UNIVERSITY OF CALIFORNIA
SANTA CRUZ

ARGUMENT FACETS IN SOCIAL MEDIA DIALOGUE

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

COMPUTER SCIENCE

by

Amita Misra

June 2018

The Dissertation of Amita Misra
is approved:

Professor Marilyn Walker, Chair

Professor Snigdha Chaturvedi

Professor Jean E Fox Tree

Professor Pranav Anand

Tyrus Miller
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Argument Facets in Social Media Dialogue

by

Amita Misra

Argumentation is an interactive process and frequently occurs in conversations. Social networks and online debates provide a two-way communication platform with a huge amount of opinion and argument rich information. However, the massive amount of information available today is overwhelming to our brain and its ability to efficiently absorb and comprehend. Identifying and integrating arguments across these discussions requires computational methods that can facilitate users to systematically search, analyze and summarize arguments as well as reason about the relationship among arguments. We develop techniques to recognize the specific arguments and counter arguments people tend to advance and group them across discussions as *Facets*. This entails two sub-tasks: (i) *Extract and summarize important arguments*, and (ii) *Discover similar repeated argument aspects that appear across multiple dialogues on a topic.* We present a systematic approach to leverage pyramid based summarization framework to identify *central propositions* as those arguments that people find most salient, and then *rank and select arguments* in social media dialogue, which is a novel method for ranking arguments in conversational data. Our results show that adding contextual knowledge from a dialogue improves argument extraction. We introduce a new task of Argument Facet Similarity (AFS), where we develop a new corpus aimed at identifying various *facets* across opinionated dialogue. A graded argument similarity model was defined that takes as input two sentential arguments

and returns a scalar value that categorizes their similarity (AFS). The prediction output obtained from the proposed model improves the results obtained from previous work that defines similarity of short texts as Semantic Textual Similarity (STS).

This thesis is dedicated to my Parents.

Acknowledgments

I take this opportunity to express my heartfelt gratitude to all my teachers and friends who have helped and inspired me during my doctoral study. First and foremost, I am greatly indebted to my advisor, Marilyn Walker for her mentoring, guidance, and consistent support. I couldn't have asked for a more committed, more enthusiastic, or a more passionate advisor. The energy and enthusiasm she has for research is contagious and motivational, and helped me navigate even the toughest times in the Ph.D. journey. Thank you for encouraging and guiding me through the subtleties of research writing, and for helping me develop my potential as a researcher and writer.

To Pranav Anand, who helped at various phases of this research and as my committee member. His door was always open for discussions whenever I approached him. After meeting with him, I always came out with my mind full of new ideas and ready to explore new dimensions. I owe a lot to him for making this possible.

I gratefully acknowledge the others members of my Ph.D. committee, Jean E Fox Tree and Snigdha Chaturvedi, for their time and valuable feedback on a preliminary version of this thesis.

To all the members of the NLDS lab who have been my colleagues and friends during my Ph.D. years. The numerous discussions with an amazing group of talents, Stephanie Lukin, Zhichao Hu, Lena Reed, Rob Abbott, Reid Swanson, Kevin Bowden, Geetanjali Rakshit, and Jiaqi Wu helped to clarify my ideas and concepts. To Elahe Rahimtoroghi and Shereen Oraby, for your thoughts, well-wishes, messages, visits, chit-chats, the editing advice, and being there

whenever I needed a friend.

Finally, I extend my deepest gratitude to my family. My husband Vijay, who always encouraged me, made me believe in myself and realize my strengths. My children, Aryan and Ishita, for being patient all these years and who provided the much needed love, enthusiasm and joy. This journey would not have been possible without them. Thank you!

Chapter 1

Introduction

Argumentation is a verbal, social, and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of propositions justifying or refuting the proposition expressed in the standpoint.

(Van Eemeren and Grootendorst, 2004)

As a verbal activity, argumentation includes some form of communication (words and sentences) to state, deny, accept, or reject a proposition. Argumentation is also a social process, as arguments are usually directed towards other people, and are exchanged when people differ on a standpoint. Thus, argumentation presupposes a controversial standpoint about which the participants disagree. It can be seen as a means of resolving differences of opinion. It involves two or more individuals responding to one another's claim to justify their position or to contradict someone else's. The constellation of propositions advanced in argumentation is often referred to by the term argument (Van Eemeren and Grootendorst, 2004). Arguments unfold based on the contributions of participants. Since the aim is to convince another person of the acceptabil-

ity of the standpoint, a reasonable critic is assumed; otherwise, it does not make sense to further advance the argumentation. The goal is to either prove or disprove that particular standpoint.

Closely related to the Van Eemeren and Grootendorst's view on argumentation is Jacobs and Jackson's position based on conversational approach: argumentation in everyday conversations. They describe an argument as a "disagreement managing device in conversations." Specific speech acts can be paired into adjacency pairs such as question-answer, offer-accept/decline, and request grant/refuse. Arguments are characterized by the projection, production, suppression, or resolution of disagreement to regulate the occurrence of disagreeable speech acts (Jacobs and Jackson, 1982). Walton's conception of argumentation takes into account the language use, dialogical context, and the original intention of the speaker (i.e., the illocutionary force of a speech act) (Walton, 1992). Based on a dialogic model of communication 'pro' and 'con' arguments that resemble zero-sum games in which one participant wins and the other loses, receive more attention from the audience (Salmon and Zeitz, 1995).

As per Van Eemeren et al. (2014), a plethora of argumentation theories exist. It is an interdisciplinary field pervaded by insights from philosophy, logic, psychology, communications studies, linguistics, with different interpretations, analyses, and uses of arguments. Theorists working mainly within Artificial Intelligence developed formal computational models of argumentation over complex argumentation structures, and made substantial progress in providing abstract and structured formal models to represent and reason over argumentation structures. However, there has been very limited research applying these computational models to naturally occurring argumentation in text (Cabrio et al., 2016). It is still a challenge to understand how humans reason and argue in debates, a vital issue in cognitive science. Identifying

and weighing "pro" and "con" arguments, via cognitive processes involve persuasions and emotions, which are inherently harder to formalize from a computational perspective (Villata et al., 2017). A desirable characteristic of intelligent machines is to find arguments in an automated way in human discourse (Moens et al., 2007). An increasing amount of diverse textual data is now available from a variety of sources including social networks, online debate dialogues, newspaper comments, and blogs generating a significant resource of dynamic natural language conversations from where arguments can be identified and analyzed. The enormous volume and the breadth of data available today, together with advances in computational techniques created fertile ground for the investigation of the computational aspects of human argumentation in a holistic manner. This led to the rise of a new research area called **'Argumentation Mining'**, or **'Argument Mining'**, also referred to as **'Computational Argumentation'**, defined as *The detection of an argumentative discourse structure in text or speech, and the identification and the functional classification of its composing components (Moens, 2013).*

It involves automated extraction of natural language arguments and their relations from generic textual corpora, e.g., the premises, conclusion, argumentation scheme of each argument, as well as relationships between pairs of arguments in a given document. Interest in this field rapidly increased over the past few years, as evident by a number of events like, *Frontiers and Connections between Argumentation Theory and Natural Language Processing in 2014; workshops on Argumentation Mining at Association of Computational Linguistics (ACL) 2014, ACL 2016 and Conference on Empirical Methods in Natural Language Processing (EMNLP) 2017; Dagstuhl Seminar on Debating Technologies in 2015, and on Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments in 2016*. Applications of argumentation

mining include improved argument search and retrieval, summarizing complex decisions along with underlying grounds, counterarguments, and automatic argument construction for debating. Qualitative analysis of discussions, online comments, and newspaper articles can also provide user-friendly guidance tools for policy-makers to understand citizen needs and desires, as well as help business by understanding customers. Being a young research domain, argumentation mining approaches vary considerably, both in their problem definition and target domains (Lippi and Torroni, 2015a). Initial works focused on well-structured and edited text, such as legal text (Moens et al., 2007) or scientific publications (Teufel et al., 2009), and then expanded to other domains: persuasive essays (Stab and Gurevych, 2014b; Song et al., 2014; Persing and Ng, 2015), news articles (Sardianos et al., 2015), Wikipedia articles and blog posts (Biran and Rambow, 2011; Rosenthal and McKeown, 2012; Levy et al., 2014; Aharoni et al., 2014), and social media (Goudas et al., 2014; Boltuzic and Šnajder, 2014; Habernal and Gurevych, 2015; Swanson et al., 2015; Misra et al., 2015).

Given this variety of work on argumentation mining, a range of tasks have been addressed by several studies including identifying argumentative segments in text (Moens et al., 2007; Park and Cardie, 2014; Goudas et al., 2014; Levy et al., 2014; Swanson et al., 2015; Sardianos et al., 2015; Misra et al., 2017), classification of argument components (i.e., premise vs. conclusion) (Stab and Gurevych, 2014a; Nguyen and Litman, 2015), determining argumentation structure (i.e., support vs. attack) (Cabrio and Villata, 2012; Ghosh et al., 2014), and modeling similar arguments (Boltuzic and Šnajder, 2015; Misra et al., 2015, 2016). Some recent works go more in-depth than analyzing argument components or structure and target qualitative properties. For example, Persing and Ng (2015) scored the argument strength of persuasive essays,

(Tan et al., 2016; Lukin et al., 2017) studied belief change, Habernal and Gurevych (2016a) investigated a pair of arguments for convincingness, Wachsmuth et al. (2017) adapted PageRank to determine relevant arguments.

1.1 Argumentation in Social Media

The amount of time spent on various social networking sites is increasing every year. The computer networks are the modern Agora, a virtual place for exchange of comments, opinions, arguments, and all types of debates. The scale of participation and exposure to other views in the virtual community is much larger than in the past. These debates can be documented and revisited. People can use it to make more informed decisions about important arguments of an issue, and convey them to the policy-makers. Conversations in forums on topics such as health and education, public policy, social issues, as well as on consumer interests such as technical products, and entertainment are quite popular. These forums have a number of discussions going on, where each discussion is used for a conversation on a particular topic or subject area. Users typically take a stance and argue in support or opposition to the debate topic in an interactive discussion. A conversation thread is a place where a particular discussion about a given topic takes place. An initial proposition is stated by a single creator, then argued by supporting propositions or counter-propositions from other contributors (Groza et al., 2008). Online debates are very different from discussions in traditional media settings. Online debaters are highly involved and often use informal, emotional, and colorful language to make their points. Natural, informal debates exhibit a much broader range of argumentative styles (Oraby et al.,

2015). Indirect modes of communication, such as sarcasm and irony are also common. Previous work that models these different aspects include tasks such as stance classification (Walker et al., 2012b; Hasan and Ng, 2013b), sarcasm detection (Lukin and Walker, 2013; Ranade et al., 2013b; Oraby et al., 2016), and dis/agreement classification (Abbott et al., 2011; Misra and Walker, 2013). Though these tasks are highly useful in making decisions about an issue, a summary of the entire chain of reasoning is more informative. People need to go beyond such classifications to understand how statements are justified, the sources of disagreement, which facts are correct or which opinions are most relevant to them.

Consistency is another key factor in these debates. To assert their opinion, users sometimes ignore discourse coherence required for the logical integrity of the conversation (Ranade et al., 2013b). The arguments may not be well organized or structured, and may even deviate from the exact topic. Receiving so much information along with difficulty in judging the validity can lead to information overload - i.e., having far higher information than one's cognitive abilities to assimilate. In such situations, debate summarization approaches are helpful as users need not go through the entire debate. Text and opinion summarization of product reviews had been studied extensively in the past (Kim et al., 2011; Liu, 2015) but summarizing online ideological debates was a relatively an under-explored area. Opinion summarization tends to focus on product aspect identification and sentiment polarity classification. For example, an opinion summary about 'iPod' may discuss the common aspects such as 'battery life', 'camera quality', and 'screen size'. When it comes to summarizing general online political dialogic discussions on controversial policies or social issues, the challenges can be very different as the aspects may not be known in advance. Information is often partial and inconsistent concerning matters

of fact and opinion. The arguments that are under discussion for a particular topic may evolve over time, and people often repeat information that may not be veridical.

