al. (2010). Using the extracted unlabeled question and answer pairs, we divided the words into frequent and infrequent words (IFW) according to their relative frequency in the unlabeled corpus and a

specified threshold. This allows us to abstract away from topics and issues. The threshold was set to 1000 per million. The length of patterns was set to be 5 to 7 words with only 3 to 5 slots for infrequent words. Multiple patterns were extracted from each reputation defence answer. We then computed a score for each question and answer pair according to the exact matches of the patterns of each reputation defence strategy. For example, from the *denial* answer *Mr. Speaker, at no time have we interfered with the operations of Air Canada, and I stand by my answer of yesterday*, the following example patterns were extracted:

– *at no time have we* IFW *with*

– *no time have we* IFW *with the*

– *have we* IFW *with the*

– *i* IFW *by my* IFW *of yesterday*

Each question and answer pair was first assigned a raw score for each strategy, and then the score was normalized by the sum of all strategy scores (similar to the approach in Section 2.2.3):

$$\frac{\sum_k \text{Length}(\text{pattern}_k) \times \text{Count}(\text{pattern}_k)}{\sum_i \text{Score of Label}_i}$$

Score of Label_i is a raw score of strategy i.

The extracted word pairs that were assigned highest scores based on the sets of features, patterns, or observed pairs of words were considered as candidates to be added to the training set.

2.2.4 Evaluation and discussion

In order to be able to examine the quality of the extracted candidates, we used a five-fold cross-validation approach for the extension and evaluation of the data. In each fold, we used 94 instances of the reputation defence dataset (described in Section 2.1.3) for test, and the remaining for data extension (extracting patterns and observed word pairs from question and

Q. Mr. Speaker, my question is for the Minister of Human Resources Development. It concerns the government's plans for the end to the TAGS program. How could the minister expect Canadians to take him seriously when he says that the government is working on plans to help out the affected communities after TAGS is finished and we know he is telling the RCMP and his own officials they should get ready for the fact that they will be doing nothing? The minister now has a copy of the leaked document before him. Will he explain why the government is making plans for a social disaster in fishing communities instead of preventing the end of assistance for fishing communities and the people in those areas?	**A.** Mr. Speaker, I have never asked the RCMP to do the sorts of things he said in his question. I understand that some of our officials need some training to be able to cope with confrontational situations and to handle more difficult situations on an individual basis. It has happened not only in relation to TAGS but across Canada. This is the way it works. Our government is doing the right thing by conducting a review of the post-TAGS situation. We are not particularly worried because we trust Canadians and we know Canadians behave properly all the time.

Table 2.11: An example of the *denial* strategy used together with the *justification* strategy; 1997-11-21, Peter Stoffer (Q) and Pierre S. Pettigrew (A).

answer pairs) and classification task. We extended the training data once with only the observed word pairs, and once with only the pattern features. In each fold, the size of the training set varies according to the assigned scores. Since each answer can express multiple reputation strategies (see the example in Table 2.11) or none, we used a threshold value to decide whether to add the candidate QA pair to the training set or not. We examined various threshold values for each approach.

The quality of the labelled QA pairs was evaluated in two ways: (1) comparison with manual annotation, and (2) the contribution of the added training data to the classification of reputation strategies.

Inter-annotator agreement

To examine whether the assigned labels are of high quality, we conducted a study with 180 random question and answer pairs on the CrowdFlower platform. The question and answer pairs were sampled from a pool of pairs that were assigned a reputation strategy label using the two approaches that were described earlier (see Sections 2.2.3 and 2.2.3).

(a) Does the answer express Concession?	(b) Does the answer express Justification?
Q. Mr. Speaker, my question is for the Minister of Labour. Former workers at Singer are arguing that the federal government did not fulfill its contract obligations toward them because it gave the company, instead of them, the Government Annuities Account surplus, that is a part of their pension funds that it was responsible for administering. Does the Minister of Labour not agree that the contract binding the parties between 1946 and 1957 is abundantly clear and that the federal government had an obligation to pay the surplus out to the workers and not to Singer? **A.** Mr. Speaker, all the federal regulations have been applied in this matter.	**Q.** Mr. Speaker, if we understand this correctly, 72% of Canada's refugee claimants have entered Canada from the United States of America, which means that 28% of refugees obviously come from refugee camps. Is the minister telling us that we are only accepting 28% of legitimate refugees to this country who actually deserve to be raised to higher levels? **A.** Mr. Speaker, the member is telling us that legitimate refugees are only people who we picked up, that everyone crossing our borders or arriving at our airports are not legitimate. He should be ashamed of himself.

Table 2.12: (a) Disagreement among six annotators, two of whom annotated it as *concession* and three as not *concession;* 1995-06-01, Claude Bachand (Q) and Lucienne Robillard (A). (b) Three of the annotators confirmed the answer as *justification* strategy and two as not *justification;* 2002-04-30, Rahim Jaffer (Q) and Denis Coderre (A).

All crowdsourced annotations				**Crowdsourced annotations with confidence $>$ 80%**			
(a) Observed word pairs				**(c) Observed word pairs**			
t $>$ .33	**t $>$ .32**	**t $>$ .31**	**t $>$ .30**	**t $>$ .33**	**t $>$ .32**	**t $>$ .31**	**t $>$ .30**
.60	.71	**.73**	.70	.80	**.85**	.77	.76
(b) Extracted patterns				**(d) Extracted patterns**			
t $>$ .90	**t $>$ .80**	**t $>$ .70**	–	**t $>$ .90**	**t $>$ .80**	**t $>$ .70**	–
.41	**.43**	**.43**	–	**.41**	.39	.38	–

Table 2.13: (a) Evaluation of automatically assigned strategies using observed word pairs against all crowd annotations; (b) Evaluation of automatically assigned strategies using extracted patterns against all crowd annotations; (c) Evaluation of automatically assigned strategies using observed word pairs against crowd annotations with confidence $>$ 80%; (d) Evaluation of automatically assigned strategies using extracted patterns against crowd annotations with confidence $>$ 80%. *t* is the threshold used for accepting the candidate labels.

Contributors were shown a question and answer pair with the assigned reputation defence strategy, as well as the description and conditions of the assigned strategy from Table 2.2. The contributors were then asked whether the assigned strategy was correct or not. We asked for at least five annotations per pair from the English-speaking countries. The contributors were presented with one test pair of question and answer and three other pairs on each page, and had to maintain 80% accuracy throughout the job. In total, the task included 66 *denial,* 5 *excuse,* 79 *justification,* and 30 *concession* questions. 81 of 180 were agreed by all 5 annotators. Only 59 answers were annotated with a confidence score below 80%. The confidence score is the agreement of the five annotators weighted by the annotators' trust scores.[10] Trust scores are determined by the annotators' accuracy on the test questions they have seen. Table 2.12 shows two examples of disagreement by the annotators. Most of the answers that caused disagreement among annotators evaded providing a response to the given question.

Table 2.13 shows what percentage of the automatically assigned strategies using word pairs and pattern acquisition approaches were correct compared to the crowdsourced annotations. We once considered all the crowdsourced data. We further removed the crowdsourced annotations with the confidence scores lower than 80%, and assessed the quality of the automatically assigned labels against higher-quality crowdsourced annotations. When compared with the crowdsourced annotations with a confidence score of at least 80%, the labels that were extracted using the observed word pairs approach with the threshold $t > .32$ shows the highest agreement. The automatically assigned labels using pattern acquisition approach show low agreement with the crowdsourced annotations.

Five-fold cross-validation

We further evaluated the quality of the data by assessing its contribution to the classification task. As mentioned earlier, we performed a five-fold cross-validation using the reputation defence dataset. The test set always came from the reputation defence dataset. We performed a

[10] https://success.crowdflower.com/hc/en-us/articles/201855939-How-to-Calculate-a-Confidence-Score

Train	**Original** 379	**t > .33** 512	**t > .32** 1238	**t > .31** 3797	**t > .30** 8495
		BOW			
F_1	51.32	54.65	55.39	52.61	55.28
Accuracy	53.35	56.74	59.10	56.32	62.00
Denial	62.40	64.86	65.69	63.29	75.77
Excuse	13.60	17.00	13.64	13.64	3.64
Justification	55.60	62.42	66.39	63.50	67.14
Concession	36.40	32.00	25.00	14.32	11.02
		BOW+Negation+VerbNet+Similarity+Senti.+Disc.			
F_1	56.92	55.62	54.83	51.86	56.42
Accuracy	57.59	57.37	57.58	55.48	62.85
Denial	65.00	64.73	64.82	63.83	76.60
Excuse	18.00	17.00	17.27	17.00	6.60
Justification	59.80	62.30	64.75	63.05	67.50
Concession	48.00	37.74	24.30	13.01	10.80
		BOW+Negation+VerbNet			
F_1	53.22	54.77	56.01	53.05	55.29
Accuracy	54.22	56.11	58.84	56.74	62.01
Denial	63.60	64.73	65.60	63.45	75.95
Excuse	17.80	14.97	17.27	13.63	3.64
Justification	56.40	60.17	65.63	63.78	67.20
Concession	39.80	36.32	27.56	16.39	10.68

Table 2.14: Classification of reputation defence strategies using the extended training data with observed word pairs. The performance of classification of each strategy is reported in terms of average F_1. t is the threshold used for accepting the candidate labels.

multi-class classification using a class-weighted Support Vector Machine model with a linear kernel[11] and the features proposed in Section 2.1.4, including the bag-of-words representations (weighted using *tf-idf*) of the answers, VerbNet verb classes, positive and negative sentiments, and negations in the answers, as well as discourse relations and similarity measure between the question and answer. We extracted the sentiments using OpinionFinder (Wilson et al., 2005) and discourse relations using End-to-End PDTB-Styled Discourse Parser (Lin et al., 2014). We further used the word2vec embeddings (Mikolov et al., 2013) for computing the similarity between the questions and answers (as described in Section 2.1.4).

Table 2.14 shows the results of the classification with the extended data using the observed word pairs approach. We used various threshold values (t) for accepting the candidates for the extension of the training data (train). Since in each fold the size of the extended data varies, we report the average size of the training sets of all folds. The baseline is the original dataset without any added data (the column specified as *original* in Table 2.14). The average F_1 measure of each reputation defence strategy is also presented. As shown in the table, by adding the automatically assigned labels to the training set, the performance of the classification of the *denial* and *justification* strategies improves; however, the data extension does not improve the classification of the *excuse* and *concession* strategies. Examining the extended data, we find that most of the added instances are *denial* and *justification* instances, and only a few pairs of questions and answers are annotated with the *excuse* and *concession* strategies. The reputation defence dataset consists of the total of only 36 *excuse* and 95 *concession* annotations; thus it is expected that the extended dataset includes very few of these strategies. Using the automatically added labels, the average F_1 measure of *denial* and *justification* reaches about 75% and 67%, respectively.

When we added the discourse relation and sentiment features, we did not observe any improvement in classification for the extended data. This can be due to having noise in the automatically assigned labels, and also the noisy nature of discourse relations and sentiment

[11] LibSVM implementation (Pedregosa et al., 2011).

annotations.

Table 2.15 presents the results of the classification with the extended data using pattern acquisition approach. Extending the data using this approach does not result in a high-quality dataset and the performance of the classification drops very quickly. To improve the quality of the labels, we further examined whether removing the patterns that appeared in all the other strategies help. For example, for denial, we removed the patterns that appeared in non-denial examples. After removing the patterns that were shared between different strategies, we computed the scores introduced in Section2.2.3; however, we did not observe any improvements. Reputation defence strategies do not apply to all question and answer pairs (see the example in Table 2.16), and although we removed the few question and answer pairs annotated with *none* from the seed examples, we might be able to find these cases using a threshold value for accepting the candidate labels.

2.2.5 Conclusion

We presented two approaches to automatically induce a corpus of reputation defence strategies. We considered pattern-based representation of reputation defence strategies and the observed pairs of words from the cross-product of questions and answers. We evaluated the generated data using the two proposed approaches against crowd annotation, and also assessed its contribution in the classification task. The observed word pairs approach resulted in a higher quality dataset. We found that the extended dataset using the observed word pairs contributes positively to the performance of the classifier, even though it contains noisy and weak labels.

2.3 Characterizing the language of reputation defence

Goffman (1967) defines *face*, or *reputation*, as "the positive social value a person effectively claims for himself by the line others assume he has taken during a particular contact. Face is an image of self delineated in terms of approved social attributes". Criticisms and persuasive attacks

	Original	**t > .90**	**t > .80**	**t > .70**
Train	379	453	486	573
		BOW		
F_1	51.32	48.52	47.63	47.51
Accuracy	53.35	49.99	49.15	48.94
Denial	62.40	56.94	54.73	57.54
Excuse	13.60	13.60	11.64	17.00
Justification	55.60	53.44	53.76	52.53
Concession	36.40	34.83	33.60	29.35
		BOW+Negation+VerbNet+Similarity+Senti.+Disc.		
F_1	56.92	49.00	49.10	49.53
Accuracy	57.59	50.84	50.62	51.26
Denial	65.00	56.60	56.01	56.61
Excuse	18.00	13.60	9.40	12.53
Justification	59.80	54.00	54.15	54.65
Concession	48.00	38.60	39.73	40.17
		BOW+Negation+VerbNet		
F_1	53.22	49.81	48.55	48.18
Accuracy	54.22	51.25	49.90	49.36
Denial	63.60	58.10	57.24	57.79
Excuse	17.80	18.10	18.10	27.42
Justification	56.40	54.32	53.30	51.66
Concession	39.80	36.94	34.58	30.89

Table 2.15: Classification of reputation defence strategies using the extended training data with patterns. The performance of classification of each strategy is reported in terms of average F_1. t is the threshold used for accepting the candidate labels.

Q. Mr. Speaker, a week after the latest escalation in the conflict in Bosnia, when 370 peacekeepers, including 55 Canadians, were taken hostage by Serbian forces, there has been a flurry of statements and meetings which failed to produce any concrete results leading to the release of the hostages. This morning, the International Red Cross said that the Bosnian Serbs told them they would release the hostages unconditionally, either today or tomorrow. Could the Deputy Prime Minister confirm the statement by the Red Cross that the Bosnian Serbs will release the 370 peacekeepers who are being kept hostage sometime during the next few hours, although Bosnian Serb leader Radovan Karadzic said yesterday that no hostages could be released without guarantees that all air strikes would be suspended?	**A.** Mr. Speaker, we received communications mentioning that a few hostages might be released today, but at 11.13 a.m., we were unable to confirm whether that was the case.

Table 2.16: An example of an answer where *none* of the strategies apply; 1995-06-02, Gilles Duceppe (Q) and Sheila Copps (A).

pose threats to reputation or face and they are common in all social interactions. Allegations are often made against organizations (e.g., companies and governments) and individuals (e.g., medical practitioners and politicians), and various argumentation tactics and persuasive strategies are used in response to these allegations to attempt to defend the respondent's reputation and thereby save face. Previous studies on reputation defence mostly use manual content analysis, such as the studies by Benoit and Henson (2009) and Zhang and Benoit (2009) on political cases, and Penman (1990) and Tracy (2011) on courtroom cases. While these studies reveal much about reputation defence strategies in various social settings, they do not analyze in detail the actual language used in the defence of reputation.

Here, we examine political speeches and investigate whether we can detect the language of reputation defence. We created a corpus of reputation defence,[12] in which the annotations are based on the structure of parliamentary debate. This corpus is based on the oral question period of a Westminster-style parliamentary system, specifically that of Canada, where the government

[12]The data is freely available at `http://www.cs.toronto.edu/~nona/data/data.html`

of the day is held accountable for its actions and tries to defend its reputation.[13] Using this naturally annotated data lets us avoid the subjectivity of manual analysis, any interpretation by the annotators, and any annotation inconsistencies. We investigate whether we can predict the language of reputation defence and whether the context in which the reputation defence occurs can help in identifying this language. We first perform experiments on a sampled dataset from Canadian parliamentary proceedings of 1994–2014. We then explore the performance of our approaches on two different governments. We show that the context of reputation defence is effective in its recognition.

2.3.1 Related work

Reputation defense is more broadly related to Aristotelian ethos (Aristotle, 2007) or one's credibility that is reflected through the use of language. Previous studies on face-saving and reputation management focused on identifying various persuasive strategies and their effectiveness (Benoit, 1995; Coombs and Holladay, 2008; Burns and Bruner, 2000; Sheldon and Sallot, 2008). In the NLP field, Duthie and Budzynska (2018) focused on extracting ethos from the United Kingdom's parliamentary debates; they used a set of features, such as sentiments and part-of-speech tags, to extract negative and positive references. Here, instead, we are interested in studying whether we can classify a speech as reputation defence or not, and whether the context can improve this classification.

2.3.2 Reputation defence

The main purpose of the oral question period in a Westminster-style parliamentary system is to hold the government accountable for its actions and to highlight the inadequacies of the government.[14] Members of the opposition and government backbenchers both may ask questions,

[13] `https://www.ourcommons.ca/About/Compendium/Questions/c_d_principlesguidelinesoralquestions-e.htm`

[14] The Westminster system originated in the United Kingdom and is used in Commonwealth nations, such as Canada, Australia, India, and New Zealand. The tradition of question time for government accountability is practiced under different names in these countries; in the United Kingdom, it is known as *oral questions*, in Canada

and government ministers must respond. The questions asked by the opposition members are confrontational, intended to criticize or embarrass the government, and are considered reputation threats; the answers to these questions by government ministers try to defend the government's choices and the ministers' reputations. Therefore, these questions and answers are a rich dataset for characterizing the language of reputation attack and the language of reputation defence. Government backbenchers can also pose questions. However, these questions are most often friendly and promotional questions, and the answers given to these questions try to promote the government's plans. Thus these questions and their answers are ordinary reputation-building or reputation-enhancing pairs. They thus act as negative examples.

This dichotomy between the two types of questions in Parliament is supported by qualitative studies such as those of de Ayala (2001), Ilie (2006), and Bates et al. (2012). de Ayala (2001) describes Question Time in the U.K. House of Commons as a "face-threatening genre" and examines politeness strategies used in the face-threatening language of a set of questions. Bates et al. (2012) analysis shows that government backbenchers ask either questions that allow the minister to talk about the government's policies and positions, or questions that are straightforward to answer. While concerns with reputation are of particular importance not only for politicians but are salient in all social encounters, gathering a dataset of reputation threats and defences from encounters other than parliamentary settings is challenging. Hence, we use the available parliamentary proceedings for characterizing these languages.

The following question posed by the opposition in the Canadian Parliament and the Minister's reply to it is an example of a reputation threat and the defence made in response. In the example, the [Deputy] Prime Minister is confronted by an opposition member with a persuasive attack, and he tries to defend and justify the actions of the government:[15]

Example 2.3.1 ***Q.*** *Mr. Speaker, the former finance minister continues to amaze the crowds with his dance of the veils, with the ethics counsellor standing just off stage catching whatever is shed.*

as *oral question period*, in Australia and New Zealand as *question time*, and in India as *question hour*.

[15] 2003-02-20, John Reynolds (Q) and John Manley, Deputy Prime Minister, representing the Prime Minister (A).

The first layer was the blind trust that no one could see through. Next came blind management. Now we are down to the last and flimsiest layer, the supervisory agreement. Could the Prime Minister explain why the former finance minister was allowed the opportunity for hands on management by the ethics counsellor while all other ministers adhered to the stricter blind trust or blind management agreements?

***A.** Mr. Speaker, the arrangements that were in place were those that were appropriate to the circumstances and, in fact, reflect the views of the Parker commission that reviewed these matters in the past. The former minister complied entirely with the requirements before him.*

The next example shows a non-threatening question and answer pair, where the question is posed by a government backbencher.[16]

Example 2.3.2 ***Q.** Mr. Speaker, my question is for the Minister of the Environment. Recently we have been reading more and more articles in the media concerning high levels of sulphur in fuels, air pollution and health problems that result from these high levels. On this issue could the minister tell the House what actions are being taken to deal with the issue of high sulphur levels in fuels in Canada?*

***A.** Mr. Speaker, the announcement I made earlier this year covers gasoline, diesel and fuel oils outside road fuels. It will reduce the amount of sulphur in gasoline from its average now of 360 parts per million to 30 parts per million. In on road diesel, the figure will go from 500 parts per million to 15. The dates for this are the end of 2004 for gasoline and June 1, 2006, for diesel.*

2.3.3 Data

We extracted our Canadian data from the Lipad[17] dataset of the Canadian parliamentary proceedings (Hansard) from 1994 to 2014. This data consists of the proceedings of the 35^{th} to 41^{st} Canadian parliaments. We focused on only the first question and answer pair of each

[16] 2001-06-04, Shawn Murphy (Q) and David Anderson (A).

[17] LInked PArliamentary Data, `https://www.lipad.ca`

Party	Parliaments	Opposition	Government
Liberal	36, 37, 38	11,090	1,736
Conservative	39, 40, 41	11,504	2,004

Table 2.17: Corpus statistics; *Party* shows the governing party; *Opposition* shows the number of questions asked by the opposition members and their respective answers, *Government* shows the number of questions asked by the government backbenchers and their respective answers.

Feature	Ratio	Text
Anger	1.15	**Opp:** Prime Minister has the **annoying** habit of blindly exonerating ...
Negative emotion	1.35	**Opp:** We all know there is a **nasty** trade dispute going on between ...
Positive emotion	0.69	**Gov:** ... presenting new and exciting **opportunities** ...
Achievement	0.82	**Gov:** ... foundation has **successfully** concluded agreements with ...
Cognitive processes	1.20	**Opp:** ... Minister of the Environment **ought to** read the U.S. ...

Table 2.18: Ratios of linguistic features in opposition questions to government backbenchers' questions. Text shows an example for each feature. **Opp** shows an opposition question and **Gov** shows a government backbencher's question.

topic of discussion during the oral question period of parliament sessions in order to minimize dependency on the broader topical context. We created a balanced corpus by randomly sampling the same number of questions posed by the opposition members (reputation threats) as those asked by the government backbenchers (friendly non-threats). This resulted in 9,048 pairs of questions and answers on more than 1,600 issues over the 20-year period.

To further analyze reputation defence strategies used by different governments, we extracted the question and answer pairs from parliaments with different governing parties. The Liberal Party was the government in the 36^{th}, 37^{th}, and 38^{th} Parliaments, and the Conservative Party was the government in the 39^{th}, 40^{th}, and 41^{st} Parliaments. This allows us to examine the language of reputation defence used by different political ideologies. Furthermore, by training and testing models on parliaments with different governing parties, we can ensure that the models are not affected by the ideology of the speaker and the topic of day or interest of the accuser. Table 2.17 shows the statistics of these datasets, which, unlike the 1994–2014 dataset, are not balanced.

2.3.4 Reputation threat analysis

A principled analysis of the language of face-threats or accusations themselves falls outside the scope of this work, but here we characterize the differences between the questions asked by opposition members (reputation threats) and questions asked by government backbenchers (friendly non-threats). We randomly sampled 3,400 questions asked by the oppositions and 3,400 questions asked by the government backbenchers. We performed our analysis using Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010), which is widely used in social science studies. Table 2.18 presents the ratio of averages between reputation threats and non-threat questions for a set of LIWC features, including *anger*, *negative* and *positive* emotions, *achievement*, and *cognitive processes*. Ratios greater than 1.0 indicate features that are more prominent in reputation threats and ratios less than 1.0 indicate features that are more prominent in non-threats. The results show that, unsurprisingly, anger and negative emotions used more in reputation threats than non-threats, whereas positive emotions are used more in non-threats. These features are motivated by theories, such as Brown and Levinson (1987) and Partington (2003) that recognize varying degrees of politeness in threatening or saving the addressee's face. Achievements are used more in non-threats and cognitive processes are used more in reputation threats. This is consistent with theories (Mulholland, 2003) that recognize mentioning the consequences of the fault as one mode of accusation.

2.3.5 Our approach

Convolutional Neural Networks (CNN) process information from various parts of a sentence in parallel using a set of filters that take into account fixed-size sequences of words (LeCun et al., 1998). These models are able to extract the most informative ngrams. Convolutional Neural Networks (CNN) have been shown to be effective for classification tasks (Kim, 2014). Here, we used a CNN model to represent the question and answer pairs for binary classifications of face-saving language. We first represented each word in the question and the answer with its

associated pre-trained embedding. We then applied a convolution operation to each possible window of x words from the question and the answer to produce a feature map, similar to the approach of Kim (2014). We then applied a sliding Max Pooling and concatenated the representation of the question and the answer. We used 20 and 10 filters for the five-fold cross-validation and cross-parliament experiments, respectively. We used filter windows of 3 and 4, a dropout of 0.8, and mini-batch sizes of 32 and 50 for five-fold cross-validation and cross-parliament experiments, respectively.

Recurrent neural networks have been used effectively in NLP for sequence modeling. Here, we further used two long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks[18] with 128 units to represent questions and answers, separately. The LSTM layers were then passed to a dropout layer (Hinton et al., 2012) with a rate of 0.6. We then merged the two representations. For all our Neural Network models, we initialized our word representations using the publicly available GloVe pre-trained word embeddings (Pennington et al., 2014)[19] (300-dimensional vectors trained on Common Crawl data), and restricted the vocabulary to the 5,000 most-frequent words. The models were trained with binary cross-entropy with the Adam optimizer (Kingma and Ba, 2014) for 10 and 5 epochs for five-fold cross-validation and cross-parliament experiments, respectively. We also tried encoding the questions and answers using a layer of Gated Recurrent Units (GRU) (Cho et al., 2014) with shared parameters, but this model performed worse than the other models, and for brevity we do not report the results here.

We further trained an SVM classifier (using the scikit-learn package (Pedregosa et al., 2011)) with all possible combinations of words extracted from the cross-product of questions and answers to capture the interaction between reputation threat and reputation defence. The features are tuples of word pairs from question and answer pairs. We removed word pairs that occurred fewer than 80 times in the datasets. Our use of this set of features is inspired by the effectiveness of word pairs in classifying discourse relations (Biran and McKeown, 2013; Pitler

[18]Using `https://keras.io/`

[19]`https://nlp.stanford.edu/projects/glove/`

et al., 2009) regardless of their sparsity issue.

2.3.6 Evaluation and discussion

We approach the recognition of the face-saving language as a binary supervised classification task. Our baselines are majority class (which is always answers given to the opposition questions), an SVM model trained with answer unigram vectors (weighted using *tf–idf*, represented with the notation '-Answers' in the result tables), and one layer of GRU to model answer sequences. Since reputation defence is expressed in response to the reputation threat, we further considered the question as the context of the reputation defence and trained an SVM model with question and answer unigrams (weighted using *tf-idf*, represented by the notation '-Questions&Answers' in the result tables). For comparison, we further include the results of an SVM model trained on only unigrams from questions ('-Questions'). We also use one layer of GRU to model the concatenation of question and answer pairs as one sequence. The SVM model trained on word pairs is represented with the notation '-Questions$\times$Answers' in the result tables.