PostID:Turn
S1:1 That's the best you could do? There's the other three and here's a bonus one : ' To disarm the people is the best, most effectual way, to enslave them. ' - George Mason
S2:1 Is George Mason some kind of anarchist? It seems to me that the por-gun people are more and more pushing towards the 'every man for himself, the government can't protect us, we have to defend our property and way of life'. Shouldn't the government be improved instead? Besides, where I live we have very restrictive gun-laws and I ve never feeled 'enslaved' in any way. Ok, a comment to these 'myths '. First of all, where did you get these facts? Second of all, no matter how you try, now matter how much statistics you dig forth, no matter how many **references to the constitution** you do, you can never, ever convince me that harsh gun-restrictions would **not reduce the amount of murders and the amount of armed crime.** Besides, all people doesn't share your attitude of : **a good criminal is a dead criminal**, so those arguments get a bit meaningless.
S1:2 Nope. He was a great man who didn't believe in your cradle to grave liberalism. The founding fathers did not envision or want the all-seeing, all-knowing Federal government that you worship today. Nope. That government is best which governs the least, because its people discipline themselves. - Thomas Jefferson Restrictive? Where do you live? Britain? That would explain a lot. The people there are used to being 'subjects' rather than 'citizens'."When the people fear the government you have tyranny. When the government fears the people you have liberty" - Thomas Jefferson Again.. if you're so convinced that gun-control works, then why don't you place a sign in your front yard or on your door that reads :" DOING MY PART FOR GUN-CONTROL : THIS IS A GUN-FREE HOME! " You'd be wrong. MOST people are NOT coddlers of criminals like yourself.
S2:2 A sign like that would hardly be nessecary where I live. We have very restrictive gun-laws and hardly anyone owns a gun. And no, people does not discipline themselves. You might claim I have to much trust in the government, I would claim you have to much trust in your fellow man. **When you give freedom to the people you give freedom to criminals too.** I don't trust guns to only be used in 'selfdefense'. I want my government to uphold justice and to protect me from criminals. That's one of their purposes. And that is impossible without a certain amount of control over the citizens.

Figure 1.1: Excerpt from a Gun Control Dialogue-1.

In our approach, we are interested in finding nonveridical arguments, thus our meth-

ods find things that multiple people are saying, e.g. “Its cheaper to kill someone with the death penalty than to keep them in prison for life”, even though the actual evidence suggests that this is not true due to the costs of public defenders and the court system. The intuition behind our approach is that the there are similarities present in the way human beings reason and argue and these commonalities can be identified to extract arguments. For example, in debates concerning Gay Marriage, arguments are exchanged on both sides, with religious themes more typically appearing on one side, and economic benefits on the other. For a debate on Gun Control, we may find crime and safety aspect on one side and constitutional rights on the other. For instance, consider two back and forth dialogue exchanges in Figure 1.1 and Figure 1.2 for the topic Gun Control from the debate website 4Forums[1]. Though these two dialogue snippets are from two separate discussion threads with non-overlapping author pairs and different lexical realizations, they still present similar issues and concerns, as depicted in **bold**. Systematically extracting these arguments could provide an overview of the topics and arguments represented in the debate.

1.2 Thesis Objective

Debate websites such as *Procon.org* list arguments for controversial issues. Figure 1.3 shows a sample for the topic of Gay Marriage. The arguments on various sides of an issue are organized as top “pro” and “con” points. Each point contains the main proposition, reasons to support a viewpoint, and counter-arguments for the opposing views. This set is manually curated by editors to summarize the main points of the debate. A deep understanding and analysis

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

PostID:Turn
S1:1 To reduce the numer of shootings associated with crime of course. Your strident claim that **"Gun Control is Not Crime Control!"** is specious. Countries with few guns still have crime. Whoever argued differently? My argument is more like " **gun control is shooting control"**. Guess what? Countries with fewer guns, especially handguns, have fewer people shot. Short of banning, how can we reduce the number of shootings we have every year? That's my interest. I'm not looking to **rewrite the Constitution or disarm law-abiding citizens.** Stop mischaracterizing my motives.
S2:1 It's a little hard to stop when you continue to support ideas that don't work, and **have proven unconstitutional and thus illegal.** Who exactly are the people responsible for most of the shootings that take place in America? I would suggest trying to find out anything about their backgrounds, see what records they have with the police. Are these shooters individuals with records of criminal actions, felony arrests, parole violations, drug dependency, etc? Or are the people responsible for the shootings everyday individuals who have clean records and absolutely no indication that something might be wrong. We already have the **link showing that 80% of America's crimes are directly connected to gang members with criminal records.** To me that suggests something, it suggests that the **ones who're doing the most killing are the ones who are breaking the laws and paying no heed to the various rules and regulations that you demand we follow.** So unless you have conclusive evidence that the vast majority of shootings in the USA are directly the fault of the law-abiding gun owners, all your suggestions amount to saying that you need more laws to deal with those who break the laws.
S1:2- Your high school diploma doesn't qualify you as a legal scholar. Assuming you completed high school. I'm in no position to make demands. And yes, laws only affect those that abide by them. Gun laws, traffic laws, theft laws. Laws do give society a way to punish the transgressors if they can be identified and apprehended. This whole argument that **gun laws only affect the law-abiding is stupidity** itself IMO. Maybe just better laws, but the idea of more laws seems to work with lots of other societal problems. Ask a deadbeat dad.

Figure 1.2: Excerpt from a Gun Control Dialogue-2.

of beliefs and ideas is required to create such argument summaries. To enable computational systems to explain and justify their choices is challenging, and the current work takes a step in that direction. We envision a system that supports discovery, extraction, and eventually grouping of similar arguments that gives a concise summary of the vast variety of viewpoints across threaded discussions. As a process, argumentation involves to support and defend one's claim or

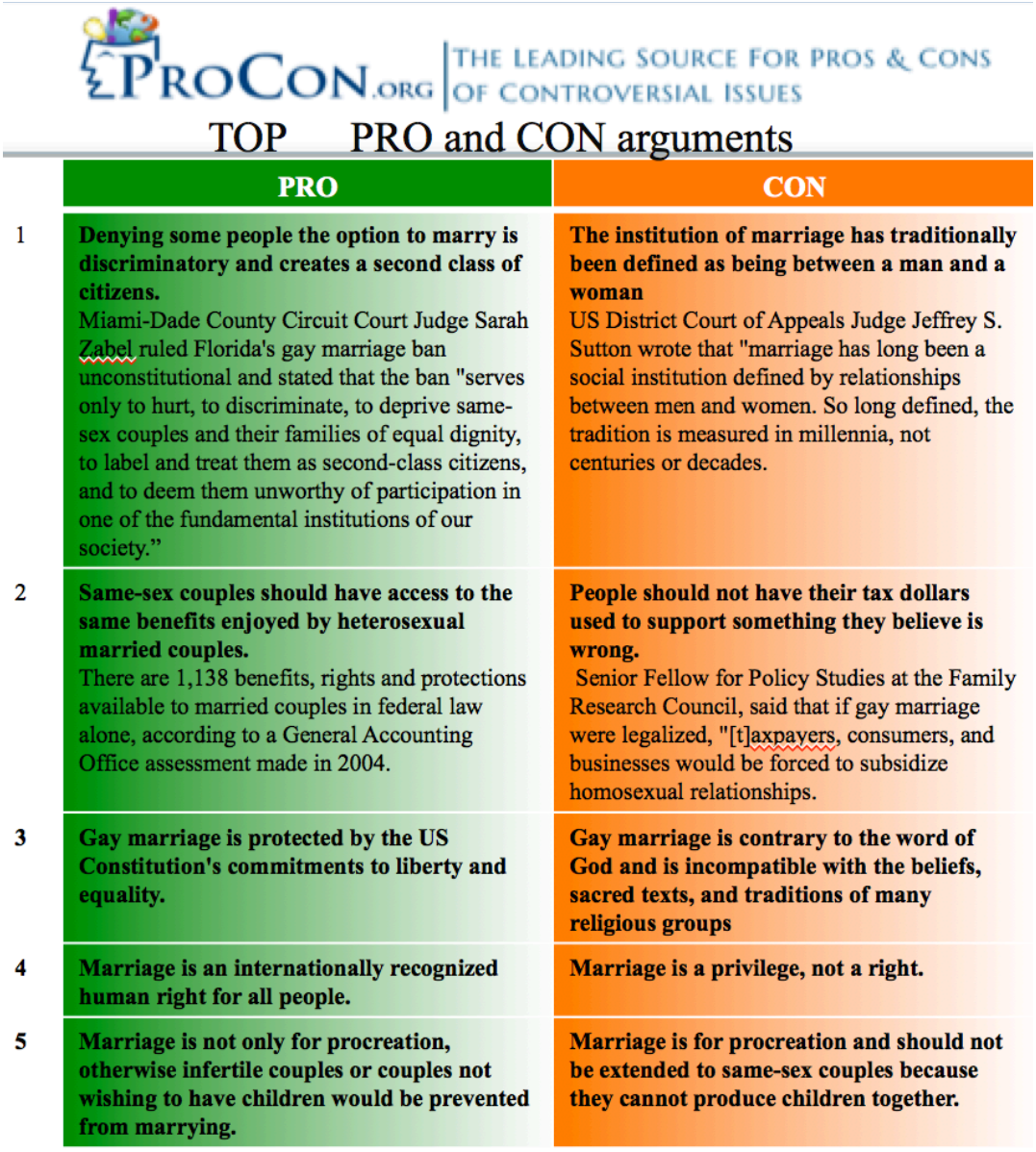


Figure 1.3: A sample Pro Con Argument Summary for Gay Marriage

refute the opponent's and hence evolves with a dialogue among the contributors. The proposed thesis aims to automatically extract meaningful information from these dialogues that reflect the arguments exchanged in a debate that can be used to summarize the issues under discussion.

Bearing the challenges presented above in the field of argument mining, this research will investigate the following research questions:

- **Q1**: What are the central propositions associated with different stances on an issue?
- **Q2**: What are the most important arguments that can be used to summarize an argumentative dialogue?
- **Q3**: Can we develop computational tools to extract the most important arguments in an argumentative dialogue in an online debate?
- **Q4**: What are the abstract objects under discussion in an online debate?

1.2.1 Argument Dialogue Corpus

An argument mining pipeline with machine learning and AI techniques typically involves building a corpus, defined as

A systematic collection of naturally occurring texts of both written and spoken language. It is systematic for two reasons: (1) the structure and contents of the corpus follow certain extralinguistic principles (the principles on the basis by which the texts included were chosen); (2) the exact composition of the corpus including the sampling process is available (Nesselhauf, 2004). The corpus is manually annotated, relevant to the task, and contains training instances for a predictor. Inter-annotator agreement measures are typically employed to demonstrate consistency and measure the quality of the obtained annotations provided by multiple annotators. This is a hard task for the argument mining problems as fuzziness of argument components, their boundaries, and how they relate to each other can be quite complicated and controversial

even for humans (Mochales and Ieven, 2009). Another consideration is the cost and the time required for annotations, limiting the amount of data that can be annotated for training.

Since there are two sides to an argument, the purpose and its characteristics may be well understood when given in dialogue. For building a corpus for our research, we utilized the publically available Internet Argument Corpus, a collection of corpora for research in a political debate (Walker et al., 2012c). To create dialogue chains between a pair of authors, we selected a subset of the corpus from an internet forum called 4forums.com. The website is tailored to US audience, and some US-centric topics are brought up frequently. People initiate discussions (threads) and respond to others posts. Each thread has a tree-like dialogue structure. Figure 1.2 shows a snippet of a dialogue chain depicting alternate turns. Our basis of selection of these dialogues was the number of turns in a dialogue, author of the dialogue, and the word length focusing on the topics of *Gun control, Gay Marriage,* and *Abortion*. Having these linguistic features as a stimulus to our task of dialogue selection led to an interesting and diverse collection.

With the dialogue corpus resource in hand, we investigate our first research question Q1 about central propositions for a given dialogue and define the term central propositions as:

Central Propositions: Debate propositions that are most important in a dialogue. These propositions in a dialogic debate connect to the subject matter of the debate in general, including arguments for or against a position as well as data and/or anecdotes used to justify a position. Central propositions exclude factors related to the dialogic context itself, such as insults toward the other party, phatic communication, and metadialogic conversation about the dialogue itself.

The next task was to label the central propositions in a dialogue, but how? No one

really knows how to get the central propositions for a given dialogue. The notion of important argument is subjective to begin with. Hence, we do not ask annotators to select central propositions for a dialogue. Instead, we use salience as a proxy for centrality. An important assumption towards seeking central propositions serves as a starting point for this thesis and can be stated as:

Central propositions in a dialogue are exactly those arguments that people find most salient in a dialogue.

We adapt the Pyramid method of summary evaluation to find salience (Nenkova and Passonneau, 2004; Nenkova et al., 2007). In this method, the content units are organized in tiers based on their frequency in multiple human summaries. The intuition is that higher the frequency, more important it is, and thus belongs to higher tiers in the pyramid. We hope that pyramid method is a more objective and reliable way to find central propositions as compared to what has been done in related work (spam filtering by hand or first getting rid of non-argumentative material, as defined by a theory). For instance, Aharoni et al. (2014) introduced an argument structure to label claims and evidence from 33 controversial topics from the website Idebate.org, and reported an inter-annotator agreement (IAA) of Cohen's kappa 0.39. The high degree of subjectivity in the argument and claim recognition task leads to low IAA and reliability. Instead, we use pyramid method for our annotation task. This reduces the subjectivity in the task, thus improving the overall reliability.

We used human summarization as a probe to discover saliency and conducted a series of summary collection experiments using crowdsourcing. Multiple summaries were collected for each dialogue. We trained undergraduate linguistics to perform pyramid annotation of sum-

maries, giving a ranked set of summary labels. Repeated elements in the summaries end up in the higher tiers as the central propositions for individual dialogues.

Our first hypothesis $\mathbf{H_1}$ addresses the next research question **Q2** about most important arguments in the dialogue and can be stated as:

$\mathbf{H_1}$: It is possible to map the ranked labels from the pyramid annotation back to the dialogue with good reliability.

1.2.2 Argument Extraction and Quality

In Social media argumentative discussions, subjective perspectives or differences of opinion are exchanged on a much larger scale with a broader audience as compared to formal settings. Anyone can initiate and participate in that conversation. Arguments are made spontaneously, are implicit and even wordy, a characteristic of natural dialogue. As every conversation can be significantly different regarding context or purpose, every conversation thread may not have the same quality level or the number of arguments. We explored two different methods that help to assess the argument quality in social media argumentative discussions:

1. A definition of important argument based on the text segments that a majority of readers are attentive to. This is captured by summarizing a dialogue and then linking each sentence to the best-mapped label in the pyramid structure. Each sentence gets a rank based on the rank of the pyramid label it gets mapped to indicating the quality. Higher the rank, better is the argument quality.

2. Improved upon the argument quality definition from Swanson et al. (2015). The Argu-

ment Quality (AQ) is based on the notion of argument clarity and is used to find text segments that clearly express an argument facet.

Both these methods can be used to extract high quality arguments. The first method ranks sentences within a dialogue using pyramid scheme. The second method gives a ranking across all conversations based on contextual knowledge and inference required to determine if the sentence expressed a clear argument for a given topic.

The next hypothesis addresses the research question **Q3** of automatic extraction of arguments from a dialogue. We will investigate the effectiveness of linguistic and contextual features and evaluate the prediction output of our computational model using the training data obtained from $\mathbf{H_1}$.

The computational model by Palau and Moens (2009) captured the general characteristics of argumentation using features such as punctuations, number of tokens, number of sub-clauses, and modal verbs for legal domain. Ranade et al. (2013a) showed that features related to topic and sentiment information are useful for ranking debate sentences. (Levy et al., 2014; Rinott et al., 2015) showed that the discussion topic is an important feature when distinguishing relevant claims/evidence from the irrelevant. Topic features such as cosine similarity using Word2Vec (Mikolov et al., 2013b) and WordNet (Miller, 1995) were employed. Using surrounding text as the context-rich representation of the input for feature extraction has been studied by Biran and Rambow (2011). They considered the discourse relations using Rhetorical Structure Theory (Mann and Thompson, 1988) within the context segment to characterize contextual features, while other works enrich the feature space by extracting features separately from surrounding text spans. Bex and Walton (2016) discuss the role of dialogical context in

argumentation, where the context can be determined by speech acts. Based on these observations, we state our next hypothesis as:

H_2: Adding linguistic and contextual features improves argument extraction.

We used features based on topic, context and overall sentence structure. Topic features are represented using words from the lexical resource named Linguistic Inquiry Word Count (LIWC) (Pennebaker et al., 2001). Features that are characteristic of dialogical contexts are recognized using the dialogue act classifier from NLTK (Loper and Bird, 2002). We used Readability metrics to determine text quality. The readability score is calculated using surface characteristics such as the average number of syllables or the average number of words per sentence. These measures approximate the difficulty levels of the text, for example, vocabulary or syntax.

We then formulated a binary classification task to identify argumentative sentences and conducted experiments to investigate which feature and classification models are effective, using F-score, a standard evaluation metric based on precision and recall. We also examined the performance of the traditional document summarization approaches using several off the shelf summarizers, as our baselines. The final experimental setup runs a statistical analysis to determine which feature combinations make a significant improvement.

1.2.3 Argument Facet Similarity

A computational model for argument extraction gives a ranked set of arguments for a particular dialogue, giving a large collection of arguments, but how do we determine the central concepts across dialogues which drive the discussion? Concepts which manifest in a majority of

dialogues represent important aspects of the debate. It would be more useful to organize these important arguments according to these concepts and present this information to the reader, leading us to the last research question **Q4**: the abstract objects under discussion in an online debate.

Predefined Arguments	**Stance**
It is discriminatory to refuse gay couples the right to marry.	Pro
Gay couples should be able to take advantage of the fiscal and legal benefits of marriage.	Pro
Marriage is about more than procreation, therefore gay couples should not be denied the right to marry due to their biology.	Pro
Gay couples can declare their union without resort to marriage.	Con
Gay marriage undermines the institution of marriage, leading to an increase in out of wedlock births and divorce rates.	Con
Major world religions are against gay marriages.	Con
Marriage should be between a man and a woman.	Con

Figure 1.4: Pre-defined Argument list from Boltuzic and Šnajder (2014).

Boltuzic and Šnajder (2014) introduced an argument recognition task of mapping user comments to a predefined set of arguments. These predefined arguments sets, also called as **'Aspects'** or **'Frames'** or **'Argument tags'** in related work are similar to our abstract objects in research question **Q4** as they highlight the issues under discussion in a debate.

However, the authors assume that this topic-dependent set of arguments has been prepared in advance. The Corpus CompArg[2] contains user comments compiled from the website Procon.org, and predefined argument set from Idebate.org. The task is framed as a multi-class classification problem using entailment-based features and an off the shelf tool for Seman-

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

tic Textual Similarity (STS), defined as the degree of semantic equivalence between two texts (Agirre et al., 2012). Boltuzic and Šnajder (2015) used only the argumentative sentences dataset from Hasan and Ng (2014) and presented an unsupervised approach to recognize these aspects. They grouped similar arguments into clusters using hierarchical clustering based on STS. Naderi and Hirst (2016) used a similar term, namely **'Frame'**, to refer to this predefined argument list. The authors used word and sentence vector representations to measure STS between the statements and the frames in parliamentary speeches.

A principal problem with the approach of the predefined argument list is that, in general, these aspects of a given topic are not known in advance. A change in the political or social environment may change these aspects. These are multifaceted issues where facets evolve with changing scenarios.

Much of the previous work concerned with modeling aspects uses a notion of similarity based on STS or entailment. Similarity can have many levels of granularity such as lexical, structural, or conceptual. The features and relations establishing similarity could be application-specific. For argumentation domain, we have people on same sides of an issue discussing the same aspect but different stance. To measure argument similarity, we first defined the term argument facet as:

Facet: A facet is a low-level issue that often reoccurs in many arguments in support of the author's stance or in attacking the other author's position. There are many ways to argue for your stance on a topic. For example, in a discussion about the death penalty, you may argue in favor of it by claiming that it deters crime. Alternatively, you may argue in favor of the death penalty because it gives victims of the crimes closure. On the other hand, you may argue against the

death penalty because some innocent people will be wrongfully executed or because it is a cruel and unusual punishment. Each of these specific points is a facet.

Our argument similarity measure, which we call as Argument Facet Similarity (AFS) reflects both the aspects and the stance. Two tasks similar to AFS are Semantic Textual Similarity and Entailment (TE). TE is a binary decision and directional while AFS is graded. For instance, consider these two example pairs for the AFS annotation for the Gun Control topic.

- There is no reason law-abiding citizens need to defend themselves as they are untrained, do not have full knowledge of the law and eventually cause more harm than good by owning such weapons.

- Gee, maybe if the law abiding had guns, they could defend themselves?

These two statements are discussing the same facet that may be stated as, "guns and self defense by law abiding citizens". One argument is in favor while the other is against. These two statements contradict each other, none entails the other and may not be equivalent as per STS. However, AFS would label them as roughly equivalent as both the arguments are discussing the same facet but have opposite stance. This indicates that the task we want to address, AFS, is different from other textual similarity based problems, like STS, and calls for a need for defining this new task for identifying whether two sentences, like the ones shown above, discuss the same argument facets. For this purpose, we used a 6 point graded scale to characterize sentence similarities and collected annotations for this task using crowdscourcing. The level of agreement between workers would indicate the quality of the annotations obtained, and whether the definition can be transformed to a computational model that can predict argument similarity.

We formulated AFS as a supervised regression task and showed that the prediction output of our proposed AFS model improved the output obtained using STS. We used a combination of features such as ROUGE (Lin, 2004), Ngramoverlap, Word2Vec representations and LIWC to predict AFS. We state our next hypothesis as:

H_3: The proposed features yield better performance on AFS than those used for STS.

1.3 Overview of Contributions

The work presented in this thesis is focused on inducing the recurring **facets** in a particular topic domain via supervised learning over several dialogic interactions. Many different threads of recent research on argument mining have a strong parallel with this goal. However, they differ from our approach as they all assume a finite set of topic-specific labels that are determined in some form by the researchers themselves. In contrast, we seek to uncover popular facets via clustering similar central propositions across the dialogues. This research contributes to the current state of the knowledge in the following ways:

- The first corpus of summaries, pyramids, and central propositions of spontaneously-produced written dialogue of high social and political importance. The pyramid labels back-linked to the original dialogue sentences with high reliability among human annotators used as "ground truth" for training and testing the summarization framework.

- A pyramid based summarization to arrive at a well-motivated and theoretically grounded definition of important argument that also reflects the argument quality in social media

dialogue, which is a novel method for ranking arguments in conversational data.

- A computational method for an end-to-end argument summarization of dialogue exchanges from social media debates significantly beating the traditional summarization baselines. We observed that using contextual and linguistic information with support vector machines (SVM) and neural network-based models achieved better performance for argument summarization. Both topic-specific and topic-neutral features provided useful knowledge. LIWC categories were helpful in conveying the topic of the communication. Word categories such as death, money, and religion reveal the topical content of the input. On the other hand, the complexity of a given text, as measured by Readability, can be used to recognize important arguments in a topic-independent way. We further find that adding contextual knowledge derived from LIWC categories of previous sentences and coreference resolution enhances the model's predictive power.

- For the first time, we presented an application of the pyramid summarization scheme to the task of **Facet induction;** and introduced a new task of **Argument Facet Similarity (AFS)** aimed at identifying **Facets** across opinionated dialogues. A new corpus of paired arguments with human-annotated AFS scores using crowdsourcing. The average score over all the annotators correlated at a high value of 0.7 with our gold standard annotation indicating that the AFS similarity task is well-defined, and understandable by minimally trained annotators on Mechanical Turk. Our experimental results showed that we can identify AFS with a higher correlation as opposed to a baseline Semantic Textual Similarity (STS) using a combination of features based on distributional similarity, Word2Vec,

LIWC, and dependency parser. Our results also demonstrated that using concatenation for learning similarity with vector representations works much better than reducing a pair of vectors to a single score using cosine similarity.

1.4 Thesis outline

The remaining thesis is organized as follows: Chapter 2 discusses the related literature. We start by presenting the various corpora and the subtasks in argument mining. Since we compare our work to STS, we present the work related to STS. We further investigate previous work in summarization and describe how it is limited for the current task. Chapter 3 describes in detail the argumentative dialogue corpus. We start with the dialogue collection experiments from IAC. This is followed by summarization, pyramid annotations, and mapping pyramid labels to dialogue text. Chapter 4 describes the computational method to summarize dialogic arguments from the above corpus. It describes the machine learning framework, the performance of baseline summarizers, the experimental set up using LSTM and SVM models. We demonstrate how adding contextual and linguistic features improves the performance of LSTM and SVM. Chapter 5 begins with a description of the AFS task, guidelines for the crowdsourced corpus annotation (includes the number of workers per task, annotator agreements and task quality). We run AFS on two different types of argument sets. The first one contains data from the pyramid annotations. For the second set of experiments, we use the argument corpus from Swanson et al. (2015). Chapter 6 concludes with a summary of our contributions, limitations, and future work.

Chapter 2

Previous Work

Argumentation is an interactive process and frequently occurs in conversations. It involves making and explaining the choices and the decisions we make. Each choice may have its pros and cons. With the use of social networking sites such as Twitter, Facebook, Forums, and Blogs, the options are nearly unlimited. Through dissemination of thoughts, opinions and extensive exchange of arguments on the social media, we now contribute to popular opinion. However, the massive amount of information available today is overwhelming to our brain and its ability to efficiently separate relevant information from the irrelevant. Concise, easy to read summaries of this argumentative text can help us to efficiently absorb and comprehend this content.