In the cross-parliament setting, we used the 36^{th}, 37^{th}, and 38^{th} parliaments with Liberal governments and the 39^{th}, 40^{th}, and 41^{st} parliaments with Conservative governments. We first performed a five-fold cross-validation on the Liberal and Conservative governments individually (three parliaments each), and then performed a cross-parliament classification. For all datasets and models, we randomly used 10% of the training data as the development set. We evaluated the performance of reputation defence classification using the metrics *Accuracy, Precision, Recall,* and F_1. Table 2.19 shows the results of five-fold cross-validation on a balanced set from all parliaments in the period 1994–2014, on just the Liberal governments, and on just the Conservative governments. Both CNN and LSTM models improve the classification compared to the baselines. In general, we can see that all the models that rely only on the answer or reputation defence perform poorer than the models that rely also on the questions. The best model achieves an accuracy and F_1 measure of above 98% on the parliaments with Conservative

Model	Accuracy	F_1	Precision	Recall
(1) Canada 1994–2014; Opposition: 4,524; Government: 4,524				
Majority	50.00			
Unigrams-Answers	76.57	76.57	76.59	76.57
Unigrams-Question&Answers	88.00	88.00	88.01	88.00
Unigrams-Questions	90.10	90.10	90.11	90.10
1 GRU(128)-Answers	81.60	82.64	77.27	89.99
1 GRU(128)-Questions&Answers	**94.39**	**94.23**	**93.94**	**94.91**
CNN(128)-Questions&Answers	91.40	91.16	90.54	92.41
2 LSTMs(128)-Questions&Answers	92.26	91.92	93.34	91.04
Word-pairs-Questions×Answers	91.46	91.46	91.47	91.46
(2) Parliaments 36, 37, 38; Opposition: 11,090; Government: 1,736				
Majority	86.47			
Unigrams-Answers	88.57	88.26	88.10	88.57
Unigrams-Questions&Answers	92.77	92.59	92.50	92.77
Unigrams-Questions	93.59	93.43	93.43	93.59
1 GRU(128)-Answers	90.89	94.91	91.53	98.70
1 GRU(128)-Questions&Answers	**95.72**	**97.52**	96.52	98.66
CNN(128)-Questions&Answers	94.50	96.87	95.12	**98.81**
2 LSTMs(128)-Questions&Answers	94.11	96.52	**97.23**	95.99
Word-pairs-Questions×Answers	95.06	94.95	94.98	95.06
(3) Parliaments 39, 40, 41; Opposition: 11,504; Government: 2,004				
Majority	85.16			
Unigrams-Answers	87.27	86.95	86.82	87.27
Unigrams-Questions&Answers	95.87	95.75	95.78	95.87
Unigrams-Questions	97.45	97.41	97.42	97.45
1 GRU(128)-Answers	91.05	94.93	91.63	98.63
1 GRU(128)-Questions&Answers	**98.33**	**99.02**	98.77	**99.30**
CNN(128)-Questions&Answers	97.10	98.31	97.50	99.20
2 LSTMs(128)-Questions&Answers	97.11	98.27	**98.98**	97.63
Word-pairs-Questions×Answers	97.48	97.43	97.45	97.48

Table 2.19: The performance of different models for binary classification of reputation defence language using five-fold cross-validation on (1) a balanced set from 1994–2014; (2) three Liberal governments; (3) three Conservative governments.

governments. The highest accuracy and F_1 measure on the Liberal dataset is above 95% and 97%, respectively.

Table 2.20 shows the results of the cross-parliament classification. We trained the models on all Liberal parliaments, and tested them on all Conservative governments, and then vice versa. The SVM model trained using question-and-answer unigrams is a strong baseline. Both the CNN and LSTM models improved F_1 measure compared to the baseline models. On the cross-parliament classification setting, again the models trained on both questions and answers perform better. The overall performance of the neural net models across parliaments is poorer than the classification performance within parliaments. This can be explained by the differences in framing strategies used in the language of defence by the two parties, which each defend their actions and choices from their own point of view. While GRU performed better than CNN model in within parliament setting, it performed inferior in cross-parliament setting. This can be also explained by the use of framing by the parties and that RNNs can encode the semantics of the entire input, including the frames used, so within parliaments, they perform better.

The SVM model trained on the words extracted from the cross-product of questions and answers (word-pairs) achieves the best accuracy, reaching an accuracy and F_1 measure above 92% across parliaments. These results show that reputation defence language can be detected with high accuracy regardless of differences in ideologies and framing strategies.

An error analysis shows that most errors occurred in the classification of answers to non-threat questions. One reason for this is that while the government ministers do not defend themselves in the answers in response to the government backbenchers, they do try to enhance their image. Consider the following example[20]:

Example 2.3.3 ***Q.*** *Mr. Speaker, my question is for the Minister of the Environment. Over the weekend, the leader of the Bloc Québécois had the temerity to claim that the 2005 budget did not serve the interests of the people in Quebec. I know full well that the environment is very important to the people in my riding. Could the minister tell the House how the environmental*

[20] 2005-05-31, David Smith (Q) and Stéphane Dion (A).

initiatives contained in the budget will benefit Quebec?

A. *Mr. Speaker, Quebeckers are impatiently awaiting the greenest budget since Confederation. Very successful contacts have been established with the Government of Quebec for the use of the partnership fund. Projects are sprouting up all over for the climate fund, for new investments, for national parks and for investment in renewable and wind energy. Mayors are waiting for green investments for cities and municipalities through the new deal, the green municipal fund, the EnerGuide program for cities and so on. Quebec must not be blocked, but greened even more.*

We further examined the cases where a reputation defence was erroneously assigned a non-defence label. These cases require real-world knowledge to determine that they are indeed reputation defence. Here is an example[21]:

Example 2.3.4 ***Q.*** *Mr. Speaker, this country was built upon common interests by and for the people here. We cannot allow the House of Commons to introduce a bill which, in reality, provides a recipe for destroying this country. Does the government realize that this draft bill is an avowal of failure by this government as far as the future of the federation is concerned?*

A. *No, Mr. Speaker. This bill is a follow-up to the Supreme Court judgment referring back to the political stakeholders the responsibility to establish the conditions of clarity under which they would agree to negotiate the secession of a province from Canada, and it seems to me that one of those stakeholders is the Canadian House of Commons.*

The models that rely on only the answer have particular difficulty in distinguishing these cases.

2.3.7 Analyzing the language of defence

To help discover more about the underlying structure of the data, we conducted an exploratory feature analysis. We created two balanced datasets from the two governments, where each dataset consists of 3,400 question and answer pairs (1,700 questions asked by opposition

[21] 1999-12-13, André Bachand (Q) and Stéphane Dion (A).

Model	Accuracy	F_1	Precision	Recall
Train 36, 37, 38 (Opp: 11,090; Gov: 1,736) and test 39, 40, 41 (Opp: 11,504; Gov: 2,004)				
Majority	85.16			
Unigram-Answers	82.22	82.63	83.10	82.22
Unigrams-Questions&Answers	89.60	89.23	89.02	89.60
Unigrams-Questions	91.56	91.07	91.04	91.56
GRU(128)-Answers	84.02	91.21	85.23	98.25
GRU(128)-Questions&Answers	83.48	90.83	85.65	96.84
CNN(128)-Questions&Answers	85.86	92.32	86.53	**99.10**
2 LSTMs(128)-Questions&Answers	85.27	91.88	86.10	98.66
Word-pairs-Questions×Answers	**93.59**	**93.36**	**93.33**	93.59
Train 39, 40, 41 (Opp: 11,504; Gov: 2,004) and test 36, 37, 38 (Opp: 11,090; Gov: 1,736)				
Majority	86.47			
Unigram-Answers	86.95	85.44	84.87	86.95
Unigrams-Questions&Answers	90.34	89.10	89.40	90.34
Unigrams-Questions	91.14	90.52	90.42	91.14
GRU(128)-Answers	86.29	92.58	86.49	**99.71**
GRU(128)-Questions&Answers	85.58	92.14	86.49	98.75
CNN(128)-Questions&Answers	86.75	92.73	87.67	98.55
2 LSTMs(128)-Questions&Answers	86.72	**92.78**	87.10	99.45
Word-pairs-Questions×Answers	**92.95**	92.31	**92.62**	92.95

Table 2.20: The performance of different models for binary classification of reputation defence in the cross-parliament setting. **Opp** shows the number of opposition members' questions and their respective answers and **Gov** shows the number of government backbenchers' questions and their respective answers.

Model	Accuracy	F_1	Precision	Recall
Train 36, 37, 38 and test 39, 40, 41 (balanced, 3400 instances train and 3400 test)				
Majority	50.00			
Unigrams-Answers	67.94	67.92	67.99	67.94
+NRC Emotion (anger+pos+neg)	69.77	69.70	69.94	69.77
+Bigrams	73.41	73.33	73.73	73.41
+Vagueness cue words	73.85	73.75	74.22	73.85
Word-pairs-Questions×Answers	83.97	83.95	84.14	83.97
Train 39, 40, 41 and test 36, 37, 38 (balanced, 3400 instances train and 3400 test)				
Majority	50.00			
Unigrams-Answers	71.24	70.68	72.99	71.24
+NRC Emotion (anger+pos+neg)	71.71	71.14	73.57	71.71
+Bigram	73.71	72.91	76.88	73.71
+Vagueness cue words	73.88	73.91	76.98	73.88
Word-pairs-Questions×Answers	83.77	83.67	84.82	83.77

Table 2.21: The performance of different models for binary classification of reputation defence in the cross-parliament setting with the balanced data (1700 instances of each class).

members and 1,700 questions asked by government backbenchers). The question and answer pairs were selected randomly. In this setting, we focused only on the text of the answers or reputation defence.

We consider emotions, such as positive, negative, and anger. For extracting these features, we used the NRC Word-Emotion Association Lexicon (NRC Emotion lexicon)[22]. This lexicon provides manually assigned association scores for basic emotions including *anger, fear, joy, sadness, disgust, anticipation, trust, surprise*, and sentiments (*positive* and *negative*) (Mohammad and Turney, 2013). It consists of 14,182 unigrams that are manually annotated through crowdsourcing. We compute the total association scores of the lexicon words in the answer for each class of emotions and sentiments.

We further examined the NRC VAD Lexicon[23] for our analysis. This lexicon provides valence (positiveness–negativeness / pleasure / displeasure), arousal (active–passive), and dominance (dominant–submissive) scores for 20K English words (Mohammad, 2018). These dimensions have been used for analysis of human interaction (Burgoon and Hale, 1984). We use the total score of each dimension in the answer as a feature. We also consider vagueness cue words (Bhatia et al., 2016; Lebanoff and Liu, 2018) that can indicate obscuring language. This set of features (40 cue words) is represented by the frequency of the vagueness cues in the answer. The use of these features is motivated by theories such as that of Fraser (2012) that suggest that hedge words can be used to avoid face-threatening acts. We also use bigrams as additional features. We performed the classification using SVM. The results of the binary classification of face-saving language on the balanced data of the cross-parliament setting is presented in Table 2.21.

The only emotion that contributed to the classification was anger. The positive impact of anger on the classification performance is in line with theories such as those of Mulholland (2003) and Benoit (1995) that find that attacking the accuser is a type of face-saving strategy. Both positive and negative sentiments also improved the performance of the classification, as

[22] http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm
[23] http://saifmohammad.com/WebPages/nrc-vad.html

		Predicted	
		Non-defence	Defence
Actual	Non-defence	1,360	340
	Defence	549	1,151

Table 2.22: Confusion matrix for the best performing model that relies only on features extracted from answers, including unigrams and bigrams, NRC emotions (anger+pos+neg), and vagueness cues. Trained on 36,37,38 (3,400 instances) and tested on 39,40,41 (3,400 instances).

		Predicted	
		Non-defence	Defence
Actual	Non-defence	1,368	332
	Defence	213	1,487

Table 2.23: Confusion matrix for the model trained on word pairs. Trained on 36,37,38 (3,400 instances) and tested on 39,40,41 (3,400 instances).

did vagueness cues and bigrams. However, using valence, arousal, and dominance hurt the performance.

The confusion matrices for the best model trained on the features extracted from the answers (unigrams and bigrams + NRC Emotions including negative and positive sentiments and anger + vagueness cues) and the model trained on word pairs are presented in Tables 2.22 and 2.23, respectively. Both models are trained on 3,400 instances from the 36^{th}, 37^{th}, and 38^{th} parliaments and tested on 3,400 instances from the 39^{th}, 40^{th}, and 41^{st} parliaments.

2.3.8 Conclusion

Face-saving language is employed in everyday human interaction. In this study, we introduced the task of automatically recognizing the language of face-saving. We created a corpus of reputation-defence language on various issues from parliamentary proceedings that is freely available. We examine various approaches to classify this language and find that both lexicalized model and neural net models perform well, but the frames used by different parties can lead neural net models astray. We showed that the context of reputation defence is important for this classification task. Our results supported our annotation decision based on the adversarial structure of the parliament and showed that our corpus is appropriate for analyzing the language

of reputation defence. A practical application of our model will be to analyze human behavior and to examine the effectiveness of reputation defence in various social settings.

2.4 Automated Fact-Checking of Claims in Argumentative Parliamentary Debates

Governments and parliaments that are selected and chosen by citizens' votes have *ipso facto* attracted a certain level of trust. However, governments and parliamentarians use combinations of true statements, false statements, and exaggerations in strategic ways to question other parties' trustworthiness and to thereby create distrust towards them while gaining credibility for themselves. Creating distrust and alienation may be achieved by using ad hominem arguments or by raising questions about someone's character and honesty (Walton, 2005). For example, consider the claims made within the following question that was asked in the Canadian Parliament:

Example 2.4.1 [Dominic LeBlanc, 2013-10-21] *The RCMP and Mike Duffy's lawyer have shown us that the Prime Minister has not been honest about this scandal. When will he come clean and stop hiding his own role in this scandal?*

These claims, including the presupposition of the second sentence that the Prime Minister has a role in the scandal that he is hiding, may be true, false, or simply exaggerations. In order to be able to analyze how these claims serve their presenter's purpose or intention, we need to determine their truth.

Here, we will examine the linguistic characteristics of true statements, false statements, dodges, and stretches in argumentative parliamentary statements. We examine whether falsehoods told by members of parliament can be identified with previously proposed approaches and we find that while some of these approaches improve the classification, identifying falsehoods by members of parliament remains challenging.

2.4.1 Related work

Vlachos and Riedel (2014) proposed to use data from fact-checking websites, such as PolitiFact for the fact-checking task and suggested that one way to approach this task would be using the semantic similarity between statements. Hassan et al. (2015) used presidential debates and proposed three labels — *Non-Factual*, *Unimportant Factual*, and *Check-worthy Factual* sentence — for the fact-checking task. They used a traditional feature-based method and trained their models using sentiment scores using AlchemyAPI, word counts of a sentence, bag of words, part-of-speech tags, and entity types to classify the debates into these three labels. They found that the part-of-speech tag of cardinal numbers was the most informative feature and word counts was the second most informative feature. They also found that check-worthy actual claims were more likely to contain numeric values and non-factual sentences were less likely to contain numeric values.

Patwari et al. (2017) used primary debates and presidential debates for analyzing check-worthy statements. They used topics extracted using LDA, entity history and type counts, part-of-speech tuples, counts of part-of-speech tags, unigrams, sentiment, and token counts for their classification task. Ma et al. (2017) used a kernel-based model to detect rumors in tweets. Wang (2017) used the statements from PolitiFact and the 6-point scale of truthfulness; he compared the performance of multiple classifiers and reported some improvement by using metadata related to the person making the statements.

Rashkin et al. (2017) examined the effectiveness of LIWC (Linguistic Inquiry and Word Count) and stylistic lexicon features in determining the reliability of the news corpus and truthfulness of the PolitiFact dataset. The only reliability measurement reported on the PolitiFact dataset is by Wang (2017), who manually analyzed 200 statements from PolitiFact and reached an agreement of 0.82 using Cohen's kappa measurement with the journalists' labels. Jaradat et al. (2018) used a set of linguistic features to rank check-worthy claims. Thorne et al. (2018) created a dataset for claim verification. This dataset consists of 185,445 claims verified against

Label	True	False	Dodge	Stretch	Total
#	255	60	70	93	478

Table 2.24: Distribution of labels in the *Toronto Star* dataset.

Label	#
True	1,780
Mostly true	2,003
Half true	2,152
Mostly false	1,717
False	1,964
Pants-on-fire false	867
Total	10,483

Table 2.25: Distribution of labels in the PolitiFact dataset.

Wikipedia pages. Here, we do not consider any external resources and we focus only on the text of claims to determine whether we can classify claims as true, false, dodge, or stretch.

2.4.2 Data

For our analysis, we extracted our data from a project by the *Toronto Star* newspaper.[24] The *Star* reporters[25] fact-checked and annotated questions and answers from the Oral Question Period of the Canadian Parliament (over five days in April and May 2018). Oral Question Period is a session of 45 minutes in which the Opposition and Government backbenchers ask questions

Features	F_1	Accuracy	Dodge	True	False	Stretch
Majority class (True)	–	53.35				
BOW (tf-idf)	49.20	53.14	55.20	67.00	4.60	24.80
+ POS	52.92	58.15	62.40	71.00	4.80	27.40
+ NUM	53.40	58.58	**63.80**	70.80	4.80	28.80
+ Superlatives Rashkin et al. (2017)	54.24	59.42	**63.80**	71.60	9.20	30.00
+ PolitiFact predictions	**55.10**	**59.63**	63.60	**71.60**	**12.80**	**30.80**
BOW + NE	50.66	53.33	57.40	66.40	**17.20**	24.40

Table 2.26: Five-fold cross-validation results (F_1 and % accuracy) of four-way classification of fact-checking for the overall dataset and F_1 for each class.

[24] `http://projects.thestar.com/question-period/index.html`. All the data is publicly available.

[25] Bruce Campion-Smith, Brendan Kennedy, Marco Chown Oved, Alex Ballingall, Alex Boutilier, and Tonda MacCharles.

of ministers of the government, and the ministers must respond. The reporters annotated all assertions within both the questions and the answers as either *true*, *false, stretch,* (half true), or *dodge* (not actually answering the question). Further, they provided a narrative justification for the assignment of each label (we do not use that data here). Here is an example of the annotated data (not including the justifications):

Example 2.4.2 ***Q.*** [Michelle Rempel] *Mr. Speaker, [social programs across Canada are under severe strain due to tens of thousands of unplanned immigrants illegally crossing into Canada from the United States.]$_{False}$ [Forty per cent in Toronto's homeless shelters are recent asylum claimants.]$_{True}$ [This, food bank usage, and unemployment rates show that many new asylum claimants are not having successful integration experiences.]$_{False}$*

A. [Ahmed Hussen (Minister of Immigration, Refugees and Citizenship)] *Mr. Speaker, we commend the City of Toronto, as well as the Province of Ontario, the Province of Quebec, and all Canadians, on their generosity toward newcomers. That is something this country is proud of, and we will always be proud of our tradition. [In terms of asylum processing, making sure that there are minimal impacts on provincial social services, we have provided $74 million to make sure that the Immigration and Refugee Board does its work so that legitimate claimants can move on with their lives and those who do not have legitimate claims can be removed from Canada.]$_{True}$*

Here is an example of dodge annotation:

Example 2.4.3 ***Q.*** [Jacques Gourde] . . . *How much money does that represent for the families that will be affected by the sexist carbon tax over a one-year period?*

A. [Catherine McKenna (Minister of Environment and Climate Change)] *[Mr. Speaker, I am quite surprised to hear them say they are concerned about sexism. That is the party that closed 12 out of 16 Status of Women Canada offices.]$_{Dodge}$ We know that we must take action on climate change. Canadians know that we have a plan, but they are not so sure if the Conservatives do.*

For our analysis, we extracted the annotated span of the text with its associated label. The distribution of the labels in this dataset is shown in Table 2.24. This is a skewed dataset with more than half of the statements annotated as *true*.

We also use a publicly available dataset from PolitiFact, a website at which statements by American politicians and officials are annotated with a 6-point scale of truthfulness.[26] The distribution of labels in this data is shown in Table 2.25. We examine PolitiFact data to determine whether these annotations can help the classification of the *Toronto Star* annotations.

2.4.3 Our approach

We formulate the analysis as a multi-class classification task; given a statement, we identify whether the statement is true, false, stretch, or a dodge.

We first examine the effective features used for identifying deceptive texts in the prior literature.

- Tuples of words and their part-of-speech tags (unigrams and bigrams weighted by *tf-idf*, represented by POS in the result tables) (Hassan et al., 2015).
- Number of words in the statement (Hassan et al., 2015; Patwari et al., 2017).
- Named entity type counts, including organizations and locations (Patwari et al., 2017) (represented by NE in the result tables).
- Total number of numbers in the text, e.g., ***six*** *organizations heard the assistant deputy minister* (Hassan et al., 2015) (represented by NUM in the result tables).
- LIWC (Tausczik and Pennebaker, 2010) features (Rashkin et al., 2017).
- Five lexicons of intensifying words from Wiktionary: superlatives, comparatives, action adverbs, manner adverbs, modal adverbs (Rashkin et al., 2017).

[26]The dataset has been made available by Hannah Rashkin at `https://homes.cs.washington.edu/~hrashkin/factcheck.html`.

In addition, we leverage the American PolitiFact data to fact-check the Canadian Parliamentary questions and answers by training a Gated Recurrent Unit classifier (GRU) (Cho et al., 2014) on this data. The choice of this model is motivated by Rashkin et al. (2017)'s study, where an LSTM model trained on a subset of the PolitiFact data outperformed Maximum Entropy and Naive Bayes models. We will use the truthfulness predictions of this classifier — the probabilities of the 6-point-scale labels — as additional features for our SVM classifier (using the scikit-learn package (Pedregosa et al., 2011)). For training the GRU classifier, we initialized the word representations using the publicly available GloVe pre-trained 100-dimension word embeddings (Pennington et al., 2014)[27], and restricted the vocabulary to the 5,000 most-frequent words and a sequence length of 300. We added a dropout of 0.6 after the embedding layer and a dropout layer of 0.8 before the final sigmoid unit layer. The model was trained with categorical cross-entropy with the Adam optimizer (Kingma and Ba, 2014) for 10 epochs and batch size of 64. We used 10% of the data for validation, with the model achieving an average F_1 measure of 31.44% on this data.

2.4.4 Evaluation and discussion

We approach the fact-checking of the statements as a multi-class classification task. Our baselines are the majority class (truths) and an SVM classifier trained with unigrams extracted from the annotated spans of texts (weighted by *tf-idf*). We performed five-fold cross-validation. Table 2.26 reports the results on the multi-class classification task with these baselines and with the additional features described in section 2.4.3, including the truthfulness predictions of the GRU classifier trained on PolitiFact data. The best result is achieved using unigrams, POS tags, total number of numbers (NUM), superlatives, and the GRU's truthfulness predictions (PolitiFact predictions). We examined all five lexicons from Wiktionary provided by Rashkin et al. (2017); however, only superlatives affected the performance of the classifier, so we report only the results using superlatives. Figure 2.2 shows the confusion matrices for fact-checking

[27] https://nlp.stanford.edu/projects/glove/

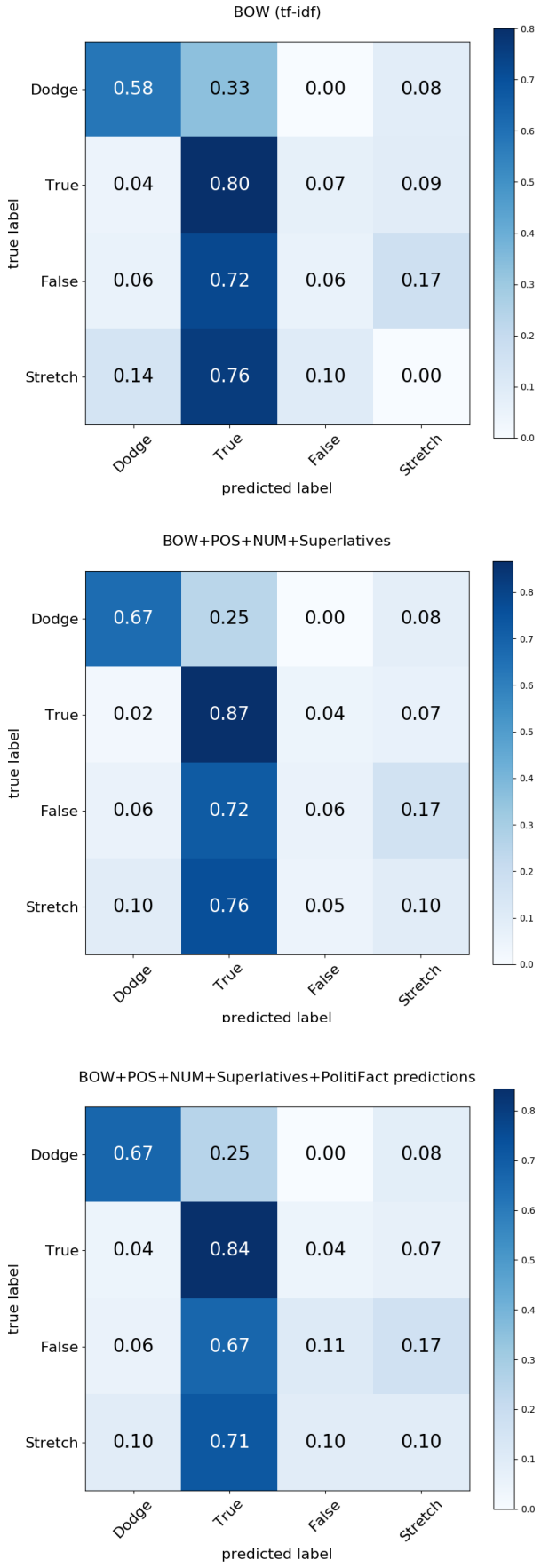

Figure 2.2: Normalized confusion matrices for fact-checking classification.

Features	Dodge	Stretch	False
True			
Majority class	54.84	52.25	58.62
BOW	76.09	54.21	58.20
BOW + NE	75.65	52.99	61.67
BOW + LIWC	52.38	49.11	53.41
BOW + PolitiFact	**77.96**	**55.73**	58.11
BOW + NE + Politifact	76.25	53.76	**63.69**
BOW + POS + NUM + Superlative + PolitiFact	77.51	54.96	55.24
False			
Majority class	53.85	**60.00**	
BOW	81.36	55.89	
BOW + NE	**82.57**	56.91	
BOW + LIWC	52.02	53.31	
BOW + PolitiFact	80.69	52.97	
BOW + NE + Politifact	82.52	55.08	
BOW + POS + NUM + Superlative + PolitiFact	78.29	54.82	
Stretch			
Majority class	57.06		
BOW	75.15		
BOW + NE	76.93		
BOW + LIWC	45.37		
BOW + PolitiFact	79.39		
BOW + NE + Politifact	77.73		
BOW + POS + NUM + Superlative + PolitiFact	**80.59**		

Table 2.27: Average F_1 of different models for two-way classification of fact-checking (five-fold cross-validation).

classification for the first fold of cross-validation.

We also report in Table 2.26 the average F_1 measure for classification of four labels in multi-class classification using five-fold cross-validation. The truthfulness predictions did not improve the classification of the *dodge* and *true* labels in multi-class classification setting. Superlatives slightly improved the classification of all labels except *dodge*.

We further perform pairwise classification (one-versus-one) for all possible pairs of labels to get better insight into the impact of the features and characteristics of labels.