The overarching goal of this thesis is to develop techniques to recognize the specific arguments and counterarguments people tend to advance, and group them across discussions as **Facets**. This entails the sub-tasks:

(i) Extract and summarize important arguments.

(ii) Discover similar repeated argument aspects that appear across multiple dialogues on a topic.

Moens et al. (2007) introduced argument mining as a research area aimed at detection of all the arguments involved in the argumentation process, their structure, argumentative relationships between their propositions, and the interactions between them. There are many formalisms for argument description with multiple framings (e.g. separating argumentative content, argument component identification, relation identification, argument similarity, and argument summarization).

Summarization is a widely researched problem in natural language processing, which entails eliminating any superfluous text and fillers, and identifying essential points that preserve the core meaning of the original input. Many summarization tasks have been studied in the past in multiple domains including single document, multi document, and conversational speech.

The framework and the methods in this research lie at the intersection of semantic textual similarity (STS), summarization and argument mining, which we discuss and compare with our work. We first describe STS in Section 2.1. Next, we present the various sub-tasks and existing corpora in argument mining in Section 2.2 . This is followed by related work in summarization in Section 2.3. Finally, the chapter concludes in Section 2.4 by pointing out the limitations in the existing literature and affirming the need for a new approach to address the goal of **facet** identification.

2.1 Semantic Textual Similarity

Agirre et al. (2012) presented a pilot shared task on Semantic Textual Similarity (STS)

Compare Two Similar Sentences

Score how similar two sentences are to each other according to the following scale. The sentences are

(5) Completely equivalent, as they mean the same thing.

(4) Mostly equivalent, but some unimportant details differ.

(3) Roughly equivalent, but some important information differs/missing.

(2) Not equivalent, but share some details.

(1) Not equivalent, but are on same topic.

(0) On different topics.

Figure 2.1: Semantic Textual Similarity Scale.

as a part of the *First Joint Conference on Lexical and Computational Semantics (SemEval) 2012*. It measured the degree of semantic equivalence between two sentences and was based on the symmetric graded equivalence between a pair of texts. The similarity was rated on a 0-5 scale (low to high similarity) by human judges using Amazon Mechanical Turk. A value of 0 meant unrelated, and 5 was a complete semantic equivalence. See the definition given in Figure 2.1. The computational approaches usually entailed recognizing and aligning semantically similar or related words across the pair, followed by an aggregated overall similarity score (Majumder et al., 2016). Pearson correlation between predicted and human similarity scores was used to determine the system's performance. Most of the top performing systems for the STS shared task employed a vast range of features, including lexical similarity, Latent semantic analysis (LSA) word similarity, and WordNet knowledge (Han et al., 2013). Since then, the STS shared task has been held annually leading to new approaches that model sentence level similarity.

STS is related to Textual Entailment (TE) as both are based on semantic equivalence. (TE) is asymmetric and binary, while STS is symmetric and graded.

2.2 Argument Mining

Although several corpora have been annotated with the goal of identifying arguments and relations, the annotation guidelines and schemes differ due to many theories of argumentation that might apply to a task (Jackson and Jacobs, 1980; Reed and Rowe, 2004; Walton et al., 2008; Gilbert, 1997; Toulmin, 1958; Dung, 1995). One single definition of argument structure may not work for every NLP task. What constitutes an argument may depend upon whether we are in a literary domain, legal domain, or a social media discussion. Factors such as writer intent and reader knowledge also have an impact, and hence argument recognition is a complex problem. It has been addressed at several levels of granularity: clause, sentence, within and across documents. Therefore, most of the argument corpora are task and domain specific. Next, we survey the argument subtasks performed and the associated corpora.

- **Argument Extraction and Quality:** Section 2.2.1.
- **Argument Relations and Tagging:** Section 2.2.2.

2.2.1 Argument Extraction and Quality

As already mentioned, several theories of argumentation exist, and each of them makes explicit assumptions about the argument structure, generation, identification, and quality evaluation. However, the definition of an argument is more controversial. Because there are

many proposals for argumentation, it is impossible to give a single, formal, universally accepted definition of argument or argument quality. There are some methods that explicitly define and measure the 'argument quality' using measures such as convincingness or persuasion while others assess quality largely as a byproduct of the argument definition, structure or schemes. Below, we describe the related work considering both 'argument definition and structure' and 'explicit quality dimensions'.

Identification, Segmentation, and Structure: These are mostly formulated as classification problems and include tasks like separating argumentative from non-argumentative text units, identifying argument components (e.g., claims, premises, conclusion), and classification using argument schemes.

Argument identification is the task of recognizing or extracting segments of text that can be classified as argumentative. Argument extraction is a quite complex procedure, and in many cases, it may be difficult for even humans to reliably annotate an argument segment, thus making it even more challenging for machines. Splitting of a given text segment into its argumentative segments and their non-argumentative counterparts is a first important step and is based on the argument definition followed.

Conceptually, an argument unit may span a clause, a complete sentence, multiple sentences, or something in between. The size of the unit depends on the domain of an argumentative text, but can also vary within a text. Formal texts, such as legal documents, usually have longer sentences and usually present premise and conclusion in subordinate sentences or independent sentences instead of subclauses and hence a minimum two proposition requirement

for an argument seems reasonable for them (Palau and Moens, 2009). Many works classify argument at sentence level (Goudas et al., 2014; Habernal and Gurevych, 2015). Argument definition by Stab and Gurevych (2014b) includes a claim that is supported or attacked by at least one premise and spans multiple sentences. (Levy et al., 2014; Lippi and Torroni, 2015b) label exact claim boundaries, which do not necessarily match a whole sentence or even a clause in the text. In contrast, dialogue or informal text in social media may contain shorter sentences where conclusion and premise can be together in a single sentence or can be implicit (i.e., enthymemes). Habernal and Gurevych (2017) showed, for instance, that 48% of the claims in user-generated web discourse are implicit.

There is no consensus yet on an annotation scheme for argument components, or on the minimal textual units to be annotated. A number of frameworks including lexical, syntactic, semantic, and discourse structure have been used to guide the selection of an argument.

Palau and Moens (2009) presented one of the first computational approaches by framing argument mining as a three-step process: (i) detection of the argumentative proposition, (ii) proposition function classification, and (iii) detection of the argumentation structure. At least two propositions are required to classify a text unit as argumentative. They work on the legal domain in the Araucaria corpus (Reed and Rowe, 2004) and the European Court of Human Rights (ECHR) corpus (Palau and Moens, 2008). The Araucaria corpus contains 641 documents, 1899 argumentative sentences, and 827 sentences without arguments. It includes data from various sources such as newspapers, parliamentary records, judicial summaries, and discussion boards. Arguments are annotated in an XML-based format called "AML" (Argument Markup Language). The ECHR contains 47 documents, 1449 non-argumentative sentences,

763 premises, and the number of conclusions is 304. Palau and Moens (2009) reported an accuracy of 73.75% when detecting arguments in the Araucaria corpus using a multinomial Naive Bayes classifier and a maximum entropy model. The accuracy increases to 80% when the task is performed on the ECHR corpus. In the second step, a support vector machine is used to classify each argumentative proposition found into a premise or a conclusion, with an F-measure of 68.12% and 74.07% respectively. The classifiers employ a rich set of features including n-grams, POS tags, list of keywords indicative of argumentative discourse, and parse trees. To determine the argumentation structure, manually derived rules tailored to a legal domain are grouped into a context-free grammar giving an accuracy of 60%.

Rooney et al. (2012) applied kernel methods and classified each proposition as a claim, a premise or non-argumentative, in Araucaria and reported an overall accuracy of 65%.

In subsequent work on the Araucaria, Feng and Hirst (2011) proposed an approach for identifying the five most commonly used argumentation schemes, which are templates for various kinds of arguments (e.g., argument from example, argument from cause to effect, practical reasoning, argument from consequences and argument from verbal classification) using Walton et al. (2008)'s scheme set. They experimented with different combinations of general and scheme-specific features. However, reliable identification of argument components is a prerequisite for the features used. Depending upon the particular scheme, they report an accuracy between 63-91% for one-against-others classification and 80-94% for pairwise classification.

Florou et al. (2013) constructed a corpus of Greek text segments and annotated them as argumentative or not. As only 69 segments were argumentative, 345 positive examples were obtained by oversampling. 332 samples were manually annotated as non-argumentative giving

a final corpus of 677 segments. They employed decision trees, used morpho-syntactic features such as mood and tense of the verb along with discourse markers, and achieved an F1 score of 0.76.

Goudas et al. (2014) and Sardianos et al. (2015) performed binary classification to identify argumentative sentences followed by a structured prediction task to identify clause-level argument segments (premises and claims) using BIO encoding for Greek language texts. Goudas et al. (2014) annotated 16,000 sentences from 204 social media documents about renewable energy and found only 760 of them to be argumentative. Employing a rich feature set, they reported an accuracy of 0.774 using logistic regression for argument identification. Boundaries of argument components were determined with a conditional random field model that achieved an F1 score of 0.42. Sardianos et al. (2015) annotated a dataset of 1191 argument segments from a variety of sources such as news, blogs, and social media. The inter-annotator agreement using the F1 measure was 75.50%. With part-of-speech tags, and distributed representations of words, their best system achieved an accuracy of 0.32 for the task of boundary detection of a text segment that encloses a claim or a premise.

Biran and Rambow (2011) investigated the use of discourse relations for identifying justifications for subjective claims in interactive written dialogues. 309 blog threads from LiveJournal.com were annotated for claims and premises (justifications) with an inter-rater agreement of kappa value 0.69. The authors show that the discourse contribution from multiple discourse relations in the justifications characterizes the argumentation in support of a given claim. Because the instances being classified are a pair of textual units, features usually involve information from both elements (i.e., source and target) of the pair (e.g., word pair, discourse

indicators in source and target) and the relative location.

Rosenthal and McKeown (2012) examined methods for identifying opinionated claims expressing a belief. A corpus of 285 blog posts from LiveJournal.com and 51 Wikipedia discussion pages was annotated by two annotators at the sentence level to identify claims. The average agreement was reported using kappa, 0.50 on LiveJournal, and 0.557 on Wikipedia discussion forums. For classification, features based on sentiment and committed belief (Prabhakaran et al., 2010) were used along with n-grams and POS tags. Whereas sentiment and POS tags were more important for LiveJournal, committed belief and n-grams features worked better for Wikipedia discussion forums.

Roitman et al. (2016) described a claim-oriented document retrieval approach to extract relevant Wikipedia articles that support or contest a given controversial topic. A two-step retrieval approach was used first to create an initial pool of articles that were relevant to the topic using a state of the art information retrieval method. At the second step, articles in the initial pool were re-ranked according to their potential to contain as many relevant claims as possible using a lexicon based set of handcrafted claim discovery features (e.g., controversy related terms such as dispute and criticism, or special Wikipedia annotations that indicate a controversial content). An overall improvement of 10% in claim recall was achieved by the re-ranking.

Aharoni et al. (2014) introduced an argument structure comprising only two components: a claim, and the associated supporting evidence used to support or contest a controversial topic. 33 controversial topics were selected from the website Idebate.org to create a dataset for context-dependent claim detection. 1,392 related claims from 321 Wikipedia articles were an-

notated by trained annotators with an inter-annotator agreement of Cohen's kappa 0.39. Levy et al. (2014) proposed a context-dependent claim detection on this corpus, while a context-independent approach was applied in Lippi and Torroni (2015b) to the same dataset. Levy et al. (2014) identified the most probable claim and the exact boundaries of the claim within the sentence for a given topic. The system pipelines a cascade of classifiers, each trained by exploiting a set of highly engineered features, incorporating knowledge coming from lexical databases such as WordNet, sentiment analysis tools, and a module to score sentence subjectivity. An initial classifier detects sentences that contain a relevant claim for the given topic. To detect the boundaries of a claim, each sentence is segmented into different sub-sentences, and a logistic regression classifier selects the most probable claim from a list of several candidate sub-sentences. The last component ranks all the identified claims using another logistic regression classifier giving the most relevant claims for the given topic. Lippi and Torroni (2015b) exploited structured parsing to detect claims without using any contextual information. An SVM classifier was trained using a tree kernel that measures the similarity between two sentence-level parse trees. An F1 of 0.168 is reported, a slight improvement from an F1 of 0.16 by Levy et al. (2014).

Rinott et al. (2015) expanded the dataset to 58 topics, 2,294 claims, and 4,960 associated evidences. Each evidence is further classified as study (quantitative analysis), expert (testimony by a person) or anecdotal (specific events). The evidence can include several sentences. All consecutive segments up to three sentences within a paragraph are considered as evidence candidates. Information from the topic of the debate and a given claim are used in order to rank the retrieved evidence. The proposed system achieved macro averaged mean reciprocal ranks

between .03 and 0.20 depending on the type of evidence.