Therefore, we created three rather balanced datasets of truths and falsehoods by randomly resampling the *true* statements without replacement (85 *true* statements in each dataset). The same method was used for comparing *true* labels with *dodge* and *stretch* labels, i.e., we created three relatively balanced datasets for analyzing *true* and *dodge* labels and three datasets for analyzing *true* and *stretch* labels. This allows us to compare the prior work on the 6-point scale truthfulness labels on the U.S. data with the Canadian 4-point scale.

Table 2.27 presents the classification results using five-fold cross-validation with an SVM classifier. The reported F_1 measure is the average of the results on all three datasets for each pairwise setting. *Dodge* statements were classified more accurately than the other statements with an F_1 measure as high as 82.57%. This shows that the answers that do not provide a response to the question can be detected with relatively high confidence. The most effective features for classifying *false* against *true* and *dodge* statements were named entities.

The predictions obtained from training the GRU model on the PolitiFact annotations, on their own, were not able to distinguish *false* from *true* and *stretch* statements. However, the predictions did help in distinguishing *true* against *stretch* and *dodge* statements. None of the models were able to improve the classification of *false* against *stretch* statements over the majority baseline.

Overall, *stretch* statements were the most difficult statements to identify in the binary classification setting. This could also be due to some inconsistency in the annotation process, with *stretch* and *false* not always clearly separated. Here is an example of *stretch* in the data:

Example 2.4.4 [Catherine McKenna] *Carbon pricing works and it can be done while growing the economy. . . . Once again, I ask the member opposite, "What are you going to do?" [Under 10 years of the [Conservative] Harper government, you did nothing.]*$_{\textbf{Stretch}}$

Elsewhere in the data, essentially the same claim is labelled *false*:

Example 2.4.5 [Justin Trudeau] *The Conservatives promised that they would also tackle environmental challenges and that they would do so by means other than carbon pricing. . . . They have no proposals, [they did nothing for 10 years.]*$_{\textbf{False}}$

We further performed the analysis using the two predictions of *more true* and *more false* from the PolitiFact dataset; however, we didn't observe any improvements. Using the total number of words in the statements also did not improve the results.

While Rashkin et al. (2017) found that LIWC features were effective for predicting the truthfulness of the statements in PolitiFact, we did not observe any improvements in the performance of the classifier in our classification task on Canadian Parliamentary data. Furthermore, we did not observe any improvements in the classification tasks using sentiment and subjectivity features extracted using OpinionFinder (Wilson et al., 2005; Riloff et al., 2003; Riloff and Wiebe, 2003).

2.4.5 Comparison with PolitiFact dataset

In this section, we perform a direct analysis with the PolitiFact dataset. We first train a GRU model (used a sequence length of 200, other hyperparameters the same as those of the experiment described above) using 3-point scale annotations of PolitiFact (used 10% of the data for validation). We treat the top two truthful ratings (true and mostly true) as true; half true and mostly false as stretch; and the last two ratings (false and pants-on-fire false) as false. We then test the model on three annotations of true, stretch, and false from the *Toronto Star* project. The

	F_1	True	Stretch	False
Majority	63			
GRU (All)	40	53	29	0
GRU (DQ)	50	75	13	8

Table 2.28: 3-point scale comparison of the PolitiFact data and *Toronto Star* annotations. **All**: GRU model is trained with all PolitiFact data and tested on *Toronto Star* annotations. **DQ**: GRU model is trained with only direct quotes from the PolitiFact data and tested on *Toronto Star* annotations.

	F_1	True	False
Majority	81		
GRU (All)	73	84	29
GRU (DQ)	72	88	8

Table 2.29: 2-point scale comparison of the PolitiFact data and *Toronto Star* annotations. **All**: GRU model is trained with all PolitiFact data and tested on *Toronto Star* annotations. **DQ**: GRU model is trained with only direct quotes from the PolitiFact data and tested on *Toronto Star* annotations.

results are presented in Table 2.28. As the results show, none of the false statements are detected as *false* and the overall F_1 score is lower than the majority baseline.

We further train a GRU model (trained with binary cross-entropy and sequence length of 200, other hyperparameters the same as above) using 2-point scale where we treat the top three truthful ratings as true and the last three false ratings as false. We then test the model on two annotations of true and false from the *Toronto Star* project. The results are presented in Table 2.29; the F_1 score remains below baseline.

The Politifact dataset provided by Rashkin et al. includes a subset of direct quotes by original speakers. We further performed the 3-point scale and 2-point scale analysis using only the direct quotes. Using only the direct quotes, also shown in Tables 2.28 and 2.29, did not improve the classification performance.

2.4.6 Conclusion

We have analyzed classification of *truths, falsehoods, dodges,* and *stretches* in the Canadian Parliament and compared it with the truthfulness classification of statements in the PolitiFact

dataset. We studied whether the effective features in the prior research can help us characterize the truthfulness in Canadian Parliamentary debates and found out that while some of these features help us identify *dodge* statements with an F_1 measure as high as 82.57%, they were not very effective in identifying *false* and *stretch* statements. The truthfulness predictions obtained from training a model on the annotations of American politicians' statements, when used with other features, helped slightly in distinguishing truths from other statements. In future work, we will take advantage of journalists' justifications in determining the truthfulness of the statements as relying on only linguistic features is not enough for determining falsehoods in parliament.

2.5 Conclusion

In this chapter, I have presented four different studies, each investigating face-saving strategy from a different angle:

- The experiment in Section 2.1, evaluated the annotation of reputation defence strategies in parliamentary question and answers. Drawing from Communication Studies, we created an annotation guideline for annotating the most agreed upon reputation defence strategies proposed in the literature. While the language of parliamentary discourse is quite complex, almost 70% of the QA pairs were agreed upon by two or more minimally trained annotators. One strategy that was not among the most agreed upon strategies, but appeared frequently in our data was *dodge* strategy. This strategy also caused frequent disagreement among the annotators. The annotation process of this data is time-consuming because the speeches are quite lengthy and the annotation requires the interpretation of their complex meaning. Using the reliable set of the data (at least three annotators agreed), we extracted a set of features to automatically identify these strategies. We achieved an accuracy of about 0.58 (baseline 0.37) with mostly lexical features. These results illustrate the feasibility of annotating and automatically identifying reputation defence strategies. The results also show that leveraging the face-threatening act improves the classification.

When using this dataset in machine learning experiments, we face two challenges: (1) the dataset is not suitable for approaches that require large training sets (e.g., deep learning architectures). (2) *excuse* strategy is not a popular strategy and we have very few training examples of this strategy, therefore, the models are not able to predict it successfully.

- Next, we investigated whether we can improve the results of the classification of reputation defence strategies. We use our manually annotated corpus to identify the lexical information important in recognizing these strategies and automatically annotate the unlabeled data. We approached this task in two ways: (1) we extracted all word pairs from the cross-product of QA pairs and score the unlabeled QAs based on the simple occurrences of word pairs found in each strategy. (2) we split words into frequent and infrequent groups based on their relative frequency in unlabeled data, and then extracted patterns from each strategy using our manually annotated dataset, where we replace infrequent words with a place-holder, and then score the unlabeled data based on the occurrences of the patterns found in each strategy. The results showed that by adding the automatically labeled data through the word-pair approach to the training set, the classification of *denial* and *justification* improves. We observed that *concession* and *excuse* strategies are too hard or too underrepresented in the data to be successfully modeled through the two proposed approaches.

- In Section2.3, we investigated whether we can predict the language of face-saving. We compared different machine learning algorithms and investigated the impact of lexical features. We found out that models that include the information about the face-threatening act improve the classification results. While neural net models perform well within parliament settings, different frames used by different parties can lead these models astray across parliaments. We further found that regardless of differences in ideologies and framing strategies, we can detect the language of face-saving with high accuracy. The error analysis shows that models have difficulty in recognizing the face-saving language

when the answers are not signaled by indicative linguistic cues, and world knowledge is required to determine that they are indeed reputation defence language.

- Statements with various truthfulness degrees can be used as face-threatening or face-saving acts. In the last section, we examine whether we can automatically detect *true, false, stretch,* and *dodge* statements. We used a corpus of parliamentary statements manually annotated by Toronto Star journalists. We examined whether models trained on the truthfulness labels of the U.S. data (PolitiFact corpus) can help our classification task. We further examined various linguistic features that have been shown to be effective for determining the truthfulness of statements in the prior research. We found that the truthfulness predictions using the U.S. data on their own are not very effective to distinguish false from true and stretch statements in our data. We further found that the linguistic features were not very effective in identifying false and stretch statements. External knowledge sources can be leveraged to facilitate identifying false and stretch statements. Dodge statements were detected more accurately compared to the other labels (F_1 of 82.6% in binary classification setting), suggesting that lexical models are effective in detecting dodge statements.

Chapter 3

Framing

Researchers have taken different approaches to operationalize the concept of framing. Some work used various kinds of topic models to analyze frames. Tsur et al. (2015) interpreted various contexts of a specific topic as frames, and employed topic models and time series to infer them. In a similar study, Nguyen et al. (2015) modeled issues and frame topics using hierarchical topic models. They used bill texts, votes, and floor speeches of the U.S. Congress for their predictions.

The prior work on the analysis of issue-specific frames mostly focused on a limited list of issues and frames. Boltužić and Šnajder (2014) addressed the task of tagging user postings with a pre-existing set of frames for the two topics of *Pledge of Allegiance* and *gay marriage*. Their supervised classification model made use of entailment and semantic similarity features. To generalize their earlier work for various topics, they subsequently presented an unsupervised model to recognize frames on the topics of *abortion, gay rights, Obama,* and *marijuana* by means of textual similarity (Boltužić and Šnajder, 2015). On the same dataset, Hasan and Ng (2014) employed a probabilistic approach to classify forum posts based on users' stance and reasons. In a similar task, Misra et al. (2015) used a set of lexical and semantic similarity features to classify online forum discussions by "argument facets". These methods were all developed for forum posts, in which the language is mainly informal and occasionally, the

contributors explicitly mention their support or opposition towards the proposed thesis. Card et al. (2016) explored the use of persona features to classify entire news articles (on the issue of *immigration*) by their overall generic frames. Baumer et al. (2015) investigated various lexical and syntactic features to characterize framing language on the topic of *national health-care system* in political news stories. They found that imagery, figurativeness, and other lexical features are important in identifying framing language; however, their annotation task was subjective.

In the following sections, we first present our study on classification of sentences by generic frames (Naderi and Hirst, 2017b), and then describe our work on classification of parliamentary speeches by issue-specific frames (Naderi and Hirst, 2016; Naderi, 2016). For generic frames, we used the Media Frames Corpus (Card et al., 2015), a corpus of news articles annotated with generic frames. We explored various features and approaches to predict generic frames expressed in each sentence. We use topic features that were generally used in prior research for representing generic frames and compare them with neural net models. As it has been shown earlier in Section 1.2.2, in order to be able to determine whether a piece of text expresses a specific frame, we need to deal with the complexity of compositional semantics. Neural network models have been shown to be effective in sentence understanding and capturing similarity and analogy, so we hypothesize that these models can represent frames more effectively and use them in our experiments.

For issue-specific frames, we used the ComArg (Boltužić and Šnajder, 2014) corpus and additionally created a corpus of parliamentary speeches. Since this corpus is small and not appropriate for models that require large training data, such as neural net models, we make use of distributional representations to better capture the meaning of the statements. We explored semantic similarity features based on distributional representations to identify issue-specific frames.

Automatic analysis of framing has great potential for decision making (Hammond et al., 1998; Kahneman and Tversky, 2013) because it allows decision and policy makers to access

arguments on a specific issue, and these arguments and frames can greatly influence the choices made. It also allows us to understand human reasoning because framing shows perspectives and the way of seeing matters. Other applications of automatic analysis of frames is in the automatic retrieval (Wachsmuth et al., 2018b) and generation of arguments (Wachsmuth et al., 2018a), and possibly automatic reconstruction of enthymemes (Stede et al., 2018; Boltuzic and Šnajder, 2016) and ideology detection (Cochrane, 2013; Hirst et al., 2014).

3.1 Generic frames

Here, we study the automatic analysis of generic frames on the Media Frames Corpus (Card et al., 2015). This corpus consists of U.S. news articles manually annotated with Boydstun (2014)'s fifteen "framing dimensions", such as ECONOMICS, MORALITY, FAIRNESS, and EQUALITY. The proposed approaches so far addressed identifying the overall frame of the article; however, to analyze the articles at the argumentation level, it is important to study frames at a more detailed level than just the theme of the article. Therefore, we propose to study frames at the sentence level in this corpus. To effectively model the meaning of frames, we rely on Long Short-Term Memory Networks and Gated Recurrent Unit. While the corpus is the largest dataset available for framing analysis, it has some shortcomings, such as low inter-annotator agreement.

Given a text about a controversial issue, our goal is to classify each sentence that expresses a frame relating to the issue (and not just the entire text with a single frame, as Card et al. (2016) did). We use articles from the Media Frames Corpus (see section 3.1.1 below), and our objective in this work is to identify the generic frames expressed in the sentences of these texts. The following example, an excerpt from an article in the Media Frames Corpus (Card et al., 2015), is annotated with the primary frame QUALITY OF LIFE as the overall frame of the article. Individual sentences are annotated with frames (shown in boldface) such as FAIRNESS AND EQUALITY and CULTURAL IDENTITY. The annotations do not always cover the entire sentence, for example, only the first part of the third sentence is annotated, and the second part is not.

Additionally, in some cases, portions of texts are annotated with multiple frames.

Example 3.1.1 *Immigration1.0-171*

[Overall frame of the article: ***Quality of life****]*

*Immigrants say bias is 'swift kick' to citizenship [****Fairness and equality****]*

*When Eduardo Flores moved to Texas in 1981, he was content straddling two cultures: working in the United States but retaining his Mexican citizenship [****Cultural identity****]. Now, the anti-immigrant sentiment spawned by California's Proposition 187 is making him have second thoughts [****Cultural identity****]: Flores wants a claim on the rights available in his adopted land. Legal immigrants like Flores throughout the Southwest have been applying for citizenship at record levels, and many say they want the right to vote to stop the spread of laws like Proposition 187. [****Legality, constitutionality and jurisprudence****]*

3.1.1 Data

The Media Frames Corpus (Card et al., 2015) consists of news articles from 13 national U.S. newspapers published between 1990 and 2012 on three topics of *immigration, smoking,* and *same-sex* marriage.[1] In this corpus, each document is annotated with overall frame (this is what Card et al. (2016) used), and in each sentence, any text that cues a frame is also annotated with that frame, as seen in Example 1 above.

To create our dataset, we first gathered the annotations that at least two annotators agreed upon; however, that process resulted in a small corpus because a majority of the articles on *smoking* were annotated only once. Therefore, we kept the cases that were annotated only once, and for the more controversial cases, where multiple frame dimensions were assigned, we kept only the annotations that were agreed upon by at least two annotators.

[1]We were able to download 4,315 articles from *smoking,* and 5,686 articles from *immigration* using the scripts provided at https://github.com/dallascard/media_frames_corpus. However, we were not able to obtain any of the *same-sex marriage* articles (according to the authors the inter-annotator agreement on the *same-sex marriage* set was much lower than the other two sets, extension of Krippendorff's alpha–that is a chance-corrected agreement, 1 represents perfect agreement and 0 represents the level of chance– 0.08 compared to 0.16 for immigration and 0.23 for smoking).

	Frame	*N* I+S	*N* I
1	Economic	7,070	2,597
2	Capacity and resources	1,516	846
3	Morality	1,185	259
4	Fairness and equality	1,368	559
5	Legality, constitutionality and jurisprudence	9,420	4,233
6	Policy prescription and evaluation	6,505	2,716
7	Crime and punishment	6,206	3,857
8	Security and defense	1,730	1,171
9	Health and safety	4,968	1,054
10	Quality of life	3,790	1,674
11	Cultural identity	4,644	2,264
12	Public opinion	2,496	937
13	Political	7,864	4,253
14	External regulation and reputation	888	438
15	Other	623	278
16	*Irrelevant*	1,256	–

Table 3.1: Frames and number of sentences for each (*N*, extracted from the Media Frames Corpus. I+S includes frames on immigration and smoking; I includes frames on only immigration

We then pre-processed the articles with a sentence splitter,[2] and gathered all the sentences annotated with the cue words for each frame. This resulted in 61,529 sentences in total. Table 3.1 shows the statistics of the resulting dataset.

The sentences were further lower-cased and all numeric tokens were converted to ⟨NUM⟩. Since frames 1, 5, 6, 7, and 13 account for more than 60% of the data, we focused on identifying these five frames; however, we also report the results based on all 15 frames, plus irrelevant category. For all classification tasks, we report 10-fold cross-validation results. For our experiments on immigration and smoking issues, in each fold, we use 30,023 sentences for training, 3,335 for validation, and 3,706 for testing. As mentioned earlier, the majority of the articles on smoking were annotated only once and the reported inter-annotator agreement on this set is very low, therefore, we further removed the irrelevant category and replicated the experiments on only the immigration set, where at least two annotators agreed upon. Table 3.1

[2]Using NLTK (Bird et al., 2009).

shows the statistics of the resulting immigration dataset. On this set, in each fold, we use 21,980 sentences for training, 2,442 for validation, and 2,713 for testing.

3.1.2 Our approach

Here, we present our deep learning–based methods for frame classification. Treating a frame as a sequence of tokens, we explore the use of long short-term memories (Hochreiter and Schmidhuber, 1997) and bi-directional LSTMs (Graves et al., 2013) (BLSTMs), and gated recurrent units (GRU) (Cho et al., 2014) to model the frames. LSTMs and gated recurrent units are types of recurrent neural network that were designed to deal with long-term dependencies, and have been used effectively in the literature to represent the meaning of long sequences for natural language understanding tasks.

To represent the frames, we use word embeddings of the sentences as an input of the model, followed by a single regular LSTM layer, and a sigmoid output layer for multi-class classification.[3] We decided to use a sigmoid function for the output layer because some sentences in our data are assigned multiple labels. We further replace the sigmoid function with a softmax function in the output layer for comparison. We have two settings for initializing our word representations: (1) publicly available GloVe pre-trained word embeddings[4] (Pennington et al., 2014) (300-dimensional vectors trained on Common Crawl data), and (2) embeddings that are constructed on the fly by the LSTM (without any pre-trained word embeddings; we use dropout of 0.2). [5]

We further explore the use of bi-directional LSTMs to represent frame sentences. A bi-directional LSTM consists of two LSTMs running on the input sequence as well as the reverse of the input sequence, thereby allowing the hidden state to capture past and future information (Graves et al., 2013). The motivation behind using this model is to allow the recurrent

[3] Using https://keras.io/

[4] http://nlp.stanford.edu/projects/glove/

[5] We further used publicly available word2vec pre-trained word embeddings (Mikolov et al., 2013) (300-dimensional vectors trained on the Google News corpus), but achieved similar results.

neural networks to decide what sentence context is important for the classification. The input layer relies on the word embeddings that we mentioned above. We took two approaches to use the pre-trained embeddings: we allowed the embedding weights to be updated during the training (with dropout of 0.2), and we also prevented the embeddings from being updated. The output of the bi-directional LSTM layer (similar to the experiments with the LSTM model and GRU model) was passed to a dropout layer (Hinton et al., 2012) with a rate of 0.2-0.5 to avoid over-fitting, and then to a sigmoid layer to predict the class label of the input sentence. Similar to the experiment with the LSTM model, we replaced the sigmoid layer with a softmax layer for comparison. We further use gated recurrent units, which have shown to improve the performance of recurrent neural networks. All models (LSTM, BLSTM, and GRU) were trained with categorical cross-entropy with the Adam optimizer (Kingma and Ba, 2014) for 5 epochs. We experiment with 128 units for all models and restrict the vocabulary to 10,000 most frequent words (for the BLSTM model, we also used the full vocabulary, but achieved a similar performance).

The baselines that we use are majority class and a random forest classifier[6] with 90 trees trained with bag-of-words representations of the sentences. We use both unigrams and bigrams, weighted using *tf-idf*. We further experiment with 20, 50, 100 topic features derived from the Gibbs-LDA++[7] implementation of Latent Dirichlet Allocation (LDA) (Blei et al., 2003). To represent the sentences with the topics, standard English stopwords were removed, and then tokens were lemmatized to their base form. To estimate the parameters, we used $\alpha = \frac{50}{K}$ (K= number of topics) and $\beta = 0.001$, and ran 1,000 Gibbs sampling iterations and estimated the model at every 100 iterations.

Further, we trained a random forest classifier with sentence vectors obtained by summing the pre-trained word embeddings.

Additionally, we used the fastText (Joulin et al., 2016) classifier based on the skip-gram model, where each word is represented as a bag of character n-grams and the classification is

[6]Using scikit-learn (Pedregosa et al., 2011).
[7]http://gibbslda.sourceforge.net/

Table 3.2: The performance of different models for classification of 5 frames on both immigration and smoking (10-fold cross-validation).

Model		Accuracy (%)
Majority Class (frame 5)		25.4
Uni-, bi-grams (tf-idf)		54.2
LDA 20-topics		53.3
LDA 50-topics		53.9
LDA 100-topics		53.2
Sum of vectors, pre-trained GloVe		60.2
fastText		62.0
LSTM (128 units) no pre-trained embeddings	10K	64.5
LSTM (128 units) GloVe	10K	66.7
LSTM (128 units) pre-trained GloVe	10K	67.5
BLSTM (128 units) no pre-trained embeddings	10K	64.6
BLSTM (128 units) GloVe	10K	66.8
BLSTM (128 units) pre-trained GloVe	10K	67.8
GRU (128 units) GloVe	10K	68.1
GRU (128 units) pre-trained GloVe	10K	**68.7**

Table 3.3: The performance of different models for 16-way classification (15 frames plus the irrelevant category) (10-fold cross-validation); B(LSTM) and GRU models use pre-trained GloVe embeddings

Model		Accuracy (%)
Majority Class (frame 5)		15.3
uni-, bi-grams (tf-idf)		38.7
50-topics		36.8
Sum of vectors, word2vec		43.2
Sum of vectors, GloVe		43.2
fastText		48.5
LSTM (128 units)	10K	52.1
BLSTM (128 units)	10K	52.5
GRU (128 units)	10K	**53.7**

Table 3.4: Confusion matrix for GRU with GloVe (5 classes) specified with frame number.

		Predicted				
		1	5	6	7	13
	1	**441**	32	28	31	35
	5	30	**705**	74	159	58
Actual	6	57	149	**268**	55	109
	7	26	78	48	**558**	33
	13	40	40	58	27	**603**

Table 3.5: The performance of different models for classification of 5 frames on only immigration (10-fold cross-validation).

Model		**Accuracy (%)**	**F_1(%)**
Majority Class (frame 13)		24.1	–
Uni-, bi-grams (tf-idf)		64.8	62.4
LSTM (128 units) GloVe	10K	70.5	69.9
LSTM (128 units) pre-trained GloVe	10K	70.5	70.2
BLSTM (128 units) GloVe	10K	70.0	69.7
BLSTM (128 units) pre-trained GloVe	10K	70.2	69.8
GRU (128 units) GloVe	10K	70.2	69.7
GRU (128 units) pre-trained GloVe	10K	**71.2**	**70.7**

performed through a hierarchical softmax.

3.1.3 Evaluation and discussion

Multi-class Classification All classification results are reported in terms of accuracy. Tables 3.2 and 3.3 present the frame detection results for the sets of 5 and 15 frames, plus irrelevant category (sixteen-way classification) respectively on both immigration and smoking issues. The models specified with "*pre-trained*" do not update the embeddings during the training process, whereas the others do update them. All the models reported here used 500 maximum string length with mini-batches of 50 (we also experimented with smaller string length and mini-batches; however, the models achieved lower accuracies). On the combined set, the best accuracy (68.7%) was obtained by the GRU model using 300-dimension GloVe word vectors without being updated, 500 maximum string length with mini-batches of 50. This was slightly

Table 3.6: The performance of different models for 15-way classification on immigration set (10-fold cross-validation)

Model	**Accuracy (%)**	**F_1 (%)**
Majority Class (frame 13)	15.7	–
uni-, bi-grams (tf-idf)	49.7	44.5
LSTM (128 units)	57.7	56.0
BLSTM (128 units)	57.4	56.0
GRU (128 units)	**58.7**	**57.1**

better than the results of uni-directional LSTM and bi-directional LSTM models, which achieve similar performance. We did not observe any performance improvement for the models when the word embeddings were updated. The models achieved very similar results with sigmoid and softmax functions. None of the models that learned the embeddings on the fly outperformed their counterparts initialized with GloVe embeddings, this shows that the semantics that are captured in word embeddings are useful for representing frames. The LSTM, BLSTM, and GRU models all outperformed the baseline random forest classifier with sentence vectors obtained by summing the pre-trained word-embeddings, this shows that this baseline classifier cannot learn the sentence representation of frames as well as language models.

Using the full vocabulary (about 30,000) did not impact the performance of the BLSTM model with GloVe embeddings. All LSTM, BLSTM, and GRU models yielded at least a 10-point improvement over the random forest classifier trained with topics. A confusion matrix for the best-performing GRU model is shown in Table 3.4. The POLICY PRESCRIPTION AND EVALUATION frame is often misclassified as the LEGALITY, CONSTITUTIONALITY, AND JURISPRUDENCE frame, which can be expected, as these frames are more likely to have overlapping expressions.

Tables 3.5 and 3.6 present the frame detection results for the sets of 5 and 15 frames on immigration corpus respectively. LSTM and BLSTM models perform similarly on the immigration set as well. The best performance (71.2%) is achieved again by GRU model, which is about 6-point above the unigram and bigram baseline.

Table 3.7: The performance of one-against the others classification achieved by GRU model on immigration set (10-fold cross-validation)

Frame	Accuracy	F_1	Majority class
1	**92.5**	**92.2**	85.3
5	**84.3**	**83.8**	76.0
6	84.9	82.6	84.6
7	**89.9**	**89.6**	78.2
13	**89.3**	**89.3**	75.9

One-against-others Classification We wanted to see how different frames were effected by the model, so we performed a one-against-others classification, where each frame is tested against the rest of frames in the corpus. Table 3.7 presents the results. We only considered the five most frequent frames. The POLITICAL and CRIME AND PUNISHMENT frames are recognized better than the other frames. While the training set for frame ECONOMIC is smaller than the training set for LEGALITY frame, ECONOMIC frame was detected more accurately. This is probably due to the unambiguous cues and phrases regarding monetary and financial expressions, such as *dollars* and *middle class* that are associated with this frame. The most ambiguous frame is POLICY PRESCRIPTION AND EVALUATION.

3.1.4 Conclusion

In order to represent frames effectively, we employed recurrent neural net models. These models achieved better performance for classification of frames compared to classifiers trained with topics. This suggests that the meaning of sentences that is captured through neural language models is important for identifying and representing frames.

We further employed our models on *immigration* subset of the data in which the annotators achieved lower agreement than the smoking subset. The results showed that the models achieved better performance in spite of a smaller training size and lower agreement. This is most probably due to the annotation process where most smoking annotations were performed by only one annotator, resulting in noisy annotations that are not suitable as a source of training data. Overall,

Table 3.8: Different expressions of frame MARRIAGE SHOULD BE BETWEEN A MAN AND A WOMAN.

Same sex couples may enter into whatever manner of relationship, arrangement or situation that they may desire, but they should not call it marriage because that is a concept that has been clearly understood for millennia.
We are calling on the government to introduce legislation to restore the traditional definition of marriage.
Just leave us, us heterosexuals, the definition of marriage as between a man and a woman, people say, and we will allow them the civil equality of civil unions.

the annotators' agreement on Media Frames Corpus is very low, this can be explained by the fact that inductive approaches to framing are difficult to be replicated. To overcome this limitation of inductive approaches, we take the deductive approach in the next section and use a predefined set of frames to identify frames at the sentence and paragraph levels in the parliamentary debates.

3.2 Issue-specific frames

Due to the significance of framing in argumentation and political discourse, we are studying automatic identification of issue-specific frames in parliamentary discourse. Politicians usually use a set of existing frames to talk about an issue. For example, MARRIAGE SHOULD BE BETWEEN A MAN AND A WOMAN can be expressed in various ways (see Table 3.8).

We hypothesize that these existing frames also used by ordinary people and appear in online forums. To test this hypothesis, we train a classifier on the annotated forums with frames and test them on parliamentary debates. We particularly focus on parliamentary debates as they are used to make decisions and set policies.

Earlier research on detection of issue-specific frames relied on forum debates. Here, we study whether frames in forum posts can help us identify frames in parliamentary discourse. Parliamentary discourse is complex in nature as the members of Parliament occasionally refer to the opposing views or use anecdotes to express their points of view. Therefore, manually

annotating this type of discourse is difficult and time consuming. We created a small corpus on the gay-marriage issue and made use of distributional representations to capture the meaning of frames. While this approach using word embeddings has been to some extent successful compared to bag-of-words representations, it cannot capture information regarding word order and syntactic relations. We further used sentence and syntactic representations; however, since the embeddings were trained on a dataset with a different genre, the model cannot benefit much from them.

3.2.1 Data

For our frame prediction task, we use user-postings manually annotated with known frames (ComArg corpus) as a training set and argumentative parliamentary speeches as a test set. The corpora that we conducted our study on are described in the following sections.

The ComArg Corpus

ComArg[8], developed by Boltužić and Šnajder (2014), is a corpus of user statements manually annotated with users' positions towards a specific topic (pro or con stance), and a set of pre-existing "arguments". These arguments are, in effect, *frames* in the sense that we introduced above, as each highlights certain aspects of the issue. The authors chose two different sources for collecting their data; the user statements are compiled from `ProCon.org`, where the statements are associated with a labeled *pro* or *con* stance, and the frames are taken from `Idebate.org`.[9] The corpus covers two topics of *gay marriage (GM)* and *Under God in Pledge (UGIP)*. Since the latter (regarding the Pledge of Allegiance) is an issue specific to the United States, we focused solely on the GM part of the corpus, which contains 198 statements and 7 pre-existing frames, shown in Table 3.9[10]. In this corpus, the pairs of statements and frames are annotated as *explicit attack, implicit attack, no mention, explicit support,* and *implicit*

[8] http://takelab.fer.hr/data/comarg/

[9] Idebate.org provides a set of manually curated frames for various issues.

[10] The third frame is modified to accommodate frames in our current corpus.

support; that is the statements *for* gay marriage can support the *pro* frames, and attack the *con* frames, and vice versa for statements opposing gay marriage. In this work, we only used the statements that explicitly (176 instances) and implicitly (98 instances) *supported* the pre-existing frames.

Table 3.9: ComArg pre-defined frames on Gay Marriage.

Frame	Stance	Description
1	con	Gay couples can declare their union without resort to marriage.
2	pro	Gay couples should be able to take advantage of the fiscal and legal benefits of marriage.
3	con	Gay marriage undermines the institution of marriage.
4	pro	It is discriminatory to refuse gay couples the right to marry.
5	con	Major world religions are against gay marriages.
6	pro	Marriage is about more than procreation; therefore gay couples should not be denied the right to marry due to their biology.
7	con	Marriage should be between a man and a woman.

Argumentative Parliamentary Statements

For our test set, we focused on debates regarding same-sex marriage in the Canadian Parliament. In 2005, Bill C-38, *An act respecting certain aspects of legal capacity for marriage for civil purposes*, to legalize same-sex marriage in Canada, was introduced in the Parliament. Later that year, the bill was passed and the legal definition of marriage was expanded under the then-Liberal government to include conjugal couples of the same sex. After the Conservative Party of Canada gained power, the debate on same-sex marriage was re-opened in the Parliament in 2006; therefore, the issue was debated extensively in the Parliament in two different periods of time (same-sex marriage was debated briefly in 1999).

We selected speeches regarding same-sex marriage made by the members of the Canadian Parliament from both periods. The corpus described here consists of two sets of debate speeches. The first set consisted of 136 *sentences* of the debate speeches and the second set consisted of 400 *paragraphs* of the debate speeches with an average of 70 words. We asked three annotators to examine the statements in the first set with respect to the position of the speaker towards

same-sex marriage, and assign *pro, con,* or *no* stance. We further asked them to examine which of the pre-existing frames (described in Section 3.2.1) support the statements, and manually annotate them with one of the frames or none; Table 3.12 shows a few examples from our corpus. To measure inter-annotator agreement, we adopted Weighted Kappa metric. Table 3.10 shows the achieved agreement for both stance and frames. For almost 90% of the statements, at least two annotators were in agreement. These statements were kept as the final dataset. Some statements cannot be judged without their context, and annotators did not agree on the stance or the frame. After discarding the statements for which the annotators were not in agreement, the final set has 121 statements. 87 of these remaining statements are supported by one of the ComArg pre-existing frames.

Unlike the first set, for the paragraph set, we asked the annotators to examine the speeches with respect to only the ComArg frames and ignore the stance. The annotation task for this set was carried out by two annotators, and to check the reliability, we computed Weighted Kappa (Table 3.10). The disagreements arose in cases where the speaker used anecdotes or examples. These ambiguous speeches were discarded to create the final dataset. The statistics of the annotated corpora are presented in Table 3.11.

Table 3.10: Inter-annotator agreement on parliamentary discourse corpus.

	Sentences	Paragraphs
Stance	0.54	-
Frame	0.46	0.70

3.2.2 Our approach

The goal of distributed representations is to discover information about the meanings of words using distributional information, for example the words that a given word co-occurs with in a sentence. Distributed word representations are used efficiently in various language understanding tasks, such as sentiment evaluation (Socher et al., 2011). Recently, embedding models such as those of Mikolov et al. (2013), Wang et al. (2015), and Kiros et al. (2015) have provided

Table 3.11: Corpus statistics.

Frame	ComArg annotations		Parliamentary annotations	
	Explicit	Implicit	sentences	paragraphs
1	16	18	14	16
2	12	18	1	14
3	1	4	0	37
4	50	81	33	55
5	28	52	10	56
6	13	17	2	2
7	56	84	27	63
None	0	0	34	123

an effective and easy way to employ word and sentence representations. These distributed representations are real-valued vectors that capture semantic and syntactic content of words and sentences. Here, we use word and sentence vector representations to measure the semantic textual similarity (STS) between the statements and the frames. Our models then use these similarity measures as features to predict a frame that supports a given statement. We used word2vec embeddings (Mikolov et al., 2013) (300-dimensional vectors) trained on Google news articles, and syntactic embeddings (Wang et al., 2015) (300-dimensional vectors) trained on the Annotated English Gigaword, to compute sentence vectors, and further compare them to skip-thought sentence vectors (4800-dimensional vectors) (Kiros et al., 2015). Different composition measures are proposed in literature; one of the simplest measures is additive models (Mitchell and Lapata, 2008), where word vectors are added together to represent a phrase or sentence representation. Here, we used additive models with word2vec and syntactic vectors to represent the statements (sentences or paragraphs) and we compared them with more complex composition functions based on neural language models. After computing the sentence vectors, we measured the similarity of the statement vector representation with the frame representation. We computed two similarity scores between statements and frames: (1) the cosine similarity of the two vectors, (2) the similarity score represented by a concatenation of the component-wise product of two vectors and their absolute difference (P&D) (Tai et al., 2015). We further studied the impact of adding the stance feature (*pro/con*) to the similarity scores as suggested by Boltužić and

Table 3.12: Examples of frame and stance annotations from parliamentary discourse corpus

Stance	Frame	Parliamentary statements
Pro	4	In my opinion, the answer is clear and simple: two people who want to live together within a civil marriage, regardless of their sexual orientation, must be able to do so without the interference of the State.
Con	7	I urge all members who have even the slightest idea that they want to maintain the definition of marriage that we have known and understood for so long to vote in favour of this so that the government can act on it.
–	1	Let me give an example. When we put something in a category, we are discriminating against everything else that is not in that category. If we have a category of things that are blue, then we are leaving out all the yellows, but that does not mean that blue is better or worse than yellow. It just means that they are different.
–	2	If someone puts a lot into a relationship, into a couple, if someone invests in a house and property, that property has to be protected and we must ensure that if both of them invested, both of them reap the benefits. If one of them dies, at a minimum the inheritance must go to the other or be handled in accordance with the person's wishes. It should not be possible to deprive someone of what he or she has built up over the years along with his or her spouse. That is not all. There is not only the legal aspect, of course, but also the emotional aspect. We have to change and progress.

nsubj(abandoning-3, We-1)
aux(abandoning-3, are-2)
root(ROOT-0, abandoning-3)
amod(liberalism-5, traditional-4)
dobj(abandoning-3, liberalism-5)

Šnajder (2014). In addition to the semantic textual similarity and stance features, we also extracted POS-tags, typed dependencies (De Marneffe and Manning, 2008), and distributed representations of the statements. Dependency relation features are extracted using the Stanford parser and they represent relationships between pairs of words. For example, for the sentence *We are abandoning traditional liberalism*, the following triples are extracted:

Our supervised model then takes these features as input, and learns to identify the frames. For supervised learning, we use SVM^{light} and $SVM^{multiclass}$ by Joachims.[11][12]

[11] http://svmlight.joachims.org/
[12] https://www.cs.cornell.edu/people/tj/svm_light/svm_multiclass.html

3.2.3 Evaluation and discussion

In the first experiment, we use the statements from ComArg as a training set and the Canadian parliamentary statements on GM as a test set for our classification task. We first remove the stop-words, and then sum the vector representations of the remaining words in the sentences to compute the sentence vectors. For syntactic embeddings, we only used the noun, adjective, and verb embeddings. In case the vector representation for a given word is not found in the embeddings, the lemma of the word is searched and retrieved.

After representing the statements and frames using word2vec, the syntactic-based embedding model, and the skip-thought model, we computed the semantic similarity of each pair with the similarity measures described in Section 3.2.2.

Our baselines are the majority class and bag-of-words (with TF-IDF vectors and rare words removed) classifiers. Table 3.13 summarizes our results. We observe that almost all models that use STS features outperform the baselines. We also observe that the P&D similarity score provides a better measure for capturing the meaning of the statement-frame pairs. Furthermore, adding the stance feature to the cosine similarity scores improves the accuracy of the classifiers; however, adding it to P&D has no impact on the accuracy of the classifiers. Although the training set of explicit statements is smaller than the training set of explicit and implicit statements, the best results are mostly achieved by training the classifier on explicit instances. Furthermore, adding the stance feature to the cosine similarity scores gives an improvement of about 20 to 40 percentage points in accuracy above the baseline.

Without using the stance feature, the best score was obtained by training the classifier on explicit and implicit instances with the P&D similarity score of word2vec vectors. While we expected to achieve better accuracy with injecting syntactic information through syntactic embeddings and skip-thought vectors, the results do not show such improvements. This can be due to multiple reasons. First, syntactic embeddings were trained on a smaller set compared to word2vec embeddings. Furthermore, we only rely on three categories of syntactic embeddings

(nouns, verbs, and adjectives), whereas even prepositions, such as *against* and *between* are informative features for predicting some frames. Skip-thought models are moreover trained on a dataset with a different genre. One of the challenges of using forum posts as a training set is that they are filled with spelling errors, and in our experiments, we did not correct any of these errors. For our paragraph corpus, since our training corpus based on ComArg is very small, we focused on the two dominant frames in ComArg corpus, frames 4 and 7, and used both explicit and implicit statements for training our models. The paragraph vectors were constructed by adding sentence vectors. For this set, in addition to STS features, we explored features based on POS-tags, the typed dependencies, and the vector representation of the statements. Despite the usefulness of the stance feature as we have seen in the first set, we decided to ignore this feature for our second experiment. The reason for this is that stance is not always known, particularly for the frames of the issues that are not highly polarized. Furthermore, we believe some frames can be used with either positions, for example:

Example 3.2.1 *Earlier this year France rejected the marriage of same sex couples because of the effect that same sex marriages have on children.*

Example 3.2.2 *Are we going to divide this country into those children who are children of certain couples and children who are not? If we truly value children in the House, then we must understand, as one of the members spoke about children, that this is about the rights of the child, regardless of what their parents do, do not do or who they are.*

Both examples are supported by the frame, *impact on children*; however, the position of the speaker in the first example is against gay marriages, whereas the second speaker supports them.

Similar to the first set, most of the models using STS features outperform the baselines in the paragraph corpus (shown in Table 3.14). The best results were achieved by the P&D similarity score of word2vec features, followed by the word2vec features extracted from the statements. Another observation is that the models based on features extracted from the statements perform better than the models based on cosine similarity features.

Table 3.13: Frame prediction results on parliamentary sentences, trained on ComArg data and tested on parliamentary debates.

Features	**Accuracy (%)**
Majority Class (argument 4)	37.93
ComArg – Explicit+Implicit	
Bag of Words (BoW)	48.2
STS (Sum of vectors, word2vec, cosine)	54.0
STS (Sum of vectors, word2vec, cosine)+stance	**72.4**
STS (Sum of vectors, word2vec, P&D)	58.6
STS (Sum of vectors, word2vec, P&D)+stance	58.6
STS (Sum of vectors, syntactic embeddings, cosine)	49.4
STS (Sum of vectors, syntactic embeddings, cosine)+stance	**68.9**
STS (Sum of vectors, syntactic embeddings, P&D)	50.5
STS (Skip-thought vectors, cosine)	48.2
STS (Skip-thought vectors, cosine)+stance	**68.9**
STS (Skip-thought vectors, P&D)	51.7
ComArg – Explicit	
Bag of Words (BoW)	52.8
STS (Sum of vectors, word2vec, cosine)	55.1
STS (Sum of vectors, word2vec, cosine)+stance	**73.5**
STS (Sum of vectors, word2vec, P&D)	57.4
STS (Sum of vectors, word2vec, P&D)+stance	57.4
STS (Sum of vectors, syntactic embeddings, cosine)	54.0
STS (Sum of vectors, syntactic embeddings, cosine)+stance	**68.9**
STS (Sum of vectors, syntactic embeddings, P&D)	56.3
STS (Skip-thought vectors, cosine)	52.8
STS (Skip-thought vectors, cosine)+stance	**68.9**
STS (Skip-thought vectors, P&D)	57.4

Table 3.14: Frame prediction results on debate paragraph corpus using ComArg corpus (Explicit+Implicit)

Features	**Accuracy (%)**
Majority Class (argument 7)	53.3
Bag of Words (BoW)	71.0
Dependency features	72.0
Sum of vectors, word2vec	72.9
Sum of vectors, syntactic embeddings	64.4
STS (Sum of vectors, cosine, word2vec)	61.8
STS (Sum of vectors, P&D, word2vec)	**75.4**
STS (Sum of vectors, cosine, syntactic embeddings)	61.4
STS (Sum of vectors, P&D, syntactic embeddings)	62.7
STS (Skip-thought vectors, cosine)	53.3
STS (Skip-thought vectors, P&D)	59.3

We further report our results on five-fold cross-validation of four most frequent frames (frames 3, 4, 5, and 7) in our paragraph corpus (shown in Table 3.15). The best results were achieved by the model based on features extracted from the statements, followed by the P&D similarity measure. BOW achieves better performance compared to the other models.

Since the members of the parliament usually refer to the opposing viewpoints and their frames during the debates, relying on all the statements in the paragraphs for extracting features for the models cause errors. The following example was not successfully tagged with the frame due to treating all the statements in the paragraph in the same way.

Example 3.2.3 *Peace River constituents are not opposed to equal rights. In fact, the majority support the legal extension of rights and benefits to same sex couples. However, most are opposed to changing the historical term 'marriage' to include these unions. Many have strongly held religious views and are extremely worried that their long-held beliefs are being threatened by the same-sex marriage act. I do not think these views are limited to my riding; I believe they are shared by a majority of Canadians.*

By comparing the predicted frames with the annotations, we noticed that in cases where anecdotes are used to frame the issue, some models were more susceptible to errors; for example:

Table 3.15: Five-fold cross-validation (4 frames).

Features	**Accuracy (%)**
Majority Class (argument 7)	29.8
Bag of Words (BoW)	65.0
POS tags	63.0
Dependency features	53.8
Sum of vectors, word2vec	**70.4**
Dependency features+ word2vec	69.0
Sum of vectors, syntactic embeddings	62.8
Sum of vectors, skip-thought	54.7
STS (Sum of vectors, cosine, word2vec)	42.3
STS (Sum of vectors, P&D, word2vec)	**67.6**
STS (Sum of vectors, cosine, syntactic embeddings)	39.7
STS (Sum of vectors, P&D, syntactic embeddings)	60.9
STS (Skip-thought vectors, cosine)	41.8
STS (Skip-thought vectors, P&D)	58.6

Example 3.2.4 *Like Canada, the Netherlands has many historic ties to other parts of the world, such as Aruba in the Caribbean which, since 1986 has been a separate entity within the Kingdom of Netherlands. After a Dutch lesbian married an Arubian lesbian in the Netherlands, they moved to Aruba and expected their marriage would be recognized there. Instead, their application to register their marriage was denied amidst significant degrees of social pressure that ultimately compelled the couple to return to the Netherlands.*

The speaker uses an anecdote to express that it is discriminatory to refuse gay couples the right to marry.

3.2.4 Conclusion

We created a small corpus of parliamentary debates annotated with a set of pre-existing frames. Our annotation study showed that annotators achieve higher agreement at the paragraph level. The results indicate that the proposed models that rely on distributional representations perform better in representing frames. Furthermore, the similarity score represented by a concatenation of the component-wise product of two vectors and their absolute difference better captures the

similarity of frames compared to the cosine similarity. This work, however, has a number of limitations. First, the size of dataset is small and limited to frames regarding one issue. To address these limitations, we propose to use manifesto frames.

3.3 Conclusion

In this chapter, we have presented two studies that investigate framing.

- The experiment in Section 3.1 operationalizes framing as a classification task of generic frames at the sentence level. We investigated various machine learning algorithms and the results showed that neural network models perform better compared to the classifiers trained on topic models and other strong baselines. Framing, similar to other language understanding tasks, is a semantically difficult problem to solve and a successful model should be able to deal with the complexity of compositional semantics and to capture similarity and analogy. Neural language models can better learn representations for sentences that allow them to judge whether any given sentence expresses the abstract representation of a frame.

- Section 3.2 studies whether and how political officials frame issues and presents a corpus of parliamentary speeches annotated with a set of existing frames at the sentence and paragraph levels. There are various challenges associated with the annotation process. The annotation task requires careful analysis of the complex speeches and knowledge of the issue at hand, and hence it is not appropriate for crowd-sourcing. We leveraged a corpus of user comments manually annotated with the existing frames to train the models and examined various embeddings to represent the statements. The results support our hypothesis that frames are transferable across genres; however, the analysis is limited to only the existing frames and the new frames that appear in the parliamentary data are not detected. Furthermore, the analysis also relies on the pre-existing frames, which may not be available for all issues. Since gay-marriage is a highly polarized issue, stance can be

an informative feature for identifying certain frames; however, stance information is not always known for frames.

Chapter 4

Argumentation Quality Assessment

In previous chapters, we have shown that, based on the goal of argumentation, various persuasive strategies, such as framing and reputation defence, can be employed in arguments. Each of these strategies relates to different aspects of arguments. For example, face-saving strategies are related to logical (e.g., false claims), rhetorical, and dialectical aspects. As mentioned earlier, various models of argumentation have been proposed in the literature, for example the models that focus on the monological aspect of an argument, the models that focus on rhetorical aspect, and the models that focus on dialogical aspect of arguments, but a comprehensive analysis of how these models contribute to the evaluation of arguments is missing.

Can we automatically assess an argument? If so, how? McPeck (2016) believes that the proper assessment of arguments should be left to the experts and those who have a working knowledge of that field. According to McPeck, the assessment of even everyday arguments regarding social issues, such as *the rights of minorities* or *tax roll-back proposals*, requires critical thinking and understanding complex domain information.[1] Then, what considerations should one have in mind when assessing an argument? Can we still find some guidelines to allow us to evaluate arguments in any discipline?

Automatic assessment of arguments requires an understanding of what constitutes the quality

[1]McPeck's perspective on argumentation assessment has been criticized by some scholars, such as Gover (2018).

of argument. The following sections describe a study of the assessment of argumentation quality that I contributed to. The first section provides an extensive review of the theoretical and practical approaches to the assessment of argumentation quality and develops a set of dimensions for the assessment of argumentation quality. The follow-up work compares practical views of argument quality with theoretical views.

These sections are joint work with colleagues from Bauhaus-Universität Weimar, Technische Universität Darmstadt, IBM Research Dublin and Haifa, Stanford University, and the University of Toronto.[2] My contributions to these studies were primarily (i) conducting the annotation study for the dimensions and (ii) conducting the crowd-sourcing annotation task. In addition, I also (iii) assisted in literature review and in the development of several of the dimensions, (iv) contributed to annotating the arguments, (v) participated in the discussion of the studies, and (vi) assisted in the writing of the papers.

4.1 Computational Argumentation Quality Assessment in Natural Language

The contents of this section were published in the Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 17), pages 176–187, April 2017. The paper can be found in Appendix A.

The work on the computational analysis of argumentation quality was started at the Dagstuhl seminar 15512 that Professor Hirst was involved in. The working group on the argumentation quality was led by Dr. Henning Wachsmuth. Dr. Wachsmuth had the idea of achieving a common understanding of what argumentation quality is and defining dimensions of argumentation quality. This understanding allows us to devise computational models to automatically assess and evaluate argumentation. Argumentation assessment is particularly of interest in educational and policy-making systems. The language of science is replete with argumentation, making it

[2] All my co-authors have given their permission to include these manuscripts in the appendix of the dissertation.

important for students to learn how to examine knowledge claims, whether they are justified by relevant evidence, and whether alternatives are accounted for. These analysis skills are also necessary for evaluating policies that influence the health and safety of a population.

As a first step to achieve such an understanding, all the major theories and studies of argumentation quality were reviewed and a taxonomy of argumentation quality dimensions was derived. The details of the taxonomy can be found in section 3 of the paper in Appendix A. Then, to evaluate how subjective and complex these dimensions are, we conducted an annotation study, which I ran. We began with a pilot study with 20 comments from the unshared task dataset (variant D) of the 3rd Workshop on Argument Mining. For the annotation task, we provided the opinionated article, the given comment with its preceding comment, and asked expert annotators to first examine whether the given comment is argumentative and if so, then, rate it based on all the dimensions using a 4-point scale (e.g., *highly acceptable, rather acceptable, rather unacceptable,* and *highly unacceptable* in addition to *cannot judge*). The annotations were performed by seven expert annotators who were among the authors of the paper. This analysis resulted in a low agreement (the highest inter-annotator agreement was .22 in terms of Krippendorff's α for *local sufficiency*). After refining the guidelines, we conducted the annotation task on a subset of *UKPConvArgRank* dataset by Habernal and Gurevych (2016a). This dataset contains 1,052 comments on various social issues annotated with convincingness through crowd-sourcing. In contrast, we examined each argument on its own merits based on all the quality dimensions using a 3-point scale (*high, average,* and *low*). In total, 304 arguments were rated by three expert annotators. The inter-annotator agreement of the ratings in terms of Krippendorff's α ranged between 0.26 for *emotional appeal* and 0.51 for the *overall quality*. The most agreed-upon dimensions in terms of α were *local acceptability* (whether the premises are believable) and *local relevancy* (whether premises contribute to the acceptance or rejection of the claim). However, in terms of full agreement and majority agreement, *local acceptability* was placed in the lower range of agreement. This can be explained by not accounting for the potential bias caused by the prior beliefs in our analysis. This may be improved by changing

the definition of *acceptability,* and, as suggested by Macpherson (2006) and Macpherson and Stanovich (2007), by providing a set of instructions to unbias against the prior knowledge or beliefs. Macpherson and Stanovich (2007) specified in their task that the participants had to assume that the premises were true, even if they appeared to be false, and then they had to decide whether the conclusion followed *logically* from the premises or not.

We further used the Pearson correlation to compare the correlations among the dimensions and dimensions with the overall quality. Not only were *cogency, effectiveness,* and *reasonableness* strongly correlated with each other, they were also strongly correlated with the overall quality of arguments. Dimensions of *logical* aspect were more correlated with those of *dialectical* aspect ranging from .68 to .78. At the *logical* level, *cogency* was highly correlated with *local sufficiency*, and at the *dialectical* level, *reasonableness* was highly correlated with *global acceptability*. The correlations of rhetorical dimensions, such as *credibility* and *appropriateness*, and *credibility* and *clarity* were lower than expected due to their subjectiveness.

Our annotation study further showed that most arguments do not provide sufficient premises for their claims and very few arguments account for anticipated counter-arguments. This rare account of counter-arguments can be due to prior beliefs that overlook the alternatives (Baron, 1995; Macpherson and Stanovich, 2007), or the lack of prior knowledge to generate rebuttals, and/or the level of reasoning and argumentation skills to generate high-quality arguments (Means and Voss, 1996; Mason and Scirica, 2006). We further observed that very few arguments create trust and construct credibility, which can also be due to the lack of prior knowledge to provide detailed reasons and display expertise, and/or due to the level of reasoning and argumentation skills (Means and Voss, 1996).

This annotation analysis can also be used as a baseline for analyzing computational models of argumentation quality assessment.

4.2 Argumentation Quality Assessment: Theory vs. Practice

The contents of this section were published in the Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 17), pages 250–255, July 2017. The paper can be found in Appendix B.

In this study, we performed a comparison of our approach to argument quality and Habernal and Gurevych (2016b)'s approach to argument convincingness. In the taxonomy derived from theory, 15 dimensions of quality within *logical, rhetorical,* and *dialectical* aspects of argumentation are defined, such as *local sufficiency, clarity,* and *global relevancy*. A corpus of comments (a subset of Habernal and Gurevych (2016a)'s dataset) was also annotated based on these dimensions by three expert annotators. In this study, we compare the annotations of this corpus with the annotation of Habernal and Gurevych (2016b)'s dataset. In Habernal and Gurevych (2016b)'s crowd-sourced dataset, arguments are compared in pairs in terms of convincingness, and the reasons of preference of one argument over another are provided by the crowd. These reasons were then categorized into 17 categories, such as *A sticks to the topic, B uses irrelevant reasons*, and *A is more credible/confident*. To compare the annotations of the two datasets, the quality dimension ratings were converted to paired comparison ratings and Kendall's τ rank correlation was computed. While most correlations made sense and were in line with their definitions, such as *appropriateness* and *sticks to the topic* with a correlation of .79, there were a few exceptions, such as *local relevancy* and *irrelevant reasons* with a correlation of .45. The details of this correlation analysis can be found in section 3 of the paper in Appendix B.