Another corpus from user-generated web content following a variant of the Toulmin (1958) model was developed by Habernal et al. (2014) to model arguments and their components. This dataset included 990 English comments to articles, blog posts, and forum posts. 524 samples were labeled as argumentative by three annotators with an inter-annotator agreement of Cohen's kappa 0.59. A final smaller corpus of 345 samples was annotated with finer-grained tags (claims, premises, backings, rebuttals, and refutations). The inter-annotator agreement was measured using Krippendorf's alpha. It ranged from 34.6 (distinguishing all classes and non-argumentative) to 42.4 (distinguishing between premises, claims, and non-argumentative). Habernal and Gurevych (2015) performed component identification on this dataset. The task was formulated as a sequence labeling task, and a BIO (beginning, inside, outside) tagset was used to identify the boundaries of components. The authors experimented with several feature sets including lexical, syntactic, topic, sentiment, discourse, and unsupervised embeddings. For separating argumentative from non-argumentative documents, the system achieved an F1 score of 0.69. However, the task of boundary recognition of various types of argument components achieved a comparatively low score in the range of 0.31-0.40.

Our Work. Social media arguments do not follow any specific rules, and most of the messages are simple and may even lack proper syntax or spelling. As opposed to formalized debates and structured documents, social media argument is informal, vague, and verbose. Inferring the unsaid information is natural for humans as we have access to a vast array of common sense knowledge that includes general information about the world. Additionally, conversants in a

dialogue exchange have some sort of common ground. This world knowledge is used by humans to determine the context in which an utterance or argument is made in a dialogue and to determine its meaning. Computers, however, are not privy to that same world knowledge. Therefore, we need to reflect the text properties and granularity to define an argument in social media dialogue. We propose an alternative view and define an argument as a single sentence that clearly expresses an argument facet in a dialogue and we are interested in identifying identical but differently expressed argument facets across dialogues (Misra et al., 2015, 2017).

Argument Quality: Substantial research has been dedicated to argument evaluation theory to determine what constitutes a good argument. Is a strong argument an effective argument which gains the adherence of the audience, or is it a valid argument, which ought to persuade the audience (Perelman and Olbrechts-Tyteca, 1969)? Walton (1989) discusses the strengths and weakness of arguments in informal logic, while as per Johnson and Blair (2006) relevance, acceptability, and sufficiency (RAS) are appropriate criteria for evaluating arguments in the sense of reasons offered in support of a claim. Correlation between argument quality perception and actual persuasiveness is studied by Hoeken (2001). Siegel (2015) states that good-making features of arguments vary by context and purpose. The criteria for a good argument in a scientific context differs in some respects from what makes an argument good in a court of law, a friendly conversation, or a public debate. People argue for different purposes such as persuasion, the epistemic development, or consensus, with each having its criteria for goodness. Blair (2004) construe argument evaluation relative to its use (persuasion, inquiry, collaboration). As a result of these varying goals of argumentation, quality is viewed from different perspectives such

as the logic of arguments, style and rhetorical effect of argumentation, or its contribution to a discussion. A convincing argument from the point of view of an argumentation theorist is not always convincing from a layperson's point of view (Van Eemeren, 1995). Since neither a general consensus on argumentation quality nor a clear measure of its evaluation exists, practical assessment approaches are based on simple measurable quantities rather than on theoretical framework (Habernal and Gurevych, 2016b), for example, how many premises support a claim (Stegmann et al., 2012), or the complexity of the analyzed argument scheme (Garcia-Mila et al., 2013). Next, we describe a few attempts in computational argumentation that go beyond analyzing argument structure and role to evaluate argument quality. Measures such as relevance, persuasion, convincingness, and context have been used.

Habernal and Gurevych (2016b,a) empirically compared informal social media arguments in terms of convincingness. A dataset is annotated to determine which argument is more convincing, along with the reasons given by annotators for its convincingness. These reasons were used to derive a hierarchical annotation scheme. 9111 argument pairs were labeled with 17 reason labels, including (attacking/abusive language/grammar issues/no credible evidence/no facts/sticks to the topic). The authors tested different models for predicting convincingness: An SVM with engineered linguistic features, and a BLSTM. Both the models performed similarly and obtained an accuracy in the range 0.76-0.78.

Wachsmuth et al. (2017) adapted the PageRank algorithm to objectively assess the relevance of an argument at web scale using commonly referred to arguments for a given conclusion. Expert annotators ranked the arguments and achieved an average Kendall's correlation τ of 0.36. Ranking based on page rank achieved a value of 0.28, better than frequency or

similarity.

A few studies have explored persuasiveness of arguments as a quality measure. Wei et al. (2016) ranked arguments by predicting the persuasiveness of debate posts written in response to a preceding comment in the conversation. They acquired discussion threads from the *ChangeMyView* subreddit. Users of this subforum state their views on specific topics, reply with counter-arguments to challenge a view and can vote on different replies to indicate which one is more persuasive than others. A pair-wise learning-to-rank model was used for ranking. In addition to argumentation features, they captured social interaction features. Since it becomes harder to understand an argument without understanding the (preceding) context, the effectiveness of social interaction based features increases when the number of comments in the thread grows. However, the interaction features are similarity-based and do not model persuasion behaviors.

Tan et al. (2016) examined the effectiveness of arguments where people challenge others to change their views expressed in Reddit's subforum *ChangeMyView* discussions. A logistic regression model is used for prediction. Successful arguments are less similar to the original post in content words but more similar in stop words. Use of first-person pronouns and formatted text were other characteristics of persuasive arguments. Authors relate singular vs. plural first person pronouns to the personality trait of Openness to Experience.

Swanson et al. (2015) extended the Internet Argument Corpus (IAC) with dialogues from *createdebate.com* to create a larger corpus consisting of 109,074 posts on the topics Gay Marriage, Gun Control, Death Penalty, and Evolution. They address argument extraction at the sentence level and train a regressor to predict an argument quality (**AQ**) score for each sentence.

The AQ score is intended to reflect how easily the speaker's argument can be understood from the sentence without any context. The annotators rated the argument quality using a continuous slider ranging from hard (0.0) to easy to interpret (1.0). They compared the sentence extraction to multi-document summarization where sentences that clearly represent an argument facet are identified. Their work is based on the hypothesis that the best candidates for high-quality arguments are marked by high semantic density cues such as POS tags, specificity as measured by Speciteller (Li and Nenkova, 2015), lexical n-grams, sentence length, word length, and certain discourse relations.

Our Work. In this era of instant communication, social media provides a major venue for argument and dialogue exchange. Arguments are exchanged back and forth in extended conversations providing a rich resource for capturing and analyzing the vast variety of viewpoints across these conversations. We employed two different methods to separately evaluate argument quality in social media dialogue. The first used common elements across summaries as indicators of argument quality (Misra et al., 2015, 2017). For the second method, we improved upon the argument quality definition from Swanson et al. (2015) for ranking arguments based on contextual knowledge needed to understand an argument (Misra et al., 2016).

2.2.2 Argument Relations and Tagging

Arguments (and their components) are usually embedded in a particular situation and cannot be understood in isolation (Peldszus and Stede, 2013). An argument seeks to establish or challenge a controversial proposition and refers to an explicitly mentioned or at least supposed

opponent. Van Eemeren and Grootendorst (2004) claim that an argument always has two sides, a proponent, and an opponent and both support and attack relations are necessary to represent realistic knowledge.

In this section, we review existing works that focus on the identification of argumentative relations between arguments and analyzing their interactions.

Cabrio and Villata (2012) trained an existing textual entailment (TE) platform, Edit Distance Textual Entailment Suite (EDITS) (Kouylekov and Negri, 2010), to infer a support or an attack relation between two arguments to determine a set of acceptable arguments in a debating portal. Dung (1995) framework is used to evaluate their acceptability. The dataset consists of 200 Text-Hypothesis pairs (entailment and contradiction). An argument is accepted if it is not attacked in the debate, i.e., all the arguments attacking it are rejected, and an argument is rejected if it has at least an argument attacking it which is accepted. The system can be used by the participants to get an overview of the debate and the accepted arguments, with an accuracy of 0.75.

Peldszus and Stede (2016) introduced an annotation scheme based on the work of Freeman's theory (Freeman, 1991) and collected 112 very short texts in a controlled text generation experiment. Each text contains roughly five argumentative relevant segments written in response to a question on some potentially controversial topic (e.g., Should intelligence services be regulated more tightly by parliament?). Writers were asked to include a direct statement of their central claim and at least one objection to that claim. The texts were annotated using Freeman's theory of using the moves of proponent and challenger in a dialectical situation. In addition to distinguishing between claims and premises, support and attack, it is also labeled

with additional properties of argument components like proponent-opponent, normal-example, and rebut-undercut relations. The texts were first written in German and then professionally translated into English. An inter-annotator agreement of Fleiss k=0.83 among three expert annotators was obtained. Using this dataset, Peldszus and Stede (2015) jointly linked several argument components in a single tree structure using a Minimum Spanning Tree model. It has a central claim as the root node connected by relations including support or attack, proponent or opponent. The edge weights are determined based on the features extracted from each component including lemma, verb morphology, POS-tags, discourse connectives and some statistics such as the relative position of the components and length. The system achieved an F-1 score of 0.720 for identifying argumentative relations and 0.869 for recognizing claims in the above micro-text corpus. However, a significant limitation is a presupposition that the components of an argument are already known in advance. Moreover, the texts in the corpus were created in a controlled setting to promote one opposing argument component in each text. Therefore, it remains mostly unknown if the results obtained can be reproduced when working with real data sets.

Argument Tagging: So far, we have studied models and relations between a pair of arguments based on structure and role, but what are the issues that are being discussed? For a given topic, a number of important arguments emerge in the debate, representative of the various aspects of the debate. A majority of users will support their stance or refute an opponent by arguing about these sub-issues, aspects, or facets. As we have a plethora of information in social media and readers are often faced with information overload, merely finding arguments in

isolation is not particularly useful. An aggregation at a higher level such as some correlation between the arguments by different people about the same topic may be more useful to get a quick overview of the various argument aspects (facets). Next, we discuss argument tagging where arguing expressions are extracted and labeled with tags reflecting the conceptual argument being made.

Conrad et al. (2012) constructed an argument mining system on monologic weblog and news data about universal healthcare. One component of their system identified **arguing segments** and the second component labeled the segments with the relevant stance-specific **argument tags**. They showed that distributional similarity features help identify arguments that belong to the same tag set and reported an accuracy of 0.522 for the tagging task. An argument tag is similar to our definition of **facet** since it represents a controversial abstract belief, expressed through arguing subjectivity. However, these subjectivity tags were manually identified by expert annotators after rigorous training.

Boltuzic and Šnajder (2014) proposed a multi-classification approach to tag arguments with a set of predefined labels. Instead of hand-generating argument tags like Conrad & Wiebe, they selected short sentential summaries of the key arguments for a given topic from a debate website, and then labeled comments on the same topic from a different website with the most closely matching summary. They compiled a corpus of user comments from online discussions (Procon.org [1]) on two specific topics and manually paired them with arguments from Idebate.org [2], with a Cohen's kappa of 0.49. Afterwards, they used a supervised model to match user-created comments to a set of predefined topic-based arguments. The final task was

[1] http://procon.org
[2] http://idebate.org

to predict the extent to which the author of the post supports or opposes the reason as measured on a five-point ordinal scale (strong attack, attack, strong support, support, none). Using textual entailment (TE) features, semantic text similarity (STS) features, and a stance alignment (SA) feature, the best models achieved a micro-averaged F1-score in the range 70.5% to 81.8%, depending on the task formulation.

Naderi and Hirst (2016) explored the tagging framework in parliamentary discourse for the Gay Marriage topic to determine the aspects of an issue. The task is formulated as 'frame identification'. Political debates were tagged with pre-existing argument-list by three annotators with an inter-annotator agreement of 0.46. An SVM classifier was trained using distributed word representations, similarity, and stance as features. An overall accuracy of 68.9% was reported.

Ghosh et al. (2014) annotated a corpus of blog comments for argument mining about technical topics and applied a theory of argument structure that is based on identifying **targets** and **callouts**, where the callout is basically an argumentative attack against a particular target proposition in another speaker's utterance. Inter-rater reliability estimates yielded a Krippendorff's alpha in the range 0.64 - 0.87.

The same problem on debate posts is tackled as a 'reason classification' problem (Hasan and Ng, 2014), with a probabilistic framework for argument recognition (reason classification) that operates jointly with the related task of stance classification. Each post is split up into sentences and each sentence is manually labeled with one argument from a predefined set of arguments. An inter-annotator agreement between 0.61 and 0.67 is reported.