In this study, we further conducted an annotation study, which I ran, using a crowdsourcing platform to examine whether the dimensions can be annotated by lay annotators. We used the same arguments with the same instructions for the annotation process. For reliability purposes, we asked for 10 judgments per argument and dimension. We used Krippendorff's α to compute the inter-annotator agreement between the crowd and expert annotations. We computed the

estimate of the crowd annotation in two ways, one using the mean of annotations and one using MACE (Hovy et al., 2013) (suggested and performed by Dr. Ivan Habernal). The inter-annotator agreement of the expert with the crowd was similar to the inter-annotator agreement of the experts. The highest agreement was observed for *global acceptability* (.54) and *appropriateness* (.54) and the lowest agreement was observed for *global sufficiency*. We further examined how many crowd annotations can form a reliable annotation by splitting the crowd annotations into two independent groups of 5. We computed the inter-annotator agreement among the expert annotations with the estimate annotation of each group. The results showed that 10 judgments can form a more reliable annotation.

4.3 Conclusion

This chapter proposed a unified approach for various perspectives on the assessment of arguments. We investigated what aspects of argumentation can contribute to its assessment and proposed 15 quality dimensions. We further developed the first corpus annotated with all the dimensions that we proposed. The annotation process highlights that the assessment of some dimensions is subjective. This can be due to different persuasive strategies used to target the audience and also the prior belief or knowledge that can impact persuasion (Baron, 1995; Durmus and Cardie, 2018), and/or evaluation of arguments objectively (Petty et al., 1981; Baron, 1995; Stanovich and West, 1997; Macpherson and Stanovich, 2007). In this study, we assessed only everyday arguments on various social issues and did not assess domain-specific arguments, for example arguments in chemistry or physics. Obviously, some of these dimensions, such as emotional appeal, do not apply to such arguments. In the following study, we compared our approach and quality dimensions with Habernal and Gurevych (2016b)'s approach of analyzing convincingness and the reasons that they derive from crowd-sourcing experiment. We found that their notion of convincingness is correlated with our overall quality and most of convincingness reasons are represented by the dimensions. We further performed a crowd-sourcing study to

examine how lay annotators can assess arguments using the quality dimensions and found that the agreement similar to that of the experts is limited.

Chapter 5

Conclusion and future work

5.1 Summary

In this dissertation, I have presented the computational analysis of a number of persuasive strategies, including face-saving and framing strategies to better understand and evaluate argumentation. Below, I summarize the contributions of this dissertation.

The analysis in Chapter 2 lays out the first computational analysis of how individuals work on their credibility through the use of language. Drawing from communication studies, I created an annotation guideline for the analysis of reputation defence strategies in parliamentary QAs and the development of an annotated corpus through crowd-sourcing. This annotation study showed the limitation of Benoit's theory to identify *dodge* as a face-saving strategy in our data. Dodging the face-threatening act may not seem to be an effective strategy, but nevertheless it is used frequently in political argumentation. I further operationalized the analysis of reputation defence strategies as a classification task and proposed effective features to identify the most agreed upon strategies in the literature. This study showed that the classification of reputation defence strategies, although challenging, is feasible to tackle on the complex texts of parliamentary debates. The results showed that the features that capture the interactions between face-threatening acts and face-saving acts, including the discourse relations and semantic

similarity, can improve distinguishing between *justification* and *denial* strategies. Inspired by the effectiveness of the word-pairs in the classification of discourse relations, I extracted all word-pairs from the cross-product of the questions and answers and used them to extend the training set for the classification of reputation defence strategies. The results confirmed my hypothesis that word-pairs are effective in the classification of face-saving strategies and showed that the classification of justification and denial strategies improved from F_1 scores of 59.8% and 65.0% to 67.5% and 76.6%.

I further demonstrated that the word pairs can be applied to the automatic identification of the language of face-saving in the context of question and answering. The results also showed that regardless of differences in ideologies and framing strategies, we can detect the language of face-saving with high accuracy.

The last study of Chapter 2 examined whether we can automatically detect true, false, stretch, and dodge statements. The results show that linguistic features are not very effective in identifying false and stretch statements and that world knowledge is necessary in determining the validity of statements. Dodge statements were detected more accurately using mostly the lexical features compared to the other labels (F_1 score of 82.6% in binary classification setting).

Chapter 3 operationalized framing in news corpora as a classification task. The results showed that classification is an effective approach to analyze framing. I further demonstrated that world knowledge that is captured in distributional semantics is beneficial in modeling frames. This chapter further examined whether frames are transferable across genres. The results showed that using a corpus of user comments annotated with a set of pre-defined frames can help us identify frames at the sentence and paragraph levels in parliamentary debates; however, the analysis is restricted to the pre-defined frames and cannot identify the extra frames in the target genre (parliamentary genre in our experiment).

Chapter 4 investigated how different models of argumentation can help in understanding and evaluation of argumentation, and proposed a set of dimensions based on the existing theories for assessing arguments at the logical, rhetorical, and dialectical levels. This chapter further

presents a benchmark corpus developed using these dimensions, and showed what aspects of argumentation are difficult to assess and challenging to reach agreement on through multiple annotation studies.

This dissertation showed that relying on the characteristics of language employed in argumentation can help identify persuasive strategies with a fair level of accuracy; however, in order to improve this analysis, it is important to account for external factors, such as world knowledge and the belief systems of audience.

5.2 Future work

There are many open questions and research directions that remain to be explored. Here, I suggest just a few ways in which this work can and should be extended.

5.2.1 Analyzing face-saving approaches in other corpora and domains

The proposed approach to face-saving strategies has so far only been applied to the parliamentary genre. Another interesting corpus on which to test this approach is the corpus of Vanderbilt[1] that provides an extensive archive of television news and interviews. This analysis would allow us to investigate how reputation management in mainstream media platforms differs from reputation management in political institutions. The interviews need to be automatically transcribed. We can then investigate whether the approaches that we took in Chapter 2 can be applied to this corpus. Another potential corpus for the analysis of these strategies is court documents. Defendants' statements express admission or denial of guilt and can be leveraged for this analysis. In the legal domain, answering the question cannot be avoided and the answers can have serious consequences. Furthermore, the defendant usually focuses on self-reputation management as opposed to the party's reputation in parliamentary institutions. This results in sparse data. Such analysis allows us to investigate how effective each strategy is in court

[1] https://tvnews.vanderbilt.edu

outcomes.

5.2.2 Examining the persuasion effect of face-saving strategies

Manual analyses of case studies suggest certain face-saving strategies are more effective and persuasive than others (Koller, 1993; Benoit and Drew, 1997); however, large-scale studies are required to test the effect of rebutting skills and face-saving strategies. Politicians are at varying levels of rebutting skills, and being able to defend their government and its actions can represent good leadership.

One way of testing this persuasion effect is using the language of reputation defence to predict whether a leader of a government (e.g., the prime minister of the UK) will be in power for the next term. The results in Chapter 2 lay the groundwork for a future system which makes use of the detected face-saving strategies to determine whether rebutting skills predict staying in power. For example, the attempts by the May government to defend Brexit could be compared with the Blair government's defence of the 2003 invasion of Iraq, both in 2003 and in 2016 after the Chilcot report.[2] Each government leader's (prime minister's) answer to a question posed by an opposition member can be taken as the basic unit of text in our study. We could then use the features that we would find effective in the detection of reputation defence and train a classifier to examine the impact of face-saving strategies.

This study can help us determine whether certain face-saving strategies are more persuasive than others and whether there is any connection between face-saving strategies and the persuasiveness of arguments. One major challenge of such a study is to account for potential variables that may affect the outcome.

5.2.3 Using manifesto data for issue-specific frame classification

As discussed earlier in Section 3.2, our analysis on issue-specific framing was limited to the *gay marriage* issue. In order to expand our analysis of identifying issue-specific framing, we can use

[2] A British public inquiry into the nation's role in the Iraq war.

the Comparative Manifesto Project dataset[3]. This dataset consists of manifestos of over 1,000 parties in 50 countries. A subset of these manifestos has been annotated with 56 categories that are grouped into 7 policy areas, including *External relations, Freedom and democracy, Political system, Economy, Welfare and quality of life, Fabric of society,* and *Social groups*.

This analysis further allows us to compare framing across issues. We can then investigate the use of recent representation methods such as that of topically driven neural language models (Lau et al., 2017), which can model both the sentence and the sentence context or the document that the sentence appears in. Furthermore, we can examine the representations provided by BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), as they capture the context information effectively and have been shown to perform well for tasks that require world knowledge.

5.2.4 Combining generic and issue-specific frames for frame classification

In Chapter 3, by improving framing analysis using distributional representations, we showed that world knowledge is necessary to better capture frame meanings; however, this dissertation did not study the interaction of generic and issue-specific frames. Future work can examine how these two related tasks can contribute to each other. This interaction can be explored by recent advances in multi-task learning (Ruder, 2017) and transfer learning models (Weiss et al., 2016). In multi-task learning, the common information between related tasks that is shared during training can improve the generalization of the models (Caruana, 1997). In transfer learning, the knowledge learned in one task can be used to improve learning in a related task. Using these approaches may help in learning and interpreting the meaning of frames more effectively as some scholars such as Narvaez (2001) believe that even humans develop better judgment and understanding of argument reasoning by expanding their domain knowledge.

[3] https://manifesto-project.wzb.eu/

5.2.5 Examining ad hominem arguments or personal attacks in argumentative QAs

In Section 2.3, we briefly analyzed the differences between the questions asked by opposition members and government backbenchers. Future work can focus on a more detailed analysis of the language of reputation threat and its impact on reputation defence language. For example, whether hostile language in reputation threats can trigger personal attacks or ad hominem arguments in reputation defences. The *Wikipedia Talk Labels* datasets (Thain et al., 2017; Wulczyn et al., 2017) contains annotations regarding personal attacks and toxic language that could be useful for identifying answers, such as *if the hon. member pardons my laughing, the question he posed shows the absolute sheer idiocy that has been going on across the way* and *a stupid question does not deserve an answer.* In parliamentary sessions, when a member uses unparliamentary language, the *Speaker* interrupts the member with a comment like *Order, I urge the hon. member to be very judicious with his language* or *I encourage the hon. member to use language that is more judicious.* We could extract the speech leading to these comments and use them for training our models. One of the challenges of this study will be data sparsity. We can investigate the use of character-level networks, such as Zhang et al. (2015)'s model to help us identify the new words that are coined in parliamentary language and are intended to insult, but not explicitly,[4] such as *fuddle duddle*[5] or *terminological inexactitude.*[6] Character-level neural models, despite being slower to process than word-level neural models, help with the out-of-vocabulary problem and may identify these rare new words that are intended to insult.

[4]MPs are expected to refrain from direct insults.

[5]https://en.wikipedia.org/wiki/Fuddle_duddle

[6]https://en.wikipedia.org/wiki/Terminological_inexactitude

Appendix A

Computational Argumentation Quality Assessment in Natural Language

Computational Argumentation Quality Assessment in Natural Language

Henning Wachsmuth
Bauhaus-Universität Weimar
Weimar, Germany
henning.wachsmuth@uni-weimar.de

Nona Naderi
University of Toronto
Toronto, Canada
nona@cs.toronto.edu

Yufang Hou
IBM Research
Dublin, Ireland
yhou@ie.ibm.com

Yonatan Bilu
IBM Research
Haifa, Israel
yonatanb@il.ibm.com

Vinodkumar Prabhakaran
Stanford University
Stanford, CA, USA
vinod@cs.stanford.edu

Tim Alberdingk Thijm, Graeme Hirst
University of Toronto
Toronto, Canada
{thijm, gh}@cs.toronto.edu

Benno Stein
Bauhaus-Universität Weimar
Weimar, Germany
benno.stein@uni-weimar.de

Abstract

Research on computational argumentation faces the problem of how to automatically assess the quality of an argument or argumentation. While different quality dimensions have been approached in natural language processing, a common understanding of argumentation quality is still missing. This paper presents the first holistic work on computational argumentation quality in natural language. We comprehensively survey the diverse existing theories and approaches to assess logical, rhetorical, and dialectical quality dimensions, and we derive a systematic taxonomy from these. In addition, we provide a corpus with 320 arguments, annotated for all 15 dimensions in the taxonomy. Our results establish a common ground for research on computational argumentation quality assessment.

1 Introduction

What is a good argument? What premises should it be based on? When is argumentation persuasive? When is it reasonable? We subsume such questions under the term *argumentation quality*; they have driven logicians, rhetoricians, linguists, and argumentation theorists since the Ancient Greeks (Aristotle, 2007). Now that the area of computational argumentation is seeing an influx of research activity, the automatic assessment of argumentation quality is coming into the focus, due to its importance for envisioned applications such as writing support (Stab and Gurevych, 2014) and argument search (Wachsmuth et al., 2017), among others.

Existing research covers the mining of argument units (Al-Khatib et al., 2016), specific types of evidence (Rinott et al., 2015), and argumentative relations (Peldszus and Stede, 2015). Other works classify argumentation schemes (Feng et al., 2014) and frames (Naderi and Hirst, 2015), analyze overall argumentation structures (Wachsmuth et al., 2015), or generate claims (Bilu and Slonim, 2016). Also, theories of argumentation quality exist, and some quality dimensions have been assessed computationally (see Section 2 for details). Until now, however, the assertion of O'Keefe and Jackson (1995) that there is neither a general idea of what constitutes argumentation quality in natural language nor a clear definition of its dimensions still holds.

The reasons for this deficit originate in the varying goals of argumentation: persuading audiences, resolving disputes, achieving agreement, completing inquiries, and recommending actions (Tindale, 2007). As a result, diverse quality dimensions play a role, which relate to the logic of arguments, to the style and rhetorical effect of argumentation, or to its contribution to a discussion. Consider the following argument against the death penalty:[1]

Everyone has an inalienable human right to life, even those who commit murder; sentencing a person to death and executing them violates that right.

Although implicit, the conclusion about the death penalty seems sound in terms of (informal) logic, and the argument is clear from a linguistic viewpoint. Some people might not accept the first stated premise, though, especially if emotionally affected by some legal case at hand. Or, they might not be persuaded that the stated argument is the most relevant in the debate on death penalty.

This example reveals three central challenges: (1) Argumentation quality is assessed on different levels of granularity; (2) many quality dimensions are subjective, depending on preconceived opinions; and (3) overall argumentation quality seems hard to measure, as the impact and interaction of the different dimensions remain unclear.

[1] Taken from www.bbc.co.uk/ethics/capitalpunishment.

This paper does *not* propose a specific approach to assess quality; rather it defines a common ground by providing a so-far-missing holistic view on argumentation quality assessment in natural language. In particular, we first briefly but comprehensively survey all major theories and computational approaches for argumentation quality. Following Blair (2012), we distinguish three main quality aspects, each associated with several quality dimensions:

- *Logical quality* in terms of the cogency or strength of an argument.
- *Rhetorical quality* in terms of the persuasive effect of an argument or argumentation.
- *Dialectical quality* in terms of the reasonableness of argumentation for resolving issues.

We organize the survey along these aspects, discussing quality at four levels of granularity: (1) *argument unit*, i.e., a segment of text that takes the role of a premise or conclusion; (2) *argument*, i.e., a composition of premises and a conclusion, some of which may be implicit; (3) *(monological) argumentation*, i.e., a composition of arguments on a given issue; and (4) *(dialogical) debate*, i.e., a series of interacting argumentation on the same issue.

To unify and to consolidate existing research, we then derive a generally applicable taxonomy of argumentation quality from the survey. The taxonomy systematically decomposes quality assessment based on the interactions of 15 widely accepted quality dimensions (including the overall quality). Moreover, we provide a new annotated corpus with 320 arguments for which three experts assessed all 15 dimensions, resulting in over 14,000 annotations. Our analysis indicates how the dimensions interact and which of them are subjective, making the corpus an adequate benchmark for future research.

In summary, the contributions of this paper are:

1. *A comprehensive survey* of research on argumentation quality assessment (Section 2).
2. *A taxonomy* of all major quality dimensions of natural language argumentation, which clarifies their roles and dependencies (Section 3).
3. *An annotated corpus* for computational argumentation quality assessment (Section 4).[2]

2 Survey of Argumentation Quality

This section briefly surveys all major existing theories and the assessment of natural language argumentation quality. While we order the discussions along the three main quality aspects, we point out overlaps and interrelations where relevant.

2.1 Theories of Argumentation Quality

We focus on the major fields dealing with argumentation quality in natural language: argumentation theory and rhetoric. Table 1 gives an overview of the quality dimensions that we detail below.

Logic Formal argumentation studies the *soundness* of arguments, requiring the truth of an argument's premises and the deductive *validity* of inferring its conclusion. In case of inductive strength, the conclusion becomes probable given the premises. While sound arguments exist in natural language, most are defeasible in nature (Walton, 2006). The desired property of such arguments is *cogency*.

A cogent (or logically good) argument has individually acceptable premises that are relevant to the argument's conclusion and, together, sufficient to draw the conclusion (Johnson and Blair, 2006). Here, *(local) acceptability* means that a premise is rationally worthy of being believed by the target audience of the argument. It replaces truth, which is often unclear (Hamblin, 1970). A premise's *(local) relevance* refers to the level of support it provides for the conclusion, and *(local) sufficiency* captures whether the premises give enough reason to accept the conclusion. In the end, sufficiency thus presupposes relevance (Blair, 2012). While acceptability is more dialectical, overall the three dimensions of cogency are, with slight variations, acknowledged to cover the logical quality of arguments.

Damer (2009) adds that a good argument also depends on the rebuttal it gives to anticipated counterarguments (a dialectical property) as well as on its structural *well-formedness*, i.e., whether it is intrinsically consistent, avoids begging the question, and uses a valid inference rule. These dimensions adopt ideas from the argument model of Toulmin (1958), including rebuttals and warrants, and from the argumentation schemes of Walton et al. (2008), whose critical questions are meant to evaluate inference rules. While not focusing on quality, critical questions particularly help identify fallacies.

Introduced by Aristotle as invalid arguments, fallacies have been brought back to attention by Hamblin (1970). In general, a fallacy has some sort of error in reasoning (Tindale, 2007). Fallacies range from resorting to inapplicable evidence types or irrelevant premises to rhetoric-related errors, such

[2]The corpus is freely available at: http://www.arguana.com

Aspect	Quality Dimension	Granularity	Sources
Logic	**Cogency**	Argument	Johnson and Blair (2006), Damer (2009), Govier (2010)
	Local relevance	Argument (unit)	Johnson and Blair (2006), Damer (2009), Govier (2010)
	Local sufficiency	Argument	Johnson and Blair (2006), Damer (2009), Govier (2010)
	Well-Formedness	Argument	Walton et al. (2008), Damer (2009)
Dialectic	Global sufficiency	Argument	Toulmin (1958), Damer (2009)
Dialectic	Local acceptability	Argument (unit)	Johnson and Blair (2006), Damer (2009), Govier (2010)
	Fallaciousness	Argument (unit)	Hamblin (1970), Tindale (2007), Walton et al. (2008)
	Local relevance	Argument (unit)	Hamblin (1970), Tindale (2007)
	Local sufficiency	Argument	Hamblin (1970), Tindale (2007)
	Validity	Argument	Hamblin (1970), Tindale (2007)
	Well-Formedness	Argument	Hamblin (1970), Tindale (2007)
	Strength	Argument	Perelman et al. (1969), Tindale (2007), Freeman (2011)
Rhetoric	**Effectiveness**	Argument(ation)	Perelman et al. (1969), O'Keefe and Jackson (1995)
	Arrangement	Argumentation	Aristotle (2007), Damer (2009)
	Appropriateness of style	Argumentation	Aristotle (2007)
	Clarity of style	Argumentation	Aristotle (2007), Tindale (2007), Govier (2010)
	Credibility	Argumentation	Aristotle (2007)
	Emotional appeal	Argumentation	Aristotle (2007), Govier (2010)
Logic	Soundness	Argument	Aristotle (2007)
Dialectic	**Convincingness**	Argumentation	Perelman et al. (1969)
	Global acceptability	Argument(ation)	Perelman et al. (1969)
	Reasonableness	Argumentation, debate	van Eemeren and Grootendorst (2004)
	Global acceptability	Argument(ation)	van Eemeren and Grootendorst (2004)
	Global relevance	Argument(ation)	van Eemeren and Grootendorst (2004), Walton (2006)
	Global sufficiency	Argumentation, debate	Cohen (2001)

Table 1: Theoretical treatment of quality dimensions in the referenced sources for the given granularities of natural language argumentation, grouped by the aspect the bold-faced high-level dimensions refer to.

as unjustified appeals to emotion. They represent an alternative assessment of logical quality. Following Damer (2009), a fallacy can always be seen as a violation of one or more dimensions of good arguments. *Fallaciousness* negatively affects an argument's *strength* (Tindale, 2007).

Argument strength is often referred to, but its meaning remains unclear: "Is a strong argument an effective argument which gains the adherence of the audience, or is it a valid argument, which ought to gain it?" (Perelman et al., 1969). Tindale (2007) sees validity as a possible but not mandatory part of reasoning strength. Freeman (2011) speaks of the strength of support, matching the idea of inductive strength. Blair (2012) roughly equates strength with cogency, and Hoeken (2001) observes correlations between evidence strength and rhetorical persuasiveness. Such dependencies are expected, as the use of true and valid arguments represents one means of persuasion: logos (Aristotle, 2007).

Rhetoric Aristotle's work on rhetoric is one of the most systematic to this day. He defines rhetoric as the ability to know how to persuade (Aristotle, 2007). Besides logos, the three means of persuasion he sees include ethos, referring to the arguer's *credibility*, and pathos, the successful *emotional appeal* to the target audience. Govier (2010) outlines how emotions interfere with logic in arguments. Pathos is not necessarily reprehensible; it just aims for an emotional state adequate for persuasion.

In overall terms, rhetorical quality is reflected by the persuasive *effectiveness*, i.e., the success in persuading a target audience of a conclusion (Blair, 2012). It has been suggested that what arguments are considered as effective is subjective (O'Keefe and Jackson, 1995). Unlike persuasiveness, which relates to the actual arguments, effectiveness covers all aspects of an argumentation, including the use of language (van Eemeren, 2015). In particular, the three means of persuasion are meant to be realized by what is said and how (Aristotle, 2007). Several linguistic quality dimensions are connected to argumentation (examples follow in Section 2.2). While many of them are distinguished by Aristotle, he groups them as the *clarity* and the *appropriateness* of style as well as the proper *arrangement*.

Clarity means the use of correct, unambiguous language that avoids unnecessary complexity and deviation from the discussed issue (Aristotle, 2007). Besides ambiguity, vagueness is a major problem impairing clarity (Govier, 2010) and can be a cause of fallacies (Tindale, 2007). So, clarity is a prerequisite of logos. Also, it affects credibility, since it indicates the arguer's skills. An appropriate style in terms of the choice of words supports credibility and emotions. It is tailored to the issue and

audience (Aristotle, 2007). Arrangement, finally, addresses the structure of argumentation regarding the presentation of the issue, pros, cons, and conclusions. Damer (2009) outlines that a proper arrangement is governed by the dimensions of a good argument. To be effective, well-arranged argumentation matches the expectations of the target audience and is, thus, related to dialectic (Blair, 2012).

Dialectic The dialectical view of argumentation targets the resolution of differences of opinions on the merit (van Eemeren and Grootendorst, 2004). Quality is assessed for well-arranged discussions that seek agreement. In contrast to the subjective nature of effectiveness, people are good in such an assessment (Mercier and Sperber, 2011). In their pragma-dialectical theory, van Eemeren and Grootendorst (2004) develop rules for obtaining *reasonableness* in critical discussions. Reasonableness emerges from two complementary dimensions, intersubjective *(global) acceptability* and problem-solving validity, but effectiveness still remains the underlying goal (van Eemeren, 2015). For argumentation, global acceptability is given when the stated arguments and the way they are stated are acceptable to the whole target audience. Problem-solving validity matches the *(global) relevance* of argumentation that contributes to resolution, helping arrive at an ultimate conclusion (Walton, 2006).

Global relevance implicitly excludes fallacious moves, so reasonable arguments are cogent (van Eemeren, 2015). Van Eemeren sees reasonableness as a precondition for *convincingness*, the rational version of persuasiveness. Following Perelman et al. (1969), persuasive argumentation aims at a particular audience, whereas convincing argumentation aims at the universal audience, i.e., all reasonable beings. This fits the notion that dialectic examines general rather than specific issues (Aristotle, 2007).

Convincingness needs *(global) sufficiency*, i.e., all objections to an argumentation are countered. The dilemma here is that the number of objections could be infinite, but without global sufficiency the required support seems arbitrary (Blair, 2012). A solution is the relaxed view of Damer (2009) that only those counter-arguments that can be anticipated are to be rebutted. For debates, Cohen (2001) speaks of dialectical satisfactoriness, i.e., whether all questions and objections have been sufficiently answered. In case a reasonable debate ends up in either form of global sufficiency, this implies that the discussed difference of opinion is resolved.

Other Although closely related, critical thinking (Freeley and Steinberg, 2009) and persuasion research (Zhao et al., 2011) are covered only implicitly here; their views on quality largely match with argumentation theory. We have not discussed deliberation, as it is not concerned with the quality of argumentation primarily but rather with communicative dimensions of group decision-making, e.g., participation and respect (Steenbergen et al., 2003). Also, we have restricted our view to the logic found in natural language. For formal and probabilistic logic, dimensions such as degree of justification (Pollock, 2001), argument strength (Pfeifer, 2013), and premise relevance (Ransom et al., 2015) have been analyzed. As we see below, such logic influenced some practical assessment approaches.

2.2 Approaches to Quality Assessment

As for the theories, we survey the automatic quality assessment for natural language argumentation. All discussed approaches are listed in Table 2.

Logic Braunstain et al. (2016) deal with logical argument quality in community question answering: Combining relevance-oriented retrieval models and argument-oriented features, they rank sentence-level argument units according to the *level of support* they provide for an answer. Unlike classical essay scoring, Rahimi et al. (2014) score an essay's *evidence*, a quality dimension of argumentation: it captures how sufficiently the given details support the essay's thesis. On the dataset of Correnti et al. (2013) with 1569 student essays and scores from 1 to 4, they find that the concentration and specificity of words related to the essay prompt (i.e., the statement defining the discussed issue) impacts scoring accuracy. Similarly, Stab and Gurevych (2017) introduce an essay corpus with 1029 argument-level annotations of *sufficiency*, following the definition of Johnson and Blair (2006). Their experiments suggest that convolutional neural networks outperform feature-based sufficiency classification.

Rhetoric Persing et al. (2010) tackle the proper arrangement of an essay, namely, its *organization* in terms of the logical development of an argument. The authors rely on manual 7-point score annotations for 1003 essays from the ICLE corpus (Granger et al., 2009). In their experiments, sequences of paragraph discourse functions (e.g., introduction or rebuttal) turn out to be most effective. Organization is also analyzed by Rahimi et al. (2015) on the same dataset used for the evidence

Aspect	Quality Dimension	Granularity	Text Genres	Sources
Logic	Evidence	Argumentation	Student essays	Rahimi et al. (2014)
	Level of support	Argument unit	Wikipedia articles	Braunstain et al. (2016)
	Sufficiency	Argument	Student essays	Stab and Gurevych (2017)
Rhetoric	Argument strength	Argumentation	Student essays	Persing and Ng (2015)
	Evaluability	Argumentation	Law comments	Park et al. (2015)
	Global coherence	Argumentation	Student essays	Feng et al. (2014)
	Organization	Argumentation	Student essays	Persing et al. (2010), Rahimi et al. (2015)
	Persuasiveness	Argument	Forum discussions	Tan et al. (2016), Wei et al. (2016)
	Prompt adherence	Argumentation	Student essays	Persing and Ng (2014)
	Thesis clarity	Argumentation	Student essays	Persing and Ng (2013)
	Winning side	Debate	Oxford-style debates	Zhang et al. (2016)
Dialectic	Acceptability	Argument	Debate portal arguments	Cabrio and Villata (2012)
	Convincingness	Argument	Debate portal arguments	Habernal and Gurevych (2016)
	Prominence	Argument	Forum discussions	Boltužić and Šnajder (2015)
	Relevance	Argument	Diverse genres	Wachsmuth et al. (2017)

Table 2: Practical assessment of quality dimensions in the referenced sources for the given granularities and text genres of natural language argumentation, grouped by the aspect the quality dimensions refer to.

approach above. Their results indicate a correlation between organization and local coherence. Feng et al. (2014) parse discourse structure to assess *global coherence*, i.e., the continuity of meaning in a text. Lacking ground-truth coherence labels, they evaluate their approach on sentence ordering and organization scoring instead. Coherence affects the clarity of style, as do the *thesis clarity* and *prompt adherence* of essays. Persing and Ng (2013) find the former to suffer from misspellings, while Persing and Ng (2014) use prompt-related keywords and topic models to capture the latter (both for 830 ICLE essays like those mentioned above). For comments in lawmaking, Park et al. (2015) develop an argumentation model that prescribes what information users should give to achieve *evaluability* (e.g., testimony evidence or references to resources).

Not only linguistic quality, but also effectiveness is assessed in recent work: Persing and Ng (2015) score the *argument strength* of essays, which they define rhetorically in terms of how many readers would be persuaded. Although potentially subjective, their manual 7-point score annotations of 1000 ICLE essays differ by at most 1 in 67% of the studied cases. Their best features are heuristic argument unit labels and part-of-speech n-grams. Recently, Wachsmuth et al. (2016) demonstrated that the output of argument mining helps in such argumentation-related essay scoring, obtaining better results for argument strength and organization. Tan et al. (2016) analyze which arguments achieve *persuasiveness* in "change my view" forum discussions, showing that multiple interactions with the view-holder are beneficial as well as an appropriate style and a high number of participants. On similar data, Wei et al. (2016) find that also an author's reputation impacts persuasiveness. Zhang et al. (2016) discover for Oxford-style debates that attacking the opponents' arguments tends to be more effective than relying on one's own arguments. These results indicate the relation of rhetoric and dialectic.

Dialectic Dialectical quality has been addressed by Cabrio and Villata (2012). The authors use textual entailment to find ground-truth debate portal arguments that attack others. Based on the formal argumentation framework of Dung (1995), they then assess global argument *acceptability*. Habernal and Gurevych (2016) compare arguments in terms of *convincingness*. However, the subjective nature of their crowdsourced labels actually reflects rhetorical effectiveness. Boltužić and Šnajder (2015) present first steps towards argument *prominence*. Prominence may be a product of popularity, though, making its quality nature questionable, as popularity is often not correlated with merit (Govier, 2010). In contrast, Wachsmuth et al. (2017) adapt the famous PageRank algorithm to objectively derive the *relevance* of an argument at web scale from what other arguments refer to the argument's premises. On a large ground-truth argument graph, their approach beats several baselines for the benchmark argument rankings that they provide.

Other Again, we have left out deliberative quality (Gold et al., 2015). Also, we omit approaches that classify argumentation schemes (Feng and Hirst, 2011), evidence types (Rinott et al., 2015), ethos-related statements (Duthie et al., 2016), and myside bias (Stab and Gurevych, 2016); their output may help assess quality assessment, but they do not actually assess it. The same holds for argument mining,

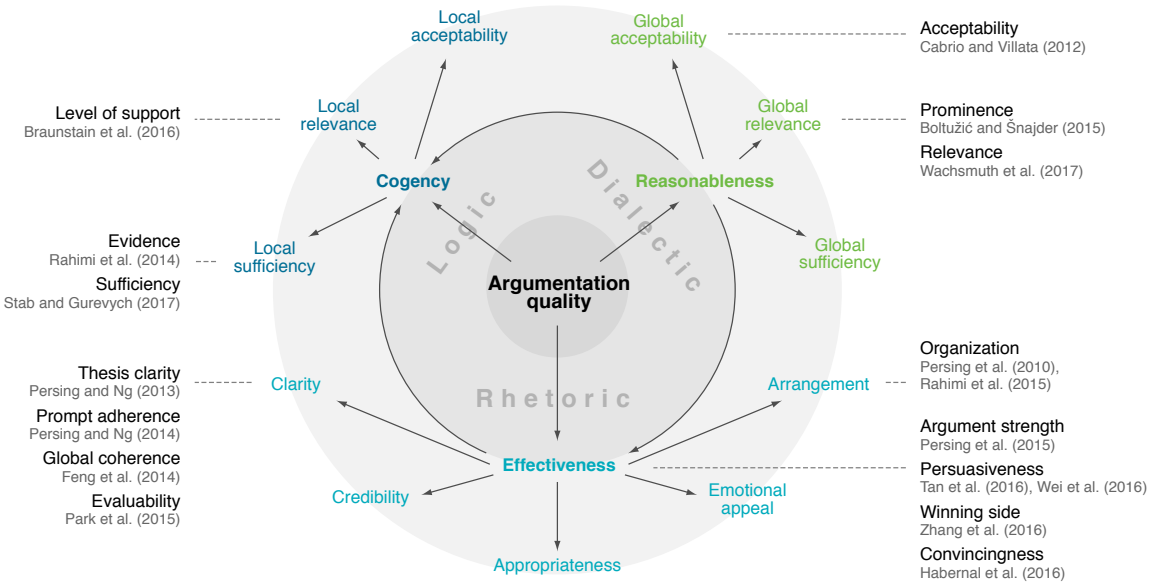

Figure 1: The proposed taxonomy of argumentation quality as well as the mapping of existing assessment approaches to the covered quality dimensions. Arrows show main dependencies between the dimensions.

even if said to aim for argument quality (Swanson et al., 2015). Much work exists for general text quality, most notably in the context of readability (Pitler and Nenkova, 2008) and classical essay scoring. Some scoring approaches derive features from discourse (Burstein et al., 1998), arguments (Ong et al., 2014; Beigman Klebanov et al., 2016; Ghosh et al., 2016), or schemes (Song et al., 2014)—all this may be indicative of quality. However, our focus is approaches that target argumentation quality at heart. Similarly, review helpfulness (Liu et al., 2008) and deception (Ott et al., 2011) are not treated, as arguments only partly play a role there. Also, only few Wikipedia quality flaws relate to arguments, e.g., verifiability (Anderka et al., 2012).

3 A Taxonomy of Argumentation Quality

Given all surveyed quality dimensions, we now propose a unifying taxonomy of argumentation quality. The taxonomy decomposes quality assessment systematically, thus organizing and clarifying the roles of practical approaches. It does not require a particular argumentation model, but it rests on the notion of the granularity levels from Section 1.

3.1 Overview of the Theory-based Taxonomy

Our objective is not to come up with a new theory, but to provide a unified view of existing theories that is suitable for quality assessment. We aim for a common understanding of the dimensions that affect quality, what interdependencies they have, and how they interact. Figure 1 illustrates the taxonomy that we propose for this purpose. The rationale behind its structure and its layout is as follows.

While Section 2 has outlined overlaps and relations between the three aspects of argumentation, we have identified one dominant high-level quality dimension of *argumentation quality* in theory for each aspect: logical *cogency*, rhetorical *effectiveness*, and dialectical *reasonableness*. The latter two benefit from cogency, and reasonableness depends on effectiveness, as discussed. Often, only one of them will be in the focus of attention in practice, or even only a sub-dimension. In particular, each high-level dimension has a set of sub-dimensions agreed upon. The sub-dimensions are shown on the outer ring in Figure 1, roughly positioned according to the aspects they refer to, e.g., *local acceptability* lies next to the other dialectical dimensions. We ordered the sub-dimensions by their interrelations (left implicit for conciseness), e.g., *appropriateness* supports *credibility* and *emotional appeal*.

Slightly deviating from theory, we match Aristotle's logos dimension with cogency, which better fits real-world argumentation. Similarly, we omit those dimensions from Table 1 in the taxonomy that have unclear definitions, such as strength, or that are covered by others, such as well-formedness, which merely refines the acceptability part of cogency (Govier, 2010). Convincingness is left out,

as it is close to effectiveness and as both the feasibility and the need of persuading the universal audience has been questioned (van Eemeren, 2015). Instead, we add *global sufficiency* as part of reasonableness. While global sufficiency may be infeasible, too (Blair, 2012), it forces agreement in critical discussions and, thereby, reasonableness.

3.2 Definitions of the Quality Dimensions

Cogency is seen as an argument property, whereas effectiveness and reasonableness are assessed on the argumentation level usually. For generality, we give informal literature-based definitions of these dimensions and all sub-dimensions here for an author who argues about an issue to a target audience:

Cogency An argument is cogent if it has acceptable premises that are relevant to its conclusion and that are sufficient to draw the conclusion.

- *Local acceptability:* A premise of an argument is acceptable if it is rationally worthy of being believed to be true.
- *Local relevance:* A premise of an argument is relevant if it contributes to the acceptance or rejection of the argument's conclusion.
- *Local sufficiency:* An argument's premises are sufficient if, together, they give enough support to make it rational to draw its conclusion.

Effectiveness Argumentation is effective if it persuades the target audience of (or corroborates agreement with) the author's stance on the issue.

- *Credibility:* Argumentation creates credibility if it conveys arguments and similar in a way that makes the author worthy of credence.
- *Emotional Appeal:* Argumentation makes a successful emotional appeal if it creates emotions in a way that makes the target audience more open to the author's arguments.
- *Clarity:* Argumentation has a clear style if it uses correct and widely unambiguous language as well as if it avoids unnecessary complexity and deviation from the issue.
- *Appropriateness:* Argumentation has an appropriate style if the used language supports the creation of credibility and emotions as well as if it is proportional to the issue.
- *Arrangement:* Argumentation is arranged properly if it presents the issue, the arguments, and its conclusion in the right order.

Reasonableness Argumentation is reasonable if it contributes to the issue's resolution in a sufficient way that is acceptable to the target audience.

- *Global acceptability:* Argumentation is acceptable if the target audience accepts both the consideration of the stated arguments for the issue and the way they are stated.
- *Global relevance:* Argumentation is relevant if it contributes to the issue's resolution, i.e., if it states arguments or other information that help to arrive at an ultimate conclusion.
- *Global sufficiency:* Argumentation is sufficient if it adequately rebuts those counter-arguments to it that can be anticipated.

3.3 Organization of Assessment Approaches

The taxonomy is meant to define a common ground for assessing argumentation quality, including the organization of practical approaches. The left and right side of Figure 1 show where the approaches surveyed in Section 2.2 are positioned in the taxonomy. Some dimensions have been tackled multiple times (e.g., *clarity*), others not at all (e.g., *credibility*). The taxonomy indicates what sub-dimensions will affect the same high-level dimension.

4 The Dagstuhl-15512 ArgQuality Corpus

Finally, we present our new annotated *Dagstuhl-15512 ArgQuality Corpus* for studying argumentation quality based on the developed taxonomy, and we report on a first corpus analysis.[3]

4.1 Data and Annotation Process

Our corpus is based on the *UKPConvArgRank* dataset (Habernal and Gurevych, 2016), which contains rankings of 25 to 35 textual debate portal arguments for two stances on 16 issues, such as *evolution vs. creation* and *ban plastic water bottles*. All ranks were derived from crowdsourced convincingness labels. For every issue/stance pair, we took the five top-ranked texts and chose five further via stratified sampling. Thereby, we covered both high-quality arguments and different levels of lower quality. Two example texts follow below in Figure 2.

Before annotating the 320 chosen texts, we carried out a full annotation study with seven authors of this paper on 20 argumentative comments from

[3]The corpus and annotation guidelines are available at http://www.arguana.com. The corpus is named after the Dagstuhl Seminar 15512 "Debating Technologies" that initialized the research in this paper: http://www.dagstuhl.de/15512

	Quality Dimension	(a) Maj. Scores 1	2	3	(b) Agreement α	full	maj.	(c) Pearson Correlation Coefficients Co	LA	LR	LS	Ef	Cr	Em	Cl	Ap	Ar	Re	GA	GR	GS
Co	**Cogency**	150	131	23	.44	40.1%	91.8%		.64	.61	**.84**	**.81**	.46	.27	.41	.32	.55	.78	.64	**.71**	.70
LA	Local acceptability	84	169	51	.46	27.0%	90.8%	.64		.51	.53	.60	**.54**	.30	.40	.54	.46	.68	.75	.46	.45
LR	Local relevance	25	155	**124**	.47	32.6%	92.4%	.61	.51		.56	.56	.39	.27	.46	.35	.50	.62	.58	.68	.45
LS	Local sufficiency	172	119	13	.44	37.2%	92.8%	**.84**	.53	.56		.73	.39	.25	.37	.23	.51	.67	.51	.68	**.74**
Ef	**Effectiveness**	184	111	9	.45	42.1%	94.4%	.81	.60	.56	.73		.48	.31	.35	.34	.54	.75	.58	.66	.71
Cr	Credibility	99	199	6	.37	37.8%	95.7%	.46	.54	.39	.39	.48		**.37**	.32	.49	.37	.52	.52	.36	.40
Em	Emotional appeal	48	**235**	21	.26	42.8%	94.4%	.27	.30	.27	.25	.31	.37		.14	.30	.20	.30	.26	.26	.22
Cl	Clarity	42	191	71	.35	29.3%	89.8%	.41	.40	.46	.37	.35	.32	.14		.45	.56	.44	.45	.38	.27
Ap	Appropriateness	43	196	65	.36	17.4%	87.5%	.32	.54	.35	.23	.34	.49	.30	.45		.48	.47	.59	.20	.20
Ar	Arrangement	91	189	24	.39	26.6%	93.4%	.55	.46	.50	.51	.54	.37	.20	**.56**	.48		.55	.51	.49	.48
Re	**Reasonableness**	126	159	19	.50	41.4%	95.7%	.78	.68	.62	.67	.75	.52	.30	.44	.47	.55		**.78**	.65	.61
GA	Global acceptability	88	161	55	.44	31.6%	95.4%	.64	**.75**	.58	.51	.58	.52	.26	.45	**.59**	.51	.78		.46	.43
GR	Global relevance	69	167	68	.42	21.7%	90.1%	.71	.46	**.68**	.68	.66	.36	.26	.38	.20	.49	.65	.46		.61
GS	Global sufficiency	**231**	72	1	.27	**44.7%**	**98.0%**	.70	.45	.45	.74	.71	.40	.22	.27	.20	.48	.61	.43	.61	
Ov	**Overall quality**	152	128	24	**.51**	44.1%	94.4%	**.84**	.66	.61	.74	**.81**	.52	.30	.45	.42	**.59**	**.86**	.71	.70	.68

Table 3: Results for the 304 corpus texts classified as argumentative by all annotators: (a) Distribution of majority scores for each dimension (2 used in case of full disagreement). (b) Krippendorff's α of the most agreeing annotator pair and full/majority agreement of all annotators. (c) Correlation for each dimension pair, averaged over the correlations of all annotators. The highest value in each column is marked bold.

the unshared task dataset of the 3rd Workshop on Argument Mining.[4] The annotators assessed all 15 quality dimensions in the taxonomy for each comment (including its overall quality). Due to simple initial guidelines based on the definitions from Section 3 and the subjectiveness of the task, the agreement of all seven annotators was low for all dimensions, namely, at most .22 in terms of Krippendorff's α. The three most agreeing annotators for each dimension achieved much higher α-values between .23 (clarity) and .60 (credibility), though.[5]

The study results were discussed by all annotators, leading to a considerably refined version of the guidelines. We then selected three annotators for the corpus annotation based on their availability. They work at two universities and one company in three countries (two females, one male; two PhDs, one PhD student). For each text in the corpus, all annotators first classified whether it was actually argumentative. If so, they assessed all dimensions using ordinal scores from 1 (low) to 3 (high).[6] Additionally, "cannot judge" could be chosen.

4.2 Corpus Distribution and Agreement

Table 3(a) lists the majority scores of each dimension for the 304 corpus texts (95%) that are classified as argumentative by all annotators, all covering the whole score range. Five dimensions have the median at score 1, the others at 2. Some seem easier to master, such as *local relevance*, which received the highest majority score 124 times. Others rarely got score 3, above all *global sufficiency*. The latter is explained by the fact that only few texts include any rebuttal of counter-arguments.

Only one of the over 14,000 assessments made by the three annotators was "cannot judge" (for *global relevance*), suggesting that our guidelines were comprehensive. Regarding agreement, we see in Table 3(b) that the α-values of all logical and dialectical quality dimensions except for *global sufficiency* lie above 0.4 for the most agreeing annotator pair. As expected, the rhetorical dimensions seem to be more subjective. The lowest α is observed for *emotional appeal* (0.26). The annotators most agreed on the *overall quality* ($\alpha = 0.51$), possibly meaning that the taxonomy adequately guides the assessment. In accordance with the moderate α-values, full agreement ranges between 17.4% and 44.7% only. On the contrary, we observe high majority agreement between 87.5% and 98% for all dimensions, even where scores are rather evenly distributed, such as for *global acceptability* (95.4%). In case of full disagreement, it makes sense to use score 2. We hence argue that the corpus is suitable for evaluating argumentation quality assessment.

Figure 2 shows all scores of each annotator for two example arguments from the corpus, referring to the question whether to ban plastic water bottles. Both have majority score 3 for *overall quality (Ov)*,

[4]Unshared task data found at: http://github.com/UKPLab

[5]We use Krippendorff's α as is suitable for small samples, multiple ratings, and ordinal scales (Krippendorff, 2007).

[6]We chose a 3-point scale to foster clear decisions on the quality; in the annotation study, we used a 4-point scale but observed that the annotators only rarely chose score 1 and 4.

Arguments	Pro Water bottles, good or bad? Many people believe plastic water bottles to be good. But the truth is water bottles are polluting land and unnecessary. Plastic water bottles should only be used in emergency purposes only. The water in those plastic are only filtered tap water. In an emergency situation like Katrina no one had access to tap water. In a situation like this water bottles are good because it provides the people in need. Other than that water bottles should not be legal because it pollutes the land and big companies get 1000% of the profit.															Con Americans spend billions on bottled water every year. Banning their sale would greatly hurt an already struggling economy. In addition to the actual sale of water bottles, the plastics that they are made out of, and the advertising on both the bottles and packaging are also big business. In addition to this, compostable waters bottle are also coming onto the market, these can be used instead of plastics to eliminate that detriment. Moreover, bottled water not only has a cleaner safety record than municipal water, but it easier to trace when a potential health risk does occur. (http://www.friendsjournal.org/bottled-water) (http://www.cdc.gov/healthywater/drinking/bottled/)														
Scores	**Co**	**LA**	**LR**	**LS**	**Ef**	**Cr**	**Em**	**Cl**	**Ap**	**Ar**	**Re**	**GA**	**GR**	**GS**	**Ov**	**Co**	**LA**	**LR**	**LS**	**Ef**	**Cr**	**Em**	**Cl**	**Ap**	**Ar**	**Re**	**GA**	**GR**	**GS**	**Ov**
Annotator A	3	3	3	2	3	3	3	3	3	3	3	3	3	2	3	3	3	3	3	3	3	2	3	3	3	3	3	3	3	3
Annotator B	2	2	3	2	1	2	2	2	2	1	2	2	2	1	2	2	3	3	2	2	3	2	3	3	2	3	3	2	2	3
Annotator C	2	3	3	2	2	2	2	3	3	3	3	3	3	2	3	3	3	3	3	3	2	1	3	3	3	3	3	3	3	3
Majority score	2	3	3	2	2	2	2	3	3	3	3	3	3	2	3	3	3	3	3	3	3	2	3	3	3	3	3	3	3	3

Figure 2: The scores of each annotator and the majority score for all considered quality dimensions of one pro and one con argument from our corpus. The arguments refer to the issue *ban plastic water bottles*.

but the pro argument shows more controversy with full disagreement in case of *effectiveness (Ef)*. Especially, *annotator B* seems to be critical, giving one point less for several dimensions. In contrast, the con argument yields majority agreement for all 15 dimensions and full agreement for seven of them. It meets main quality criteria surveyed in Section 2, such as a rebuttal or references to resources. In fact, it constitutes the only corpus text with majority score 3 for *global sufficiency (GS)*.

4.3 Correlations between Quality Dimensions

Table 3(c) compares the correlations of all dimension pairs. *Cogency* (.84), *effectiveness* (.81), and *reasonableness* (.86) correlate strongly with *overall quality*, and also much with each other.

Cogency and *local sufficiency* (.84) go hand in hand, whereas *local acceptability* and *local relevance* show the highest correlation with their global counterparts (.75 and .68 respectively). Quite intuitively, *credibility* and *appropriateness* correlate most with the acceptability dimensions. The coefficients of *emotional appeal* seem lower than expected, in particular for effectiveness (.31), indicating the limitation of a correlation analysis: As reflected by the 235 texts with majority score 2 for emotional appeal, many arguments make no use of emotions, thus obliterating effects of those which do. On the other hand, *clarity* was scored 2 in most cases, too, so the very low value there (.14) is more meaningful. Clarity rather correlates with *arrangement* (.56), which in turn shows coefficients above .50 for all high-level dimensions.

Altogether, the correlations largely match the surveyed theory. While an analysis of cause and effect should follow in future work, they provide first evidence for the adequacy of our taxonomy.

5 Conclusion

Argumentation quality is of high importance for argument mining, debating technologies, and similar. In computational linguistics, it has been treated only rudimentarily so far. This paper defines a common ground for the automatic assessment of argumentation quality in natural language. Based on a survey of existing theories and approaches, we have developed a taxonomy that unifies all major dimensions of logical, and dialectical argumentation quality. In addition, we freely provide an annotated corpus for studying these dimensions.

The taxonomy is meant to capture *all* aspects of argumentation quality, irrespective of how they can be operationalized. The varying inter-annotator agreement we obtained suggests that some quality dimensions are particularly subjective, raising the need to model the target audience of an argumentation. Still, the observed correlations between the dimensions support the general adequacy of our taxonomy. Moreover, most dimensions have already been approached on a certain abstraction level in previous work, as outlined. While some refinement may be suitable to meet all requirements of the community, we thus propose the taxonomy as the common ground for future research on computational argumentation quality assessment and the corpus as a first benchmark dataset for this purpose.

Acknowledgments

We thank all attendees of Dagstuhl Seminar 15512, particularly the rest of the quality breakout group: Wolf-Tilo Balke, Ruty Rinott, and Christian Stab. Also, we acknowledge financial support of the Stanford University, the DFG, and the Natural Sciences and Engineering Research Council of Canada.

References

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1395–1404. Association for Computational Linguistics.

Maik Anderka, Benno Stein, and Nedim Lipka. 2012. Predicting quality flaws in user-generated content: The case of Wikipedia. In *Proceedings of the 35th International ACM Conference on Research and Development in Information Retrieval*, pages 981–990.

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse* (George A. Kennedy, Translator). Clarendon Aristotle series. Oxford University Press.

Beata Beigman Klebanov, Christian Stab, Jill Burstein, Yi Song, Binod Gyawali, and Iryna Gurevych. 2016. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 70–75. Association for Computational Linguistics.

Yonatan Bilu and Noam Slonim. 2016. Claim synthesis via predicate recycling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 525–530. Association for Computational Linguistics.

J. Anthony Blair. 2012. *Groundwork in the Theory of Argumentation*. Springer Netherlands.

Filip Boltužić and Jan Šnajder. 2015. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115. Association for Computational Linguistics.

Liora Braunstain, Oren Kurland, David Carmel, Idan Szpektor, and Anna Shtok. 2016. Supporting human answers for advice-seeking questions in CQA sites. In *Proceedings of the 38th European Conference on IR Research*, pages 129–141.

Jill Burstein, Karen Kukich, Susanne Wolff, Chi Lu, and Martin Chodorow. 1998. Enriching automated essay scoring using discourse marking. In *Discourse Relations and Discourse Markers*.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 208–212. Association for Computational Linguistics.

Daniel H. Cohen. 2001. Evaluating arguments and making meta-arguments. *Informal Logic*, 21(2):73–84.

Richard Correnti, Lindsay Clare Matsumura, Laura Hamilton, and Elaine Wang. 2013. Assessing students' skills at writing analytically in response to texts. *The Elementary School Journal*, 114(2):142–177.

T. Edward Damer. 2009. *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments*. Wadsworth, Cengage Learning, Belmont, CA, 6th edition.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

Rory Duthie, Katarzyna Budynska, and Chris Reed. 2016. Mining ethos in political debate. In *Proceedings of the Sixth International Conference on Computational Models of Argument*, pages 299–310.

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

Vanessa Wei Feng, Ziheng Lin, and Graeme Hirst. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 940–949. Dublin City University and Association for Computational Linguistics.

Austin J. Freeley and David L. Steinberg. 2009. *Argumentation and Debate*. Cengage Learning, Boston, MA, 12th edition.

James B. Freeman. 2011. *Argument Structure: Representation and Theory*. Springer.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. 2016. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554. Association for Computational Linguistics.

Valentin Gold, Mennatallah El-Assady, Tina Bögel, Christian Rohrdantz, Miriam Butt, Katharina Holzinger, and Daniel Keim. 2015. Visual linguistic analysis of political discussions: Measuring deliberative quality. *Digital Scholarship in the Humanities*.

Trudy Govier. 2010. *A Practical Study of Argument*. Wadsworth, Cengage Learning, Belmont, CA, 7th edition.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, and Magali Paquot. 2009. International corpus of learner English (version 2).

Ivan Habernal and Iryna Gurevych. 2016. 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. Association for Computational Linguistics.

Charles L. Hamblin. 1970. *Fallacies*. Methuen, London, UK.

Hans Hoeken. 2001. Anecdotal, statistical, and causal evidence: Their perceived and actual persuasiveness. *Argumentation*, 15(4):425–437.

Ralph H. Johnson and J. Anthony Blair. 2006. *Logical Self-defense*. International Debate Education Association.

Klaus Krippendorff. 2007. Computing Krippendorff's alpha reliability. Technical report, Univ. of Pennsylvania, Annenberg School for Communication.

Yang Liu, Xiangji Huang, Aijun An, and Xiaohui Yu. 2008. Modeling and predicting the helpfulness of online reviews. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, pages 443–452.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, 34:57–111.