A major limitation of these prior tagging approaches is that annotators have to look through and filter a lot of irrelevant content and construct a manually curated list. A change

in political or social environment or external events could change thinking. New aspects could emerge as different sides respond to each other and raise new points.

Our Work. We envisioned a bottom-up approach. Arguments in various discussions can spontaneously arise as the conversation moves forward. These arguments are only later consolidated and grouped based on similarity. We do not work with a preexisting inventory of tags.

Boltuzic and Šnajder (2015) worked in a similar vein and performed cluster analysis using semantic textual similarity to detect similar arguments. Analysis of clustering quality and errors on manually matched cluster-classes revealed that there are difficult cases that textual similarity cannot capture and it is difficult to draw clear-cut boundaries between arguments. In contrast, we adopted a 2-step principled approach in (Misra et al., 2015, 2016):

- Step1: A classification model for Argument Extraction.
- Step2: A regression model for Argument Similarity.

2.2.3 Correlation between Argument, Disagreement, and Stance

2.2.3.1 Argument and Disagreement

In ordinary discourse, the word "argue" often means "to disagree" insistently or aggressively (Groarke, 2017). Arguments in a dialogue often involve responding and reacting to current contexts. As the conversation progresses, the participants of the discussion are not only interested in conveying their beliefs, but they are also speculating whether the other person

believes the same. When a debate arises over any issue, the participants ask questions to determine the level of agreement and then may give counter-arguments for defending or attacking the areas of disagreement. To detect agreement or disagreement can often help to understand how conflicts originate and get resolved, and the role of each participant in the conversation. Identifying these interactions can be particularly helpful in inferring controversial decisions, which can be useful for summarization. (Dis)agreement has also been linked to Speech Act Theory to indicate a participant's intention. Sentences have both a propositional content and a functional intent (Searle, 1975; Jacobs, 1989). The problem has been previously studied in transcribed speech, e.g., in meeting discussions (Hillard et al., 2003; Galley et al., 2004; Hahn et al., 2006), congressional floor-debates (Thomas et al., 2006), and broadcast conversations (Wang et al., 2011; Germesin and Wilson, 2009). Given the increasing interest in ideological dual-sided debates, a lot of research studies in recent years have been carried out on classifying disagreements. A similar task is proposed in Abbott et al. (2011) to recognize disagreement in online political forums between quoted text and a given response. By using discourse markers, generalized dependency features, punctuation, and structural features, the best system achieved an accuracy of 68%. Yin et al. (2012) focused on global (dis)agreement between a post and the root post of the discussion. Allen et al. (2014) used Rhetorical Structure Theory (RST) to build a relation graph suggesting that it captured the essential aspects of the conversational argumentative structure. Rosenthal and McKeown (2015) performed a 3-way classification (agreement/disagreement/none) between quote-response posts and showed that using conversational aspects (e.g., sentence similarity) significantly improved the results as compared to using lexical features alone.

As we are interested in finding out why people disagree and what are the reasons behind disagreement, we performed an initial pilot study to determine the effectiveness of topic independent features, e.g., discourse cues indicating agreement or negative opinions (Misra and Walker, 2013). In a similar vein, Skeppstedt et al. (2016) annotated a debate forum corpus for topic-independent expressions conveying (dis)agreement. Among the 175 annotated expressions (43 for agreement and 132 for disagreement), 163 were unique, which showed that there is a considerable variation in expressions used.

2.2.3.2 Argument and Stance

A task complementary to argument recognition is that of stance classification as users often back up their stance with arguments. It involves identifying a holistic subjective disposition towards a particular topic (Walker et al., 2012b). Stance is similar to a point of view or a perspective and identifies the "side" that a speaker is on, e.g., *for or against Gun Control*. There has been considerable previous work on stance classification in online forums and in congressional debates. Context and meta information such as author constraints have shown to be useful for stance classification (Walker et al., 2012a; Rajendran et al., 2016; Hasan and Ng, 2013a, 2014).

Somasundaran and Wiebe (2009) presented an unsupervised approach using integer linear programming paradigm while Somasundaran and Wiebe (2010) implemented a supervised classification using support vector machine. They explored the utility of sentiment and argumentation trigger lexicons as well as opinion-target pairs to classify stance in online debates on political and social topics. Discourse relations such as concessions and argumentation

triggers improve performance over sentiment features alone. For ideological debates, the best performance was approximately 64% accuracy.

(Agrawal et al., 2003; Murakami and Raymond, 2010) investigated social network structure of the debates. Agrawal's work assumed that adjacent posts always disagree, and did not use any of the information in the text. Murakami and Raymond (2010) showed that rules for identifying agreement defined on the textual content of the post can improve previous results of Agarwal's obtained with reply structures alone.

Anand et al. (2011) used relational information and dialogic structure between users and posts and worked with debates that have rebuttal links between posts. Augmenting an n-gram feature set with meta post, contextual and dependency features, they achieved accuracies ranging from 54% to 69%. Walker et al. (2012a) showed that capturing the dialogic structure between posts using information related to agreement relations between speakers improved results obtained using contextual features. Several others have modeled dialogic structure in more sophisticated ways (capturing users intent, SentiWordNet (Baccianella et al., 2010), frame-semantics, quotations and position information), reporting further improvements from such strategies (Ranade et al., 2013b; Hasan and Ng, 2013a; Sridhar et al., 2015).

More recently, Chen and Ku (2016) employed neural networks that incorporate user, topic, content, and comment information. They achieved an average accuracy of 0.842 significantly beating the previous results from (Hasan and Ng, 2013b; Sridhar et al., 2015) on ideological debates.

In our data, the opinion sharing dialogues were also characterized in terms of the stance participants hold towards the central propositions of these discussions. The dialogue

summaries that we collected also reflected the positions taken by the stance holder and the particular aspects on which speakers agree or disagree. To effectively characterize the dialogue, both the central propositions and the stance towards these propositions should be identified. However, for the purpose of this research, we do not address stance classification or dis(agreement) detection. We assumed that our dialogues are labeled with the stance annotations for each of the writers of the debate.

2.3 Summarization

Summarization is a mature research area with a range of tools and algorithms available. It has gained significantly increased attention in the last few years. In general, much of the recent progress can be attributed to the availability of large public datasets, increased computing power, and new machine learning models, such as neural network architectures (Serban et al., 2015). Based on their input type (media, genre) and expected output, many summarizations tasks have been defined.

2.3.1 Extractive Summarization

Traditional extractive summarizers produced a summary in two steps: sentence ranking and sentence selection. Human engineered features such as sentence position, length, word frequency and importance, among others, have been used to rank a sentence. The former consists of selecting most relevant units (sentences, paragraphs) from the original document while maintaining a low redundancy and then concatenating them into a new shorter document. The

later requires a deeper understanding of the concepts and is more similar to a human-written abstract. It has been observed that longer sentences score higher on metrics that rate them for importance in extractive systems (McKeown et al., 2005). Statistical metrics (e.g., word frequencies and key phrases), cluster centroids, sentence position and discourse structure have been used in the past for extractive systems.

2.3.2 Abstractive Summarization

Abstractive summarization uses natural language generation to compress the main contents of the document by using a different vocabulary from the original source document. Abstractive summarization is generally considered more difficult as it involves sophisticated techniques for meaning representation, content planning, surface realization, etc. The earlier abstractive systems determined the "importance" of a sentence constituent based on shallow features, such as syntactic role, vocabulary and it's relation to surrounding sentences.

2.3.3 Multi-Document Summarization (MDS)

Early summarization systems produced a summary of one document (news article, lectures, scientific abstracts). As the number of documents on the web increased, multi-document summarization attracted attention. It would be useful to get a brief overview of a set of documents containing similar content on the same topic or the same event. The goal was to reduce redundancy by extracting sentences that are both "central" and "diverse".

2.3.4 Speech Summarization

It includes tasks such as summarizing spontaneously spoken dialogues from meetings, voicemail messages, and broadcast news. Spoken language is often less formal than written communication. Fragmented utterances, disfluencies and errors resulting from speech recognition make speech summarization a much harder task than text summarization. Thus systems used for text summarization cannot be directly applied to speech summarization (McKeown et al., 2005). These systems exploit additional information derived from speech signals and dialogue structure (acoustic, prosodic, turn taking and dialogue acts).

2.3.5 Traditional Document Summarization

Document summarization is an automatic procedure aimed at producing fluent and coherent summaries that extract salient information and minimize redundancy. The Document Understanding Conference (DUC) provided the benchmark datasets for comparing and evaluating summarization systems (Harman and Over, 2002). This stimulated the development of computational methods for generating compressed document summaries. Both unsupervised and supervised approaches were explored. Many previous summarization systems employed an extractive approach by identifying and concatenating the most salient text units (often whole sentences) in the document. Most of the extractive systems computed salient scores of sentences and ranked them to form a summary. The important parts were often retrieved by using a range of algorithms, such as graph centrality, constraint optimization via integer linear programming, submodular functions, or support vector regression.

One of the initial methods was Maximal Marginal Relevance used by Carbonell and

Goldstein (1998). A greedy search is used to consider the trade-offs between relevance and redundancy. McDonald (2007) replaced the greedy approach with a dynamic one and used a modified objective function in order to consider whether the selected sentence is globally optimal.

Marcu (1999) developed a discourse-based approach. A relationship between the segments of the text was defined by using Rhetorical Structure Theory (RST) (Mann and Thompson, 1988). More important segments (nucleus) occupy the upper level of the tree, whereas less important segments reside deeper (satellite) in the tree. Christensen et al. (2013) jointly optimized coherence and salience by using an approximate discourse graph that represents discourse relationship between two sentences.

Cluster centroids were exploited by Radev et al. (2004) for summarization of multiple news articles. TF-IDF vector representations of the documents were used as input for clustering. Sentences central to the topic of the cluster were identified using cluster centroids. Both relevance and redundancy were taken into consideration. Graph-based representation that can capture interrelated information into the ranking component has been successfully used to determine the centrality of a sentence in LexRank (Erkan and Radev, 2004) and TextRank (Mihalcea and Tarau, 2004), with sentences as nodes and similarity as edges. DivRank (Mei et al., 2010) based on reinforced random walk in an information network balanced the prestige and diversity of the top-ranked vertices and achieved improved results on MDS. Since these methods considered the intrinsic structure of the texts instead of treating texts as simple aggregations of terms, they were able to capture and express richer information in determining important concepts (Ouyang et al., 2009).

Gillick and Favre (2009) developed an ILP model under the assumption that the value of a summary is the sum of values of the unique concepts, subject to a length constraint. Concepts are approximated by the bigrams, weighted by the number of input documents in which they appear. In subsequent work, Berg-Kirkpatrick et al. (2011) extended this model by including sentence compression using parse trees. Woodsend and Lapata (2012) jointly optimized different aspects including content selection and surface realization.

Summarization has also been formulated as a submodular maximization problem and shown promising results. Several different functions that combine coverage of relevant units and minimize redundancy have been proposed. Sipos et al. (2012) formulated the learning problem as a structured prediction problem and derived a maximum margin algorithm using the structural support vector machine framework. Lin and Bilmes (2011) used pairwise similarity to combine importance and diversity while Morita et al. (2013) formalized it using subtree extraction.

In the last few years, neural network-based methods have been developing extremely rapidly. A vast majority of the work has concentrated on extractive summarization as it is less complex and usually generates grammatically and semantically correct summaries. The earlier works explored deep networks to project sentences onto distributed representations but employed traditional frameworks for extraction (Yin and Pei, 2015; Kobayashi et al., 2015). More recent works have used deep networks in a more direct data-driven way to predict the extraction of sentences. Cheng and Lapata (2016) modeled it using an encoder-decoder framework while Nallapati et al. (2016) used a recurrent neural network (RNN)-based sequence classifier, and applied it to CNN/Daily Mail corpus (Hermann et al., 2015). Yasunaga et al. (2017) applied a

Graph Convolutional Network (GCN) and showed that unlike traditional RNN models, it can capture sentence relations across documents, thus better salience prediction and summarization for a multi-document summarization task. Progress in text-to-text generation boosted interest in abstractive summarization to generate novel text. Filippova et al. (2015) used word embeddings and Long Short Term Memory models (LSTMs) for sentence compression. Encoder-aligner-decoder models from machine translation, convolution-based encoders, and hierarchical attention models were used for compressing a document (Rush et al., 2015; Nallapati et al., 2016). These methods were able to generate summaries with high ROUGE (Lin, 2004) scores. However, these systems have typically focused on the task of headline generation and summarizing short input sequences (one or two sentences) to generate further compressed summaries. A critical issue with these models is that for longer documents and summaries they often generate unnatural summaries including repetitive and incoherent text (Paulus et al., 2017).