Nona Naderi and Graeme Hirst. 2015. Argumentation mining in parliamentary discourse. In *Principles and Practice of Multi-Agent Systems - International Workshops: IWEC 2014, Gold Coast, QLD, Australia, December 1-5, 2014, and CMNA XV and IWEC 2015, Bertinoro, Italy, October 26, 2015, Revised Selected Papers*, pages 16–25.

Daniel J. O'Keefe and Sally Jackson. 1995. Argument quality and persuasive effects: A review of current approaches. In *Argumentation and Values: Proceedings of the Ninth Alta Conference on Argumentation*, pages 88–92.

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. Association for Computational Linguistics.

Myle Ott, Yejin Choi, Claire Cardie, and T. Jeffrey Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 309–319. Association for Computational Linguistics.

Joonsuk Park, Cheryl Blake, and Claire Cardie. 2015. Toward machine-assisted participation in eRulemaking: An argumentation model of evaluability. In *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, pages 206–210.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 938–948. Association for Computational Linguistics.

Chaïm Perelman, Lucie Olbrechts-Tyteca, John Wilkinson, and Purcell Weaver. 1969. *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press, Notre Dame, IN.

Isaac Persing and Vincent Ng. 2013. Modeling thesis clarity in student essays. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 260–269. Association for Computational Linguistics.

Isaac Persing and Vincent Ng. 2014. Modeling prompt adherence in student essays. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1534–1543. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Isaac Persing, Alan Davis, and Vincent Ng. 2010. Modeling organization in student essays. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 229–239. Association for Computational Linguistics.

Niki Pfeifer, 2013. *Bayesian Argumentation: The Practical Side of Probability*, chapter On Argument Strength, pages 185–193. Springer Netherlands, Dordrecht.

Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 186–195. Association for Computational Linguistics.

John L. Pollock. 2001. Defeasible reasoning with variable degrees of justification. *Artificial Intelligence*, 133(1–2):233–282.

Zahra Rahimi, Diane J. Litman, Richard Correnti, Lindsay Clare Matsumura, Elaine Wang, and Zahid Kisa. 2014. Automatic scoring of an analytical response-to-text assessment. In *Proceedings of the 12th International Conference on Intelligent Tutoring Systems*, pages 601–610.

Zahra Rahimi, Diane Litman, Elaine Wang, and Richard Correnti. 2015. Incorporating coherence of topics as a criterion in automatic response-to-text

assessment of the organization of writing. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 20–30. Association for Computational Linguistics.

Keith J. Ransom, Amy Perfors, and Daniel J. Navarro. 2015. Leaping to conclusions: Why premise relevance affects argument strength. *Cognitive Science*, pages 1–22.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh 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. Association for Computational Linguistics.

Yi Song, Michael Heilman, Beata Beigman Klebanov, and Paul Deane. 2014. Applying argumentation schemes for essay scoring. In *Proceedings of the First Workshop on Argumentation Mining*, pages 69–78. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2016. Recognizing the absence of opposing arguments in persuasive essays. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 113–118. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2017. Recognizing insufficiently supported arguments in argumentative essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Marco R. Steenbergen, Andre Bachtiger, Markus Sporndli, and Jürg Steiner. 2003. Measuring political deliberation: A discourse quality index. *Comparative European Politics*, 1:21–48.

Reid Swanson, Brian Ecker, and Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226. Association for Computational Linguistics.

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 World Wide Web Conference*, pages 613–624.

Christopher W. Tindale. 2007. *Fallacies and Argument Appraisal. Critical Reasoning and Argumentation*. Cambridge University Press.

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

Frans H. van Eemeren and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge University Press, Cambridge, UK.

Frans H. van Eemeren. 2015. *Reasonableness and Effectiveness in Argumentative Discourse: Fifty Contributions to the Development of Pragma-Dialectics*. Argumentation Library. Springer International Publishing.

Henning Wachsmuth, Johannes Kiesel, and Benno Stein. 2015. Sentiment flow — A general model of web review argumentation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 601–611. Association for Computational Linguistics.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691. The COLING 2016 Organizing Committee.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

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

Douglas Walton. 2006. *Fundamentals of Critical Argumentation*. Cambridge University Press.

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? Ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200. 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. Association for Computational Linguistics.

Xiaoquan Zhao, Andrew Strasser, Joseph N. Cappella, Caryn Lerman, and Martin Fishbein. 2011. A measure of perceived argument strength: Reliability and validity. *Communication Methods and Measures*, 5(1):48–75.

Appendix B

Argumentation Quality Assessment: Theory vs. Practice

Argumentation Quality Assessment: Theory vs. Practice

Henning Wachsmuth * **Nona Naderi** ** **Ivan Habernal** *** **Yufang Hou** ****
Graeme Hirst ** **Iryna Gurevych** *** **Benno Stein** *
* Bauhaus-Universität Weimar, Weimar, Germany, www.webis.de
** University of Toronto, Toronto, Canada, www.cs.toronto.edu/compling
*** Technische Universität Darmstadt, Darmstadt, Germany, www.ukp.tu-darmstadt.de
**** IBM Research, Dublin, Ireland, ie.ibm.com

Abstract

Argumentation quality is viewed differently in argumentation theory and in practical assessment approaches. This paper studies to what extent the views match empirically. We find that most observations on quality phrased spontaneously are in fact adequately represented by theory. Even more, relative comparisons of arguments in practice correlate with absolute quality ratings based on theory. Our results clarify how the two views can learn from each other.

1 Introduction

The assessment of argumentation quality is critical for any application built upon argument mining, such as debating technologies (Rinott et al., 2015). However, research still disagrees on whether quality should be assessed from a theoretical or from a practical viewpoint (Allwood, 2016).

Theory states, among other things, that a cogent argument has acceptable premises that are relevant to its conclusion and sufficient to draw the conclusion (Johnson and Blair, 2006). Practitioners object that such quality dimensions are hard to assess for real-life arguments (Habernal and Gurevych, 2016b). Moreover, the normative nature of theory suggests *absolute* quality ratings, but in practice it seems much easier to state which argument is more convincing—a *relative* assessment. Consider two debate-portal arguments for "advancing the common good is better than personal pursuit", taken from the corpora analyzed later in this paper:

Argument A *"While striving to make advancements for the common good you can change the world forever. Allot of people have succeded in doing so. Our founding fathers, Thomas Edison, George Washington, Martin Luther King jr, and many more. These people made huge advances for the common good and they are honored for it."*

Argument B *"I think the common good is a better endeavor, because it's better to give then to receive. It's better to give other people you're hand out in help then you holding your own hand."*

In the study of Habernal and Gurevych (2016b), annotators assessed Argument *A* as more convincing than *B*. When giving reasons for their assessment, though, they saw *A* as more credible and well thought through; that does not seem to be too far from the theoretical notion of cogency.

This paper gives empirical answers to the question of how different the theoretical and practical views of argumentation quality actually are. Section 2 briefly reviews existing theories and practical approaches. Section 3 then empirically analyzes correlations in two recent argument corpora, one annotated for 15 well-defined quality dimensions taken from theory (Wachsmuth et al., 2017a) and one with 17 reasons for quality differences phrased spontaneously in practice (Habernal and Gurevych, 2016a). In a crowdsourcing study, we test whether lay annotators achieve agreement on the theoretical quality dimensions (Section 4).

We find that assessments of overall argumentation quality largely match in theory and practice. Nearly all phrased reasons are adequately represented in theory. However, some theoretical quality dimensions seem hard to separate in practice. Most importantly, we provide evidence that the observed relative quality differences are reflected in absolute quality ratings. Still, our study underpins the fact that the theory-based argumentation quality assessment remains complex. Our results do not generally answer the question of what view of argumentation quality is preferable, but they clarify where theory can learn from practice and vice versa. In particular, practical approaches indicate what to focus on to simplify theory, whereas theory seems beneficial to guide quality assessment in practice.

Quality Dimension	Short Description of Dimension
Cogency	Argument has (locally) acceptable, relevant, and sufficient premises.
Local acceptability	Premises worthy of being believed.
Local relevance	Premises support/attack conclusion.
Local sufficiency	Premises enough to draw conclusion.
Effectiveness	Argument persuades audience.
Credibility	Makes author worthy of credence.
Emotional appeal	Makes audience open to arguments.
Clarity	Avoids deviation from the issue, uses correct and unambiguous language.
Appropriateness	Language proportional to the issue, supports credibility and emotions.
Arrangement	Argues in the right order.
Reasonableness	Argument is (globally) acceptable, relevant, and sufficient.
Global acceptability	Audience accepts use of argument.
Global relevance	Argument helps arrive at agreement.
Global sufficiency	Enough rebuttal of counterarguments.
Overall quality	Argumentation quality in total.

Table 1: The 15 theory-based quality dimensions rated in the corpus of Wachsmuth et al. (2017a).

Polarity	Label	Short Description of Reason
Negative properties of Argument B	5-1	*B* is attacking / abusive.
	5-2	*B* has language/grammar issues, or uses humour or sarcasm.
	5-3	*B* is unclear / hard to follow.
	6-1	*B* has no credible evidence / no facts.
	6-2	*B* has less or insufficient reasoning.
	6-3	*B* uses irrelevant reasons.
	7-1	*B* is only an opinion / a rant.
	7-2	*B* is non-sense / confusing.
	7-3	*B* does not address the topic.
	7-4	*B* is generally weak / vague.
Positive properties of Argument A	8-1	*A* has more details/facts/examples, has better reasoning / is deeper.
	8-4	*A* is objective / discusses other views.
	8-5	*A* is more credible / confident.
	9-1	*A* is clear / crisp / well-written.
	9-2	*A* sticks to the topic.
	9-3	*A* makes you think.
	9-4	*A* is well thought through / smart.
Overall	**Conv**	*A* is more convincing than *B*.

Table 2: The 17+1 practical reason labels given in the corpus of Habernal and Gurevych (2016a).

2 Theory versus Practice

This section outlines major theories and practical approaches to argumentation quality assessment, including those we compare in the present paper.

2.1 Theoretical Views of Quality Assessment

Argumentation theory discusses logical, rhetorical, and dialectical quality. As few real-life arguments are logically sound, requiring true premises that deductively entail a conclusion, cogency (as defined in Section 1) is largely seen as the main logical quality (Johnson and Blair, 2006; Damer, 2009; Govier, 2010). Toulmin (1958) models the general structure of logical arguments, and Walton et al. (2008) analyze schemes of fallacies and strong arguments. A fallacy is a kind of error that undermines reasoning (Tindale, 2007). Strength may mean cogency but also rhetorical effectiveness (Perelman and Olbrechts-Tyteca, 1969). Rhetoric has been studied since Aristotle (2007) who developed the notion of the means of persuasion (logos, ethos, pathos) and their linguistic delivery in terms of arrangement and style. Dialectical quality dimensions resemble those of cogency, but arguments are judged specifically by their reasonableness for achieving agreement (van Eemeren and Grootendorst, 2004).

Wachsmuth et al. (2017a) point out that dialectical builds on rhetorical, and rhetorical builds on logical quality. They derive a unifying taxonomy from the major theories, decomposing quality hierarchically into cogency, effectiveness, reasonableness, and subdimensions. Table 1 lists all 15 dimensions covered. In Section 3, we use their absolute quality ratings from 1 (low) to 3 (high) annotated by three experts for each dimension of 304 arguments taken from the *UKPConvArg1* corpus detailed below.

2.2 Practical Views of Quality Assessment

There is an application area where absolute quality ratings of argumentative text are common practice: essay scoring (Beigman Klebanov et al., 2016). Persing and Ng (2015) annotated the argumentative strength of essays composing multiple arguments with notable agreement. For single arguments, however, all existing approaches that we are aware of assess quality in relative terms, e.g., Cabrio and Villata (2012) find accepted arguments based on attack relations, Wei et al. (2016) rank arguments by their persuasiveness, and Wachsmuth et al. (2017b) rank them by their relevance. Boudry et al. (2015) argue that normative concepts such as fallacies rarely apply to real-life arguments and that they are too sophisticated for operationalization.

Based on the idea that relative assessment is easier, Habernal and Gurevych (2016b) crowdsourced the *UKPConvArg1* corpus. Argument pairs *(A, B)* from a debate portal were classified as to which argument is more convincing. Without giving any guidelines, the authors also asked for reasons as to why *A* is more convincing than *B*. In a follow-up study (Habernal and Gurevych, 2016a), these reasons were used to derive a hierarchical annotation scheme. 9111 argument pairs were then labeled with one or more of the 17 reason labels in Table 2

		Negative Properties of Argument B										*Positive Properties of Argument A*							
Quality Dimension		**5-1**	**5-2**	**5-3**	**6-1**	**6-2**	**6-3**	**7-1**	**7-2**	**7-3**	**7-4**	**8-1**	**8-4**	**8-5**	**9-1**	**9-2**	**9-3**	**9-4**	**Conv**
Cog	**Cogency**	.86	.74	.67	.66	.85	.43	.81	.83	.84	.75	.59	.58	.62	.70	.67	.64	**.75**	.59
LA	Local acceptability	.92	.77	.86	.49	.90	**.80**	.86	.89	.89	.74	.58	.43	.73	.64	.67	.56	.73	.58
LR	Local relevance	.87	.77	.86	.70	**.95**	.45	.84	.92	**.95**	.73	.61	.56	.68	.69	.65	.70	.66	.62
LS	Local sufficiency	.79	.69	.67	.68	.74	.38	.85	.92	.84	**.79**	.63	.67	.54	.64	.52	**.78**	.70	.61
Eff	**Effectiveness**	.84	.71	.67	.66	.85	.62	**.87**	.92	.84	.71	.59	.57	.65	.66	.58	**.78**	.72	.59
Cre	Credibility	.78	.69	.71	.52	**.95**	**.80**	.66	.81	.67	.57	.51	.44	.66	.60	.71	.39	.62	.50
Emo	Emotional appeal	.80	.50	.59	.55	.70	**.80**	.70	.80	.67	.60	.36	.35	.41	.30	.42	.73	.50	.38
Cla	Clarity	.61	.70	**.91**	.41	**.95**	.58	.61	.87	.67	.60	.41	.40	.41	.68	.71	.56	.58	.44
App	Appropriateness	.94	**.86**	**.91**	.50	**.95**	.45	**.87**	.74	.36	**.79**	.57	.59	.69	.72	**.79**	.53	.57	.59
Arr	Arrangement	.81	.75	.86	.67	.85	.40	.78	.77	.67	.68	.60	**.73**	.64	**.73**	.73	**.78**	.72	.62
Rea	**Reasonableness**	.92	**.86**	.67	**.73**	.90	.49	.85	**.94**	.84	.73	.64	.56	.70	.69	.65	**.78**	.64	.63
GA	Global acceptability	**1.00**	.80	.82	.65	.76	.62	**.87**	.86	**.95**	.71	.63	.62	**.75**	.59	.67	.72	.68	.63
GR	Global relevance	.97	**.86**	.82	.63	.82	.71	.86	.82	**.95**	.75	.61	.51	.49	.66	.46	.72	.57	.61
GS	Global sufficiency	.77	.57	.59	.62	.85	.47	.75	.72	.71	.64	.59	.69	.46	.53	.39	.71	.61	.56
OQ	**Overall quality**	.94	.85	.79	.71	.90	.53	.85	.92	.84	.72	**.65**	.58	.69	.72	.61	.73	.73	**.64**
# Pairs with label x-y		34	55	18	115	11	16	64	37	10	50	536	79	68	86	34	26	39	736

Table 3: Kendall's τ rank correlation of each of the 15 quality dimensions of all argument pairs annotated by Wachsmuth et al. (2017a) given for each of the 17+1 reason labels of Habernal and Gurevych (2016a). Bold/gray: Highest/lowest value in each column. Bottom row: The number of labels for each dimension.

by crowd workers (*UKPConvArg2*). These pairs represent the practical view in our experiments.

3 Matching Theory and Practice

We now report on experiments that we performed to examine to what extent the theory and practice of argumentation quality assessment match.[1]

3.1 Corpus-based Comparison of the Views

Several dimensions and reasons in Tables 1 and 2 seem to refer to the same or opposite property, e.g., *clarity* and *5-3 (unclear)*. This raises the question of how absolute ratings of arguments based on theory relate to relative comparisons of argument pairs in practice. We informally state three hypotheses:

Hypothesis 1 The reasons for quality differences in practice are adequately represented in theory.

Hypothesis 2 The perception of overall argumentation quality is the same in theory and practice.

Hypothesis 3 Relative quality differences are reflected by differences in absolute quality ratings.

As both corpora described in Section 2 are based on the *UKPConvArg1* corpus and thus share many arguments, we can test the hypotheses empirically.

3.2 Correlations of Dimensions and Reasons

For Hypotheses 1 and 2, we consider all 736 pairs of arguments from Habernal and Gurevych (2016a) where both have been annotated by Wachsmuth et al. (2017a). For each pair *(A, B)* with *A* being more convincing than *B*, we check whether the ratings of *A* and *B* for each dimension (averaged over all annotators) show a concordant difference (i.e., a higher rating for *A*), a disconcordant difference (lower), or a tie. This way, we can correlate each dimension with all reason labels in Table 2 including *Conv*. In particular, we compute Kendall's τ based on all argument pairs given for each label.[2]

Table 3 presents all τ-values. The phrasing of a reason can be assumed to indicate a clear quality difference—this is underlined by the generally high correlations. Analyzing the single values, we find much evidence for Hypothesis 1: Most notably, label *5-1* perfectly correlates with *global acceptability*, fitting the intuition that abuse is not acceptable. The high τ's of *8-5 (more credible)* for *local acceptability* (.73) and of *9-4 (well thought through)* for *cogency* (.75) confirm the match assumed in Section 1. Also, the values of *5-3 (unclear)* for *clarity* (.91) and of *7-2 (non-sense)* for *reasonableness* (.94) as well as the weaker correlation of *8-4 (objective)* for *emotional appeal* (.35) makes sense.

Only the comparably low τ of *6-1 (no credible evidence)* for *local acceptability* (.49) and *credibility* (.52) seem really unexpected. Besides, the descriptions of *6-2* and *6-3* sound like *local* but cor-

[1] Source code and annotated data: http://www.arguana.com

[2] Lacking better options, we ignore pairs where a label is not given: It is indistinguishable whether the associated reason does not hold, has not been given, or is just not included in the corpus. Thus, τ is more "boosted" the fewer pairs exist for a label and, thus, its values are not fully comparable across labels. Notice, though, that *Conv* exists for all pairs. So, the values of *Conv* suggest the magnitude of τ without boosting.

Polarity	Label	Cog	LA	LR	LS	Eff	Cre	Emo	Cla	App	Arr	Rea	GA	GR	GS	OQ
Negative properties of Argument B	5-1	1.30	1.44	1.77	1.29	1.26	1.46	1.64	1.84	1.62	1.55	1.34	1.45	1.65	1.19	1.29
	5-2	1.51	1.73	1.97	1.39	1.41	1.66	1.82	1.96	1.89	1.72	1.55	1.72	1.74	1.21	1.48
	5-3	1.46	1.78	2.06	1.43	1.39	1.63	**1.96**	1.87	2.04	1.65	**1.63**	1.85	1.76	1.28	1.52
	6-1	1.54	**1.87**	2.22	1.43	1.44	**1.72**	1.85	**2.15**	2.12	1.79	1.62	**1.89**	1.89	1.27	1.55
	6-2	1.30	1.52	1.88	1.27	1.21	1.52	1.85	1.94	1.88	1.67	1.36	1.61	1.55	1.15	1.33
	6-3	**1.60**	1.85	**2.23**	**1.52**	**1.52**	1.65	1.79	2.00	**2.15**	**1.92**	**1.63**	1.85	**2.00**	**1.40**	**1.60**
	7-1	1.43	1.74	1.97	1.33	1.34	1.60	1.82	1.95	1.89	1.72	1.48	1.71	1.68	1.22	1.43
	7-2	1.45	1.68	1.97	1.41	1.39	1.53	1.86	1.84	1.95	1.67	1.53	1.68	1.70	1.25	1.48
	7-3	1.20	1.47	1.60	1.10	1.17	1.47	1.60	1.70	1.80	1.40	1.20	1.40	1.30	1.07	1.13
	7-4	1.43	1.71	2.02	1.37	1.34	1.71	1.79	1.95	1.97	1.65	1.55	1.75	1.75	1.23	1.46
Positive properties of Argument A	8-1	1.56	1.89	2.20	1.46	1.48	1.71	1.88	2.05	2.07	1.79	1.65	1.88	1.92	1.30	1.57
	8-4	1.65	1.97	2.27	1.53	1.61	1.73	1.86	2.12	2.14	1.89	1.73	1.92	1.96	1.37	1.64
	8-5	1.69	**2.07**	**2.39**	1.58	1.60	**1.81**	1.98	**2.19**	**2.25**	**1.99**	1.82	2.04	**2.11**	1.38	1.75
	9-1	1.54	1.86	2.22	1.49	1.43	1.67	1.84	2.09	2.03	1.74	1.63	1.85	1.92	1.30	1.54
	9-2	1.56	1.76	2.22	1.45	1.49	1.58	1.98	2.02	2.00	1.74	1.62	1.81	1.84	1.28	1.51
	9-3	1.55	1.78	2.31	1.42	1.49	1.68	**2.01**	2.18	2.10	1.79	1.63	1.83	1.97	1.27	1.50
	9-4	**1.78**	1.99	2.32	**1.64**	**1.68**	**1.81**	1.99	2.17	2.19	1.93	**1.86**	**2.05**	2.09	**1.44**	**1.79**
min(*Pos.*)−min(*Neg.*)		0.34	0.32	0.60	0.32	0.26	0.12	0.24	0.32	0.38	0.34	0.42	0.41	0.54	0.20	0.37
max(*Pos.*)−max(*Neg.*)		0.18	0.20	0.16	0.12	0.16	0.09	0.05	0.04	0.10	0.07	0.23	0.16	0.11	0.04	0.19

Table 4: The mean rating for each quality dimension of those arguments from Wachsmuth et al. (2017a) given for each reason label (Habernal and Gurevych, 2016a). The bottom rows show that the minimum maximum mean ratings are consistently higher for the positive properties than for the negative properties.

relate more with *global* relevance and sufficiency respectively. Similarly, *7-3 (off-topic)* correlates strongly with local *and* global relevance (both .95). So, these dimensions seem hard to separate.

In line with Hypothesis 2, the highest correlation of *Conv* is indeed given for *overall quality* (.64). Thus, argumentation quality assessment seems to match in theory and practice to a broad extent.

3.3 Absolute Ratings for Relative Differences

The correlations found imply that the relative quality differences captured are reflected in absolute differences. For explicitness, we computed the mean rating for each quality dimension of all arguments from Wachsmuth et al. (2017a) with a particular reason label from Habernal and Gurevych (2016a). As each reason refers to one argument of a pair, this reveals whether the labels, although meant to signal relative differences, indicate absolute ratings.

Table 4 compares the mean ratings of "negative labels" (*5-1* to *7-4*) and "positive" ones (*8-1* to *9-4*). For all dimensions, the maximum and minimum value are higher for the positive than for the negative labels—a clear support of Hypothesis 3.[3] Also, Table 4 reveals which reasons predict absolute differences most: The mean ratings of *7-3 (off-topic)* are very low, indicating a strong negative impact, while *6-3 (irrelevant reasons)* still shows rather high values. Vice versa, especially *8-5 (more credible)* and *9-4 (well thought through)* are reflected in high ratings, whereas *9-2 (sticks to topic)* does not have much positive impact.

[3] While the differences seem not very large, this is expected, as in many argument pairs from Habernal and Gurevych (2016a) both arguments are strong or weak respectively.

4 Annotating Theory in Practice

The results of Section 3 suggest that theory may guide the assessment of argumentation quality in practice. In this section, we evaluate the reliability of a crowd-based annotation process.

4.1 Absolute Quality Ratings by the Crowd

We emulated the expert annotation process carried out by Wachsmuth et al. (2017a) on *CrowdFlower* in order to evaluate whether lay annotators suffice for a theory-based quality assessment. In particular, we asked the crowd to rate the same 304 arguments as the experts for all 15 given quality dimensions with scores from 1 to 3 (or choose "cannot judge"). Each argument was rated 10 times at an offered price of $0.10 for each rating (102 annotators in total). Given the crowd ratings, we then performed two comparisons as detailed in the following.

4.2 Agreement of the Crowd with Experts

First, we checked to what extent lay annotators and experts agree in terms of Krippendorff's α. On one hand, we compared the mean of all 10 crowd ratings to the mean of the three ratings of Wachsmuth et al. (2017a). On the other hand, we estimated a reliable rating from the crowd ratings using MACE (Hovy et al., 2013) and compared it to the experts.

Quality Dimension		(a) Crowd / Expert		(b) Crowd 1 / 2 / Expert		(c) Crowd 1 / Expert		(d) Crowd 2 / Expert	
		Mean	MACE	Mean	MACE	Mean	MACE	Mean	MACE
Cog	**Cogency**	.27	.38	.24	.29	.38	.37	.05	.27
LA	Local acceptability	.49	.35	.37	.27	.49	.33	.30	.25
LR	Local relevance	.42	.39	.33	.28	.41	.39	.26	.25
LS	Local sufficiency	.18	.31	.21	.21	.34	.27	–.04	.19
Eff	**Effectiveness**	.13	.31	.19	.20	.27	.28	–.06	.20
Cre	Credibility	.41	.27	.31	.20	.43	.23	.22	.19
Emo	Emotional appeal	.45	.23	.32	.13	.41	.20	.25	.10
Cla	Clarity	.42	.28	.33	.23	.39	.27	.29	.20
App	Appropriateness	**.54**	.26	**.40**	.20	.48	.24	**.43**	.17
Arr	Arrangement	.53	.30	.36	.24	.49	.27	.35	.24
Rea	**Reasonableness**	.33	.40	.27	.31	.42	**.40**	.09	.29
GA	Global acceptability	**.54**	.40	.36	.29	**.53**	.37	.33	.28
GR	Global relevance	.44	.31	.31	.20	.50	.29	.22	.18
GS	Global sufficiency	–.17	.19	.04	.11	.00	.16	–.27	.11
OQ	**Overall quality**	.43	**.43**	.38	**.33**	.43	**.40**	.28	**.33**

Table 5: Mean and MACE Krippendorff's α agreement between (a) the crowd and the experts, (b) two independent crowd groups and the experts, (c) group 1 and the experts, and (d) group 2 and the experts.

Table 5(a) presents the results. For the mean ratings, most α-values are above .40. This is similar to the study of Wachsmuth et al. (2017b), where a range of .27 to .51 is reported, meaning that lay annotators achieve similar agreement to experts. Considering the minimum of mean and MACE, we observe the highest agreement for *overall quality* (.43)—analog to Wachsmuth et al. (2017b). Also, *global sufficiency* has the lowest agreement in both cases. In contrast, the experts hardly said "cannot judge" at all, whereas the crowd chose it for about 4% of all ratings (most often for global sufficiency), possibly due to a lack of training. Still, we conclude that the crowd generally handles the theory-based quality assessment almost as well as the experts.

However, the complexity of the assessment is underlined by the generally limited agreement, suggesting that either simplification or stricter guidelines are needed. Regarding simplification, the most common practical reasons of Habernal and Gurevych (2016a) imply what to focus on.