2.3.6 Conversational Summarization

Spontaneous conversations are more closely related to everyday experience and of more personal interest than news or scientific documents, e.g., telephone/email conversations, social media debates, and comment threads. Businesses can also benefit from summarizing meeting and call-center service dialogue and leverage the data to make better decisions. Earlier works on conversation summarization mainly focused on summarizing transcribed spoken conversations and text-based dialogues. Much work on transcribed conversation has focused on speaker identification, speech disfluencies, and filler particles (Zechner, 2001; Zhu and Penn, 2006). A wide range of methods employed, with many of them relying on notions

of salience and semantic similarity, such as Maximum-Marginal Relevance (MMR) (Zechner, 2002; Gurevych and Strube, 2004), Latent Semantic Analysis (LSA) (Murray et al., 2005), and topic modeling (Hazen, 2011). Wang and Cardie (2011) explored dialogue act and clustering for summarizing decisions in spoken meetings. (Rambow et al., 2004; Oya and Carenini, 2014) used dialogue act modeling in email thread summarization. Several studies also evaluated the effect of speech-specific acoustic/prosodic features. Murray et al. (2006) explored the usefulness of incorporating speaker and discourse information using features related to speaker activity, listener feedback, discourse cues and dialogue act length. Maskey and Hirschberg (2005) found that the best results were obtained by combining prosodic, lexical, and structural features in the broadcast news domain.

While there has been some previous work on debate summarization, these methods do not consider the dialogic nature of argumentation. Ranade et al. (2013a) summarized online debates using topic and sentiment rich features, but their unit of a summary is a single debate post, rather than an extended conversation. Wang and Ling (2016) generated abstractive one-sentence summaries for opinionated arguments from debate websites using an attention-based neural network model, but the inputs were well-structured arguments and a central claim constructed by the editors, rather than user-generated conversations.

2.4 Chapter Summary

Social media arguments are often informal, and do not necessarily follow logical rules or schemas of argumentation. Therefore, in social media, segments of text that are argumenta-

tive must first be identified. Much of the previous work does this task manually (Goudas et al., 2014; Hasan and Ng, 2014; Boltuzic and Šnajder, 2014; Habernal et al., 2014). However, a manual approach does not scale well as argument density can be low. Arbitrarily selected utterances are unlikely to be high quality arguments as users may express their arguments in a confusing or ungrammatical manner. The language is not always explicit as conversants can always clarify misunderstandings. As there are no specific rules to be followed, there are often off-topic informal discussions. For instance, Goudas et al. (2014) annotated 16,000 sentences from social media documents and found only approx. 4.8% to be argumentative.

Another close parallel with our research is the work on argument tagging that aims to identify argument aspects across discussions (Conrad et al., 2012; Boltuzic and Šnajder, 2014; Hasan and Ng, 2014; Naderi and Hirst, 2016). However, all these works assume that a predefined aspect list already exist and perform a classification task of mapping aspects to arguments. Our approach differs in two ways. First, we do believe that a **Facet** should be independent of stance: the same facet can appear on both sides of an argument. Second, we do not believe it is possible to enumerate all the possible facets for a discussion topic in advance. Rather, we are attempting to construct the facets bottom-up from a corpus of discussions. Similar arguments can be paraphrased and expressed in infinitely different ways. Grouping similar arguments on a given topic in online debates can potentially help users formulate opinions by presenting a richer picture of the controversial issues involved. Detecting similar arguments that convey the same information is our approach to finding argument facets.

Another key aspect that distinguishes our problem formulation from previous work is that we for the first time work from human summaries of dialogue, while all the previous

work in argument mining work from the source text itself. It was an open question whether the **central propositions** for a dialogue as obtained from pyramid of human summaries are identifiable as continuous spans of text in the dialogue itself. (Indeed, our Argument Dialogue corpus (described in Chapter 3) will allow us to determine how true that assumption is, and if we can reliably back-link these pyramid labels to original dialogue for argument extraction).

Work on summarization has mostly focused on monologic data and summarizing written texts, where the notion of an abstract of the text was well defined. Most previous work on summarizing spontaneously-written dialogue aimed only to extract phenomena specific to meetings, such as action items or decisions (Murray et al., 2006; Galley et al., 2004; Wang and Cardie, 2011). Other approaches, like our work, use semantic similarity metrics to identify the most central or important utterances of a spoken dialogue (Gurevych and Strube, 2004), but do not attempt to find the facets of a set of arguments across multiple dialogues. Work in open domain argumentative dialogue summarization was limited - partly because such a model should be able to capture both argumentation properties and dialogical context, and partly due to unavailability of large argumentative dialogue corpora. This had impacted the dialogue research community's ability to develop better theories, as well as good off the shelf tools for dialogue processing. Happily, an increasing amount of argument exchange occur in natural dialogue in online forums, where people share their opinions about a vast range of topics. The phenomenal growth of social networking media led to a massive amount of conversation and debate online. This data deluge yields an unprecedented opportunity to create vast datasets of naturally occurring argumentation than would otherwise be possible. These, in turn, opened new research lines for corpora based approaches for argumentative dialogue modeling. Automatic detection

and summarization of argument aspects from a dialogue was the next important step towards a better understanding of argumentative dialogue.

In the next three chapters of this thesis, we describe a novel annotated argument dialogue corpus, new methods for identifying important arguments, argument facets, as well as summarizing arguments in a written dialogue in a social media domain.

Chapter 3

Argument Dialogue Corpus

This chapter addresses the first two research questions to be answered in this thesis, **Q1:** Identifying central propositions in a dialogue, and **Q2:** The most important arguments in a dialogue.

The proliferation of online social media forums gives a wealth of new opportunities for processing dialogic conversations. As the dialogue evolves, salient arguments may emerge over time depending on the context. While some of these dialogues may be a little more than flame wars, a significant portion involves contentful, reasoned disputes on important social and political topics, as exemplified by the forum snippets in the Figure 3.1. Studying data like this will undoubtedly help us to understand dialogic and informal argumentative language in general.

Disagreement in these dialogues is frequent, ranges from 80% to 90% across topic (Misra and Walker, 2013). This data provides a useful resource for examining and modeling the arguments that flow in these disagreements. Previous work on this kind of data (Somasundaran

S1:1 Agreed She is ignoring my religious freedom and trying **to institute her religion into law. The law that will bar my family from legal protections. It won't protect her marriage but will bar me and my people from from being full citizens**. She isn't protecting marriage but bf perserving her heterosexual privledge.
S2:1 How on earth is she **impeding on you religious freedom**? She isn't trying to take away your right to any religious ceremony. With such a wide-open standard of what constitutes religious freedom that you seem to have, any legislation could be construed as imposing on religious freedom.
S1:2 Because it is her religious belief that **marriage is between a man and a woman**. My religious belief is that **marriage is between two people that love each other regardless of sex**. She is tying to place her religious belif into law over mine. Who gets hurt here? If my religious belief is put into law she can still marry the person of her choice. If her religious belief gets put into law she can still marry the person of her choice but I do not get to. So I and my people are hurt by codifing her religious belief into law. She is trying to keep gay people out of marriage and thus preserve her heterosexual privledge.
S2:2 But by that definition, either one could be viewed as impeding on religious freedom, including your view impeding on hers ! We don't define imposing on religious freedom on the basis of having different ideals. **It doesn't effect your religion or religious freedom if you don't get benefits under gay marriages**. You can argue in other ways, on other basis, but the idea that not giving gays marriage benefits is imposing on religious freedom is an empty " argument ".
S1:3 But you have not shown how she would be harmed by equal marriage? I have shown that her definition infringes upon my religious freedom. If gay marriage becomes a reality then in no way does it infringe upon her in any way. Would you care to explain how gay marriage infringes upon her in any way? It is her heterosexual privledge and her sence of entitlement that she is protecting not marriage.
S2:3 You may (or may not) can show harm in another sense, but you certainly have NOT shown **how your religious rights are taken away or harmed**, to begin with. That's the point. You certainly have not shown how her definition infringes upon your religious freedom – you are still free to be just as religious in practice and belief as ever if Musgrave's views became law. You're trying to use a religious freedom argument here and it just doesn't fit.
S1:4 Yes I have. Haven't you been poaying attention? Because it is her **religious belief that marriage is between a man and a woman. My religious belief is that marriage is between two people that love each other regardless of sex.** She is tying to **place her religious belif into law** over mine. Who gets hurt here? If my religious belief is put into law she can still marry the person of her choice. If her religious belief gets put into law she can still marry the person of her choice but I do not get to. So I and my people are hurt by codifing her religious belief into law. She is trying to keep gay people out of marriage and thus **preserve her heterosexual privledge.**
S2:4 Ok! My religious belief says that it's important to smoke in public non-smoking areas. My religious belief says that it's important to drive cars as fast as one can on city streets. My religious belief says that the death penalty should never be enforced. Gee, this it fun, using religion as an excuse and means to an end to fight legislation! Too bad for you it doesn't work that way. What you don't get is that, by your standard, EITHER would be placing one religion over another. The same justification could be made against your view because your view goes against MY religious belief. And then if you say " but your view causes harm and mine doesn't " then it's no longer a religious issue so the point is moot to begin with.
S1:5 YOu missed the point here she is trying to turn her religious belief into law that bars me from from getting married. How is she hurt if it is my religious belief of equal marriage is turned into law?
S2:5 You have no idea whether or not she is truly trying to turn her religious belief into law. She may just happen to believe that way religiously, but be pushing for the law because she thinks that's best for socety. IOW, you have no idea if religion is her real motivation for the law. There's no way to know or to police someone's real motivations.
S1:6 I am part of society and her belief is not in my best intrest. How does my being able to marry a man hurt her in any way?

Figure 3.1: Gay Marriage Dialogue-1.

and Wiebe, 2010; Abbott et al., 2011) has examined the structure of these discussions - e.g., the argumentative discourse relation a post bears to its parent (agreeing or disagreeing), or the stance that a person takes on an issue. But this data also has the promise of revealing the argumentative topography for a given issue. In particular, we aim to develop tools to summarize what various disagreement arguments are used to support or oppose the issue under debate.

We begin with our efforts to systematically create and curate a dialogue corpus with

a range of different properties in Section 3.1. Regarding the central proposition detector, we first focus on the question of extracting reliable data for central propositions, the most important propositions in a dialogue. This is notably tricky, given the subjectivity of judgments of both importance and the natural size of a central proposition. To reduce the annotator bias, we operationalized "salience" as "central proposition", and then used human summarization to discover saliency. Dialogue collection is followed by a summarization task on Mechanical Turk in Section 3.2. We demonstrate how summarization by many human summarizers can be exploited to index salience. We propose to use pyramid annotation of summaries (Nenkova et al., 2007) to find salient propositions in a dialogue. The pyramid method is a content evaluation metric for extractive document summarization. We describe the process of pyramid annotation in Section 3.3. Once a set of reference summaries are collected, the pyramid evaluation scheme provides a useful way to calculate global salience (Nenkova and Passonneau, 2004; Nenkova et al., 2007), thereby allowing us to identify *central propositions* in any dialogue.

The knowledge of central propositions helps us move towards the next goal of the thesis to summarize important arguments. Though we have identified the main propositions, these are human generated abstractive labels. In order to train a supervised model, we transform these labels to ground truth sentences in the dialogue, representing their importance in the summary. The motivation behind this approach is to provide an annotation scheme that minimizes the subjective opinions and annotator bias. We do not ask annotators to explicitly mark or rank argumentative sentences in the dialogue. We define an argument objectively rather than purely based on the annotator's intuition or judgment of an important argument. Once these annotations are performed, the argumentative sentences in the dialogue get a rank by virtue of

the annotation process. The transformation process and annotations are discussed in Section 3.5. The entire annotation process consists of the following steps:

S1: Dialogue data sources and selection criteria (Section 3.1).

S2: Mechanical Turk summarization of dialogues selected in S1 (Section 3.2).

S3: Pyramid annotation of summaries produced by S2 (Section 3.3).

S4: Select top-tier pyramid labels as the *central propositions* for individual dialogues (Section 3.4).

S5: Pyramid labels linked to dialogue sentences identify important arguments in the dialogue, which can be then utilized as the ground truth for supervised training (Section 3.5).

The goals of the above annotation process are to determine how reliably the annotators can link central propositions to sentences in dialogue, $\mathbf{H_1}$. This is essential to establish the validity of the argumentative dialogue corpus and set a human upper bound.