4.3 Reliability of the Crowd Annotations

In the second comparison, we checked how many crowd annotators are needed to compete with the experts. For this purpose, we split the crowd ratings into two independent groups of 5 and treated the mean and MACE of each group as a single rating. We then computed the agreement of both groups and each group individually against the experts.

The α-values for both groups are listed in Table 5(b). On average, they are a bit lower than those of all 10 crowd annotators in Table 5(a). Hence, five crowd ratings per argument seem not enough for sufficient reliability. Tables 5(c) and 5(d) reveal the reason behind, namely, the results of crowd group 1 and group 2 differ clearly. At the same time, the values in Table 5(c) are close to those in Table 5(a), so 10 ratings might suffice. Moreover, we see that the most stable α-values in Table 5 are given for *overall quality*, indicating that the theory indeed helps assessing quality reliably.

5 Conclusion

This paper demonstrates that the theory and practice of assessing argumentation quality can learn from each other. Most reasons for quality differences phrased in practice seem well-represented in the normative view of theory and correlate with absolute quality ratings. In our study, lay annotators had similar agreement on the ratings as experts. Considering that some common reasons are quite vague, the diverse and comprehensive theoretical view of argumentation quality may guide a more insightful assessment. On the other hand, some quality dimensions remain hard to assess and/or to separate in practice, resulting in limited agreement. Simplifying theory along the most important reasons will thus improve its practical applicability.

Acknowledgments

We thank Vinodkumar Prabhakaran and Yonatan Bilu for their ongoing participation in our research on argumentation quality. Also, we acknowledge financial support of the DFG (ArguAna, AIPHES), the Natural Sciences and Engineering Research Council of Canada, and the Volkswagen Foundation (Lichtenberg-Professorship Program).

References

Jens Allwood. 2016. Argumentation, activity and culture. In *6th International Conference on Computational Models of Argument (COMMA 16)*. Potsdam, Germany, page 3.

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse* (George A. Kennedy, translator). Clarendon Aristotle series. Oxford University Press.

Beata Beigman Klebanov, Christian Stab, Jill Burstein, Yi Song, Binod Gyawali, and Iryna Gurevych. 2016. Argumentation: Content, structure, and relationship with essay quality. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*. Association for Computational Linguistics, pages 70–75. https://doi.org/10.18653/v1/W16-2808.

Maarten Boudry, Fabio Paglieri, and Massimo Pigliucci. 2015. The fake, the flimsy, and the fallacious: Demarcating arguments in real life. *Argumentation* 29(4):431–456.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 208–212. http://aclweb.org/anthology/P12-2041.

T. Edward Damer. 2009. *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments*. Wadsworth, Cengage Learning, 6th edition.

Trudy Govier. 2010. *A Practical Study of Argument*. Wadsworth, Cengage Learning, 7th edition.

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*. Association for Computational Linguistics, pages 1214–1223. http://aclweb.org/anthology/D16-1129.

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)*. Association for Computational Linguistics, pages 1589–1599. https://doi.org/10.18653/v1/P16-1150.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pages 1120–1130. http://aclweb.org/anthology/N13-1132.

Ralph H. Johnson and J. Anthony Blair. 2006. *Logical Self-defense*. Intern. Debate Education Association.

Chaïm Perelman and Lucie Olbrechts-Tyteca. 1969. *The New Rhetoric: A Treatise on Argumentation* (John Wilkinson and Purcell Weaver, translator). 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)*. Association for Computational Linguistics, pages 543–552. https://doi.org/10.3115/v1/P15-1053.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh 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*. Association for Computational Linguistics, pages 440–450. https://doi.org/10.18653/v1/D15-1050.

Christopher W. Tindale. 2007. *Fallacies and Argument Appraisal. Critical Reasoning and Argumentation*. Cambridge University Press.

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

Frans H. van Eemeren and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge University Press.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Alberdingk Tim Thijm, Graeme Hirst, and Benno Stein. 2017a. 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*. Association for Computational Linguistics, pages 176–187. http://aclweb.org/anthology/E17-1017.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017b. "PageRank" for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, pages 1117–1127. http://aclweb.org/anthology/E17-1105.

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

Zhongyu Wei, Yang Liu, and Yi Li. 2016. Is this post persuasive? Ranking argumentative comments in online forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, pages 195–200. https://doi.org/10.18653/v1/P16-2032.

Bibliography

Rob Abbott, Marilyn Walker, Pranav Anand, Jean E. Fox Tree, Robeson Bowmani, and Joseph King. How can you say such things?!?: Recognizing disagreement in informal political argument. In *Proceedings of the Workshop on Languages in Social Media*, pages 2–11, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argumentation Mining*, pages 64–68, 2014.

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. Unit segmentation of argumentative texts. In *Proceedings of the 4th Workshop on Argument Mining*, pages 118–128, 2017.

Aristotle. *On Rhetoric: A Theory of Civic Discourse* (translated by G.A. Kennedy). Oxford University Press, 2007.

Jonathan Baron. Myside bias in thinking about abortion. *Thinking & Reasoning*, 1(3):221–235, 1995.

Stephen R. Bates, Peter Kerr, Christopher Byrne, and Liam Stanley. Questions to the prime minister: A comparative study of pmqs from thatcher to cameron. *Parliamentary Affairs*, 67 (2):253–280, 2012.

Eric Baumer, Elisha Elovic, Ying Qin, Francesca Polletta, and Geri Gay. Testing and comparing computational approaches for identifying the language of framing in political news. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1472–1482, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

Monroe C. Beardsley. *Practical Logic*. New York, Prentice-Hall, 1950.

Kaspar Beelen, Timothy Alberdingk Thijm, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, Maarten Marx, Nona Naderi, Ludovic Rheault, Roman Polyanovsky, and Tanya Whyte. Digitization of the Canadian parliamentary debates. *Canadian Journal of Political Science*, 50(3):849–864, 2017.

William L. Benoit. *Accounts, Excuses, and Apologies: A Theory of Image Restoration Strategies*. State University of New York Press, Albany, 1995.

William L. Benoit and Shirley Drew. Appropriateness and effectiveness of image repair strategies. *Communication Reports*, 10(2):153–163, 1997.

William L. Benoit and Jayne R. Henson. President Bush's image repair discourse on Hurricane Katrina. *Public Relations Review*, 35(1):40–46, 2009.

Jamal Bentahar, Bernard Moulin, and Micheline Bélanger. A taxonomy of argumentation models used for knowledge representation. *Artificial Intelligence Review*, 33(3):211–259, 2010.

Philippe Besnard and Anthony Hunter. *Elements of argumentation*. MIT Press Cambridge, 2008.

Jaspreet Bhatia, Travis D. Breaux, Joel R. Reidenberg, and Thomas B. Norton. A theory of vagueness and privacy risk perception. In *Requirements Engineering Conference (RE), 2016 IEEE 24th International*, pages 26–35. IEEE, 2016.

Or Biran and Kathleen McKeown. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 69–73, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Or Biran and Owen Rambow. Identifying justifications in written dialogs. In *Proceedings of the 2011 IEEE Fifth International Conference on Semantic Computing*, ICSC '11, pages 162–168, Washington, DC, USA, 2011. IEEE Computer Society.

Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. O'Reilly Media, Inc., 2009.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

Filip Boltužić and Jan Šnajder. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, 2014.

Filip Boltužić and Jan Šnajder. Identifying prominent arguments in online debates using semantic textual similarity. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 110–115, Denver, CO, June 2015. Association for Computational Linguistics.

Filip Boltuzic and Jan Šnajder. Fill the gap! analyzing implicit premises between claims from online debates. In *Proceedings of the Third Workshop on Argument Mining*, pages 124–133, 2016.

Amber Boydstun. Issue framing as a generalizable phenomenon. *Association for Computational Linguistics*, page 71, 2014.

Susan L. Brinson and William L. Benoit. The tarnished star: Restoring Texaco's damaged public image. *Management Communication Quarterly*, 12(4):483–510, 1999.

Penelope Brown and Stephen C. Levinson. *Politeness: Some Universals in Language Usage*, volume 4. Cambridge University Press, 1987.

Judee K. Burgoon and Jerold L. Hale. The fundamental topoi of relational communication. *Communication Monographs*, 51(3):193–214, 1984.

Judith P. Burns and Michael S. Bruner. Revisiting the theory of image restoration strategies. *Communication Quarterly*, 48(1):27–39, 2000.

Elena Cabrio and Serena Villata. Natural language arguments: A combined approach. In *20th European Conference on Artificial Intelligence (ECAI 2012)*, 2012.

Dallas Card, E. Amber Boydstun, H. Justin Gross, Philip Resnik, and A. Noah Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444. Association for Computational Linguistics, 2015.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A. Smith. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420, Austin, Texas, November 2016. Association for Computational Linguistics.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259,*, 2014.

Dennis Chong and James N. Druckman. Framing theory. *Annual Review of Political Science*, 10, 2007.

David E. Clementson. Effects of dodging questions: How politicians escape deception detection and how they get caught. *Journal of Language and Social Psychology*, 37(1):93–113, 2018.

Christopher Cochrane. The asymmetrical structure of left/right disagreement: Left-wing coherence and right-wing fragmentation in comparative party policy. *Party Politics*, 19(1): 104–121, 2013.

W. Timothy Coombs. Choosing the right words: The development of guidelines for the selection of the "appropriate" crisis-response strategies. *Management communication quarterly*, 8(4): 447–476, 1995.

W. Timothy Coombs. An analytic framework for crisis situations: Better responses from a better understanding of the situation. *Journal of public relations research*, 10(3):177–191, 1998.

W. Timothy Coombs and Sherry J. Holladay. Comparing apology to equivalent crisis response strategies: Clarifying apology's role and value in crisis communication. *Public Relations Review*, 34(3):252–257, 2008.

Soledad Peŕez de Ayala. FTAs and Erskine May: Conflicting needs? Politeness in question time. *Journal of Pragmatics*, 33(2):143–169, 2001. ISSN 0378-2166. doi: https://doi.org/10.1016/S0378-2166(00)00002-3.

Marie-Catherine De Marneffe and Christopher D. Manning. The stanford typed dependencies representation. In *COLING 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8. Association for Computational Linguistics, 2008.

Claes H. de Vreese. News framing: Theory and typology. *Information Design Journal + Document Design*, 13(1):51–62, 2005.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-

training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Sharon D. Downey. The evolution of the rhetorical genre of apologia. *Western Journal of Communication*, 57(1):42–64, 1993.

Phan Minh Dung. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357, 1995.

Esin Durmus and Claire Cardie. 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, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1094. URL `https://www.aclweb.org/anthology/N18-1094`.

Rory Duthie and Katarzyna Budzynska. A Deep Modular RNN Approach for Ethos Mining. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4041–4047. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

Robert M. Entman. Framing: Towards clarification of a fractured paradigm. *McQuail's Reader in Mass Communication Theory*, pages 390–397, 1993.

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

Vanessa Wei Feng and Graeme Hirst. Text-level discourse parsing with rich linguistic features. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), pages 60–68, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

Eirini Florou, Stasinos Konstantopoulos, Antonis Koukourikos, and Pythagoras Karampiperis. Argument extraction for supporting public policy formulation. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 49–54, 2013.

Bruce Fraser. Pragmatic competence: The case of hedging. In Gunther Kaltenböck, Wiltrud Mihatsch, and Stefan Schneider, editors, *New Approaches to Hedging*, pages 15–34. Brill, 2012.

James B. Freeman. *Dialectics and the macrostructure of arguments: A theory of argument structure*. Walter de Gruyter, 1991.

William A. Gamson and Andre Modigliani. Media discourse and public opinion on nuclear power: A constructionist approach. *American Journal of Sociology*, pages 1–37, 1989.

Debanjan Ghosh, Aquila Khanam, Yubo Han, and Smaranda Muresan. Coarse-grained argumentation features for scoring persuasive essays. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 549–554, Berlin, Germany, August 2016. Association for Computational Linguistics.

Erving Goffman. *Interaction Ritual: Essays on face-to-face interaction.* Aldine, 1967.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. Argument extraction from news, blogs, and social media. In Aristidis Likas, Konstantinos Blekas, and Dimitris Kalles, editors, *Artificial Intelligence: Methods and Applications: 8th Hellenic Conference on AI, SETN 2014, Ioannina, Greece, May 15-17, 2014. Proceedings*, pages 287–299. Springer International Publishing, Cham, 2014.

Trudy Gover. *Problems in Argument Analysis and Evaluation*. Windsor: Windsor Studies in Argumentation, 2018.

Floriana Grasso. Towards a framework for rhetorical argumentation. In *EDILOG 02: Proceedings of the 6th workshop on the semantics and pragmatics of dialogue*, pages 53–60, 2002.

Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. Hybrid speech recognition with deep bidirectional LSTM. In *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, pages 273–278. IEEE, 2013.

Ivan Habernal and Iryna Gurevych. 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, August 2016a. Association for Computational Linguistics.

Ivan Habernal and Iryna Gurevych. 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. Association for Computational Linguistics, 2016b.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. Argumentation mining on the web from information seeking perspective. In Elena Cabrio, Serena Villata, and Adam Wyner, editors, *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 26–39. CEUR-WS, 2014.

John S Hammond, Ralph L Keeney, and Howard Raiffa. The hidden traps in decision making. *Harvard business review*, 76(5):47–58, 1998.

William Forrest Harlow, Brian C Brantley, and Rachel Martin Harlow. BP initial image repair strategies after the Deepwater Horizon spill. *Public Relations Review*, 37(1):80–83, 2011.

Kazi Saidul Hasan and Vincent Ng. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar, October 2014. Association for Computational Linguistics.

Naeemul Hassan, Chengkai Li, and Mark Tremayne. Detecting check-worthy factual claims in presidential debates. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 1835–1838. ACM, 2015.

Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

Graeme Hirst. Computation, information, cognition – the nexus and the liminal. chapter Views of text-meaning in computational linguistics: Past, present, and future, pages 270–279. Cambridge Scholars Publishing, Newcastle-upon-Tyne, 2007.

Graeme Hirst, Vanessa Wei Feng, Christopher Cochrane, and Nona Naderi. Argumentation, ideology, and issue framing in parliamentary discourse. In *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, Italy, July 2014. Published as CEUR Workshop Proceedings, volume 1341.

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

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, 2013.

Cornelia Ilie. Parliamentary discourses. In Keith Brown, Anne H. Anderson, Laurie Bauer, Margie Berns, Graeme Hirst, and Jim Miller, editors, *Encyclopedia of Language and Linguis-*

tics, pages 188–196. Elsevier, Oxford, second edition, 2006. doi: https://doi.org/10.1016/B0-08-044854-2/00720-3.

Israa Jaradat, Pepa Gencheva, Alberto Barrón-Cedeño, Lluís Màrquez, and Preslav Nakov. Claimrank: Detecting check-worthy claims in arabic and english. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 26–30. Association for Computational Linguistics, 2018.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

Jaana Juvonen. The social functions of attributional face-saving tactics among early adolescents. *Educational Psychology Review*, 12(1):15–32, 2000.

Daniel Kahneman and Amos Tversky. Choices, values, and frames. In *Handbook of the Fundamentals of Financial Decision Making: Part I*, pages 269–278. World Scientific, 2013.

Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Skip-thought vectors. *arXiv preprint arXiv:1506.06726*, 2015.

Dan Klein and Christopher D. Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 423–430, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

Michael Koller. Rebutting accusations: When does it work, when does it fail? *European Journal of Social Psychology*, 23(4):373–389, 1993.

Jey Han Lau, Timothy Baldwin, and Trevor Cohn. Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 355–365, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Logan Lebanoff and Fei Liu. Automatic detection of vague words and sentences in privacy policies. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, November 2018. Association for Computational Linguistics.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Greg Leichty and James L. Applegate. Social-cognitive and situational influences on the use of face-saving persuasive strategies. *Human Communication Research*, 17(3):451–484, 1991.

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

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Singapore, August 2009. Association for Computational Linguistics.

Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20(2):151–184, Apr 2014.

Lisa Lyon and Glen T. Cameron. A relational approach examining the interplay of prior reputation and immediate response to a crisis. *Journal of Public Relations Research*, 16(3): 213–241, 2004.

Jing Ma, Wei Gao, and Kam-Fai Wong. Detect rumors in microblog posts using propagation structure via kernel learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 708–717, Vancouver, Canada, July 2017. Association for Computational Linguistics.

Fabrizio Macagno and Aikaterini Konstantinidou. What students' arguments can tell us: Using argumentation schemes in science education. *Argumentation*, 27(3):225–243, 2013. ISSN 1572-8374.

Robyn Macpherson. *Predictors of belief bias in critical thinking tasks*. PhD thesis, University of Toronto, 2006.

Robyn Macpherson and Keith E. Stanovich. Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learning and individual differences*, 17(2): 115–127, 2007.

Daniel Marcu and Abdessamad Echihabi. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 368–375, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

Lucia Mason and Fabio Scirica. Prediction of students' argumentation skills about controversial topics by epistemological understanding. *Learning and instruction*, 16(5):492–509, 2006.

John E. McPeck. *Critical thinking and education*. Routledge, 2016.

Mary L. Means and James F. Voss. Who reasons well? two studies of informal reasoning among children of different grade, ability, and knowledge levels. *Cognition and instruction*, 14(2): 139–178, 1996.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Markus J. Milne and Dennis M. Patten. Securing organizational legitimacy: An experimental decision case examining the impact of environmental disclosures. *Accounting, Auditing & Accountability Journal*, 15(3):372–405, 2002.

Amita Misra, Pranav Anand, Jean E. Fox Tree, and Marilyn A. Walker. Using summarization to discover argument facets in online idealogical dialog. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 430–440, 2015.

Jeff Mitchell and Mirella Lapata. Vector-based models of semantic composition. In *Association for Computational Linguistics*, pages 236–244, 2008.

Raquel Mochales and Marie-Francine Moens. Study on the structure of argumentation in case law. In *Proceedings of the 2008 Conference on Legal Knowledge and Information Systems: JURIX 2008: The Twenty-First Annual Conference*, pages 11–20, Amsterdam, The Netherlands, 2008. IOS Press.

Sanjay Modgil, Francesca Toni, Floris Bex, Ivan Bratko, Carlos I. Chesñevar, Wolfgang Dvořák, Marcelo A. Falappa, Xiuyi Fan, Sarah Alice Gaggl, Alejandro J. García, María P. González, Thomas F. Gordon, João Leite, Martin Možina, Chris Reed, Guillermo R. Simari, Stefan Szeider, Paolo Torroni, and Stefan Woltran. The added value of argumentation. In Sascha Ossowski, editor, *Agreement Technologies*, pages 357–403. Springer Netherlands, Dordrecht, 2013.

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

Saif Mohammad. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 174–184. Association for Computational Linguistics, 2018.

Saif M. Mohammad and Peter D. Turney. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465, 2013.

Dima Mohammed. Institutional insights for analysing strategic manoeuvring in the British Prime Minister's Question Time. *Argumentation*, 22(3):377–393, 2008.

Arjun Mukherjee and Bing Liu. Discovering user interactions in ideological discussions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 671–681, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

Joan Mulholland. *A Handbook of Persuasive Tactics: A Practical Language Guide*. Routledge, 2003.

Nona Naderi. Argumentation mining in parliamentary discourse. In Pat Bondy and Laura Benacquista, editors, *Argumentation, Objectivity, and Bias: Proceedings of the 11th International Conference of the Ontario Society for the Study of Argumentation (OSSA)*, pages 1–9, Windsor, Canada, May 2016.

Nona Naderi and Graeme Hirst. Argumentation mining in parliamentary discourse. In Matteo Baldoni et al., editor, *Principles and Practice of Multi-Agent Systems*, pages 16–25. Springer International Publishing, 2016.

Nona Naderi and Graeme Hirst. Recognizing reputation defence strategies in critical political exchanges. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 527–535, Varna, Bulgaria, 2017a.

Nona Naderi and Graeme Hirst. Classifying frames at the sentence level in news articles.

In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 536–542, Varna, Bulgaria, September 2017b. INCOMA Ltd.

Nona Naderi and Graeme Hirst. Automatically labeled data generation for classification of reputation defence strategies. In Darja Fišer, Maria Eskevich, and Franciska de Jong, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France, may 2018a. European Language Resources Association (ELRA).

Nona Naderi and Graeme Hirst. Automated fact-checking of claims in argumentative parliamentary debates. In *Proceedings of the First Workshop on Fact Extraction and Verification (to appear).*, Brussels, Belgium, November 2018b. Association for Computational Linguistics.

Nona Naderi and Graeme Hirst. Using context to identify the language of face-saving. In *Proceedings of the 5th Workshop on Argument Mining (to appear)*, Brussels, Belgium, November 2018c. Association for Computational Linguistics.

Darcia Narvaez. Moral text comprehension: Implications for education and research. *Journal of Moral Education*, 30(1):43–54, 2001.

Huy Nguyen and Diane Litman. Context-aware argumentative relation mining. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1127–1137, Berlin, Germany, August 2016. Association for Computational Linguistics.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. Tea Party in the house: A hierarchical ideal point topic model and its application to Republican legislators in the 112th Congress. 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 1438–1448. Association for Computational Linguistics, 2015.

Alan Partington. *The linguistics of political argument: The spin-doctor and the wolf-pack at the White House*. Routledge, 2003.

Ayush Patwari, Dan Goldwasser, and Saurabh Bagchi. TATHYA: A multi-classifier system for detecting check-worthy statements in political debates. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2259–2262. ACM, 2017.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: a survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31, 2013.

Robyn Penman. Facework & politeness: Multiple goals in courtroom discourse. *Journal of Language and Social Psychology*, 9(1-2):15–38, 1990. URL `https://doi.org/10.1177/0261927X9091002`.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

Isaac Persing and Vincent Ng. 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, June 2016. Association for Computational Linguistics.

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

Emily Pitler, Annie Louis, and Ani Nenkova. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691, Suntec, Singapore, August 2009. Association for Computational Linguistics.

John L. Pollock. Defeasible reasoning. *Cognitive Science*, 11(4):481–518, 1987.

John L. Pollock. A theory of defeasible reasoning. *International Journal of Intelligent Systems*, 6(1):33–54, 1991.

Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2931–2937, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

Chris Reed and Glenn Rowe. Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(04):961–979, 2004.

Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 105–112, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 25–32, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh Khapra, Ehud Aharoni, and Noam Slonim. 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. Association for Computational Linguistics, 2015a. doi: 10.18653/v1/D15-1050.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni, and Noam Slonim. 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, September 2015b. Association for Computational Linguistics.

Sara Rosenthal and Kathleen McKeown. Detecting opinionated claims in online discussions. In *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference on*, pages 30–37. IEEE, 2012.

Sara Rosenthal and Kathy McKeown. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177, Prague, Czech Republic, September 2015. Association for Computational Linguistics.

Glenn Rowe and Chris Reed. Argument diagramming: The araucaria project. In *Knowledge Cartography*, pages 163–181. Springer, 2008.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Halford Ross Ryan. Kategoria and apologia: On their rhetorical criticism as a speech set. *Quarterly Journal of Speech*, 68(3):254–261, 1982.

Jodi Schneider, Krystian Samp, Alexandre Passant, and Stefan Decker. Arguments about deletion: How experience improves the acceptability of arguments in ad-hoc online task groups. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1069–1080. ACM, 2013.

Karin Kipper Schuler. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, 2005.

Holli A. Semetko and Patti M. Valkenburg. Framing European politics: A content analysis of press and television news. *Journal of Communication*, 50(2):93–109, 2000.

Catherine A. Sheldon and Lynne M. Sallot. Image repair in politics: Testing effects of communication strategy and performance history in a faux pas. *Journal of Public Relations Research*, 21(1):25–50, 2008.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics, 2011.

Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 46–56, Doha, Qatar, October 2014a. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland, August 2014b. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659, 2017. doi: 10.1162/COLI_a_00295. URL https://doi.org/10.1162/COLI_a_00295.

Keith E Stanovich and Richard F West. Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology*, 89(2):342, 1997.

M. Stede, J. Schneider, and G. Hirst. *Argumentation Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2018.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. *arXiv preprint arXiv:1503.00075*, 2015.

Yla R. Tausczik and James W. Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1): 24–54, 2010.

Nithum Thain, Lucas Dixon, and Ellery Wulczyn. Wikipedia Talk Labels: Toxicity. `https://figshare.com/articles/Wikipedia_Talk_Labels_Toxicity/4563973`, 2 2017.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and verification. In *NAACL-HLT*, 2018.

Christopher W Tindale. Constrained maneuvering: Rhetoric as a rational enterprise. *Argumentation*, 20(4):447–466, 2006.

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

Karen Tracy. A facework system of minimal politeness: Oral argument in appellate court. *Journal of Politeness Research. Language, Behaviour, Culture*, 7(1):123–145, 2011.

Anna Trosborg. Apology strategies in natives/non-natives. *Journal of pragmatics*, 11(2): 147–167, 1987.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. ICWSM - A great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In William W. Cohen and Samuel Gosling, editors, *International AAAI Conference on Weblogs and Social Media*. The AAAI Press, 2010.

Oren Tsur, Dan Calacci, and David Lazer. A Frame of Mind: Using Statistical Models for Detection of Framing and Agenda Setting Campaigns. 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 1629–1638. Association for Computational Linguistics, 2015.

Frans H. van Eemeren. *Strategic maneuvering in argumentative discourse: Extending the pragma-dialectical theory of argumentation*, volume 2. John Benjamins Publishing, 2010.

Frans H. van Eemeren and Frans Hendrik Eemeren. *Examining argumentation in context: Fifteen studies on strategic maneuvering*, volume 1. John Benjamins Publishing, 2009.

Frans H. van Eemeren and Rob Grootendorst. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press, 2004.

Andreas Vlachos and Sebastian Riedel. Fact checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 18–22, Baltimore, MD, USA, June 2014. Association for Computational Linguistics.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. Argumentation Quality Assessment: Theory vs. Practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 17)*, August 2017.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765, 2018a.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 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. Association for Computational Linguistics, 2018b.

Douglas Walton. *Plausible argument in everyday conversation*. SUNY Press, 1992.

Douglas Walton. *Fundamentals of critical argumentation*. Cambridge University Press, 2005.

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

Lu Wang and Claire Cardie. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 97–106, Baltimore, Maryland, June 2014. Association for Computational Linguistics.

Tong Wang, Abdelrahman Mohamed, and Graeme Hirst. Learning lexical embeddings with syntactic and lexicographic knowledge. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 458–463, Beijing, China, July 2015. Association for Computational Linguistics.

William Yang Wang. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*, 2017.

Zhongyu Wei, Yang Liu, and Yi Li. Is This Post Persuasive? Ranking Argumentative Comments in Online Forum. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 195–200, Berlin, Germany, August 2016. Association for Computational Linguistics.

Karl Weiss, Taghi M. Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

Harald R. Wohlrapp. *The concept of argument: A philosophical foundation*, volume 4. Springer Netherlands, 2014.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. Wikipedia Talk Labels: Personal Attacks. `https://figshare.com/articles/Wikipedia_Talk_Labels_Personal_Attacks/4054689`, 2 2017.

Ernest Zhang and William L. Benoit. Former Minister Zhang's discourse on SARS: Government's image restoration or destruction? *Public relations review*, 35(3):240–246, 2009.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 649–657. Curran Associates, Inc., 2015.