3.1 Dialogue Data Sources and Selection Criteria

For this study, we started with the publicly available Internet Argument Corpus (IAC), an annotated collection of 109,553 forum posts (11,216 discussion threads) (Walker et al., 2012c). We used the portion of the IAC containing dialogues from the website `http://4forums.com`.

4Forums Structure: On 4forums, a person starts a discussion by posting a topic or a question in a particular category, such as society, politics, or religion. Forum participants can then post their opinions, choosing whether to respond directly to a previous post or to the top level topic

Figure 3.2: A sample Quote Response pair from 4 Forums showing the reply structure.

(start a new thread). Conversants may just agree or disagree with an earlier post or they may provide a reasoned argument. The corpus contains posts on topics such as *Abortion, Evolution, Existence of God, Gay Marriage,* and *Gun control* along with a range of useful annotations. First, there are annotations that collapse different discussions into a single topic for 14 topics. For example, the *Evolution* and *Gun Control* topics include discussions initiated with the range of titles in Table 3.1, which guarantees variation in the focus of the discussions even within topic. Each discussion is threaded so that we can identify direct responses. Discussions may have a tree-like structure, so a post may have multiple direct responses. In addition to the adjacency pairs yielded by threading, 4forums also provides a quote/response (Q/R) mechanism where a post may include a quote of part or all of a previous post.

4Forum Affordances: 4Forums provide additional meta information which includes participant supplied dialogue annotations. For example, an important type of affordance from 4Forums

Evolution	Evolution in school, Dinosaurs and Human Footprints, Can Evolution & Religion Coexist, Did Charles Darwin Recant, Shrinking Sun, Bombardier beetle, Moon Dust, Second Law of Thermodynamics, Magnetic Field, Nebraska Man
Gun Control	Gun Control, Trigger Locks, Guns in the Home, Right to Carry, Assault Weapons, One gun a month, Gun Buy Back, Gun-Seizure Laws, Plastic Guns, Does gun ownership deter crime, Second Amendment, Enforced Gun Control Laws, Gun Registration, Armor piercing bullets, Background Checks at Gun Shows

Table 3.1: Discussions Mapped to the Evolution and Gun Control Topics.

is the support for quoting another person's post. These represent critical dialogic information, showing explicit efforts by the conversant to focus their response on a particular aspect of another's turn. The site has additional affordances that have been captured such as Quote Text, Post Id associated with each post, Date, and Quote Author.

We used the links as shown in Figure 3.2 in the meta-data to extract a sequence of turns to build two-party dialogue chains like those in Figure 3.1. We have multiple topics in the dataset. Gay Marriage was used as a topic of exploration to provide an empirical basis for the data selection criteria.

- **Number of turns per contributor:** To gain a deeper understanding on the structure of these dialogues, we extracted all dialogue chains and varied the number of turns in a dialogue chain to see the effect in terms of number of dialogues. The graph in Figure 3.3 explains this phenomenon. We wanted dialogues in which substantive issues were discussed, so we extracted dialogues with at least three turns per conversant that present

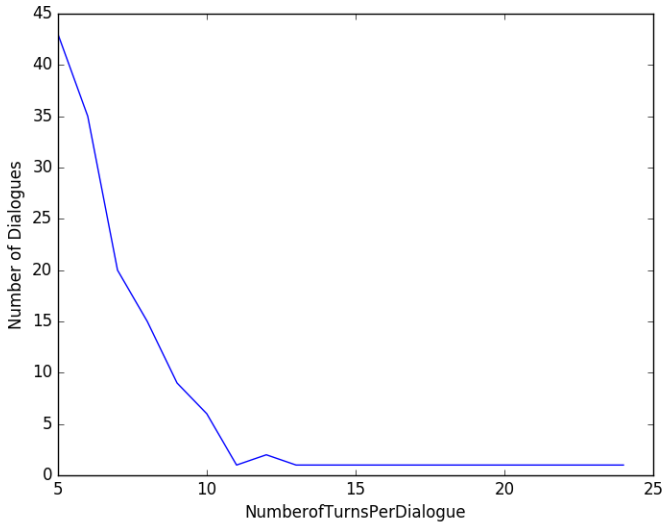


Figure 3.3: Variation in the number of dialogues as compared to the number of turns in the dialogue for the topic Gay Marriage.

at least two different perspectives on an issue.

- **Author:** Some authors post frequently and would dominate the corpus if we use random selection. To examine this trend, we plotted the number of posts by author, see Figure 3.4. To get richer, more diverse dialogues expressing different perspectives, we only select a single dialogue between any particular pair of authors from a discussion thread.

- **Word Count in a post:** Some posts are long. To make it practical to collect dialogue summaries, we extract dialogues where the number of words per turn is less than 250.

We then extracted dialogues from three controversial issues using the above-mentioned criteria. The final dialogue corpus consists of 61 dialogues on the topic of Gay Marriage, 50 on Abortion,

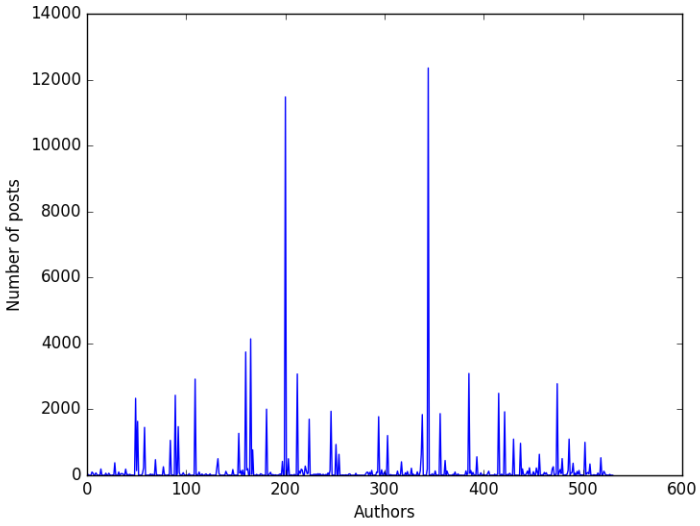


Figure 3.4: Variation in the number of posts for each author for the topic Gay Marriage.

and 50 on Gun Control.

3.2 Summarization

As discussed in Chapter 2, there was a lack of resources and corpora annotated for argumentative dialogue on the controversial ideological issues. There is no prior work on summarizing important arguments from noisy, argumentative dialogues. The few corpora annotated for argument mining do not train end-to-end summarization framework for dialogic arguments. We created a unique corpus of human-written summaries of argumentative dialogue on social media annotated with the pyramid labels.

A priori, it is not even clear what form such a summary should take. The two conversants in Figure 3.5 put forward opposing views: Should a summary give preference to one

S1:1 Certainly not yours. You should know that I am for **no marriage in government**. It should be left to a religious institution where it will actually mean something. The states should then go back to doing something that actually makes sense and doesn't reward people like Britney Spears for being white trash.
S2:1 That is all well and good, but it is not the religious ceremony and sanction that gays are looking for. They already have that ; there are churches that perform same-sex marriages. **It is the civil benefits that are at issue. Are you saying you would be in favor of foregoing ALL the legal rights and benefits you are afforded by marriage?** For example : * Assumption of Spouse s Pension * Automatic Inheritance * Automatic Housing Lease Transfer * Bereavement Leave * Burial Determination * Child Custody * Crime Victim s Recovery Benefits * Divorce Protections * Domestic Violence Protection * Exemption from Property Tax on Partner s Death * Immunity from Testifying Against Spouse * Insurance Breaks * Joint Adoption and Foster Care * Joint Bankruptcy * Joint Parenting (Insurance Coverage, School Records) * Medical Decisions on Behalf of Partner * Certain Property Rights * Reduced Rate Memberships * Sick Leave to Care for Partner * Visitation of Partner s Children * Visitation of Partner in Hospital or Prison * Wrongful Death (Loss of Consort) Benefits What do you say?
S1:2 yeah I know. I'm saying that there should be a better system. For example, if you had a best friend who you are roommates with... both hetero for the sake of argument... and never wish to get married then could they get some of the benefits you described? My Uncle actually was single until he died but he had a female friend who he was best friends with... why couldn't they get some kind of benefits if they so choose? Hospital vistiation rights and things of that nature are obviously the most important in this kinda situation. It just seems the single folks are the ones getting screwed over in the marriage deal :)
S2:2 But gay folks are not? To the single folks, I say get married.
S1:3 Ok, lets use that logic.... To the gay folks I say marry a woman. Bet you aren't happy with that :)
S2:3 If you want the benefits of marriage, get married. But you ought to be able to marry a person of your choice. Your suggestion would be to marry someone I don't want to me married to, that is a different thing. It is not using " that logic ". I thought marriage was supposed to mean something to you. Isn't it supposed to mean a commitment with rights and responsibilities? I am not proposing that anyone enters into that commitment lightly. I happen to think that's a bad idea. People get married for many reasons, some of which I would feel are not good reasons, but I would not make it illegal to do so. I am talking about the right to marry a person of your choosing.

Figure 3.5: Gay Marriage Dialogue-2.

person's views? Should a summary be based on decisions about which argument is higher quality, well structured, more logical, or which better follows theories of argumentation?

A primary objective of this research was to determine what aspects of these dialogues people are oriented to? If asked to summarize these dialogues, what would people decide to mention as important? Is it the dialogic nature or the social aspect or the stance that participants hold towards a topic?

In general, summarization provides an efficient, open-ended mechanism for determining the central aspects of a dialogue: at the individual level, each reader conveys what they subjectively deem the central or most salient aspects of the dialogue. Ranking schemes from the summarization literature can then provide a objectivity score over these summary extracts. We thus pursued a summarize-and-collate strategy: human workers would summarize each

dialogue, and then the common elements across those summaries would serve as our central propositions.

As exemplified in Figures 3.1 and 3.5 these dialogues are clearly different from traditional media. They are participative and collaborative but uncontrolled and unregulated. These dialogues can be viewed from three different perspectives. First, it is a *social* interaction between two or more individuals, who may be highly involved and often using emotional and colorful language to make their points. These debates are also personal, giving a strong sense of the individual making the argument, and whether s/he favors emotive or factual modes of expression, e.g. (*I'm saying that there should be a better system* in S1:2 in Figure 3.5). Then, there is the dialogic aspect, a sequence of utterances, linked to particular speakers with particular *dialogic* goals (such as advice giving, argument, recommendation, and persuasion). Besides, there is also, the *propositional* level, indicating the stance people hold for a general issue.

3.2.1 Mechanical Turk Summarization Task

Summarization is something that any native speaker can do; it does not require training. Thus, as an annotation of salience, it is highly amenable to crowd sourcing. The summarization task was run on Mechanical Turk. To get good quality summaries, workers completed a qualification test involving summarizing a sample dialogue. Workers were instructed to summarize according to dialogue length: dialogues under 750 words in 125 words, and those above 750 in 175 words. We collected 5 summaries for each dialogue resulting in a dataset of 305 summaries for Gay Marriage, 250 each for Abortion and Gun Control topics. Figure 3.6 provides all of the 5 summaries collected for the dialogue in Figure 3.1. Our analysis of the summaries

This is about gay marriage. S1 seems to believe that someone is making religion into law by saying marriage should be between a heterosexual couple. S2 is arguing that S1's point is not very valid because S1 is not understanding that they are not using the persons perspective. S1 believes that that person is trying to hurt them, while S2 is saying that the person may not necessarily be that way and that the person may be arguing for what they believe to be greater for society, based on their own perspective. S1's religious view of marriage is based on two people's love for each other, while the other person believes its between a man and woman. S2 says S1's argument is empty and that S1 is not showing how the other person is being hurt by gay marriage. S2 thinks S1 is not understanding the overall perspectives and just focusing on their own needs. S1 states that the person's religion put into law affects peoples ability to marry if they are gay, but that S1's religion on marriage put into law would not affect marriage for anyone.
S1 and S2 are discussing gay marriage in the context of an unknown "she" that is pushing a law to ban gay marriage. S1 believes that her religious freedom is being persecuted from people against gay marriage. S1 believes this will result in her family not having legal protection. S2 doesn't believe that S1's religious freedom is being persecuted because she is still bale to hold religious ceremonies. S2 believes the same argument can be made fro m her perspective. S1's believe of marriage is that it should be based on love and not on sex. S2 argues that the religious freedom argument could be used in any format to justify any action. S1 is angered that the religious freedom of another person will go into law and bar her from marriage so she can sustain her heterosexual privilege. S2 disagrees and thinks "she" is doing what she thinks is before society. S1 and S2 seem to get annoyed with each other by the end of dialog.
(S1) says that his religious freedom is being ignored by a politician attempting to make her religious beliefs law. This law limits legal protection of gay couples and it is not use to protect marriage but heterosexual privilege. (S2) questions how the politicians view is imposing on (S1) religious freedom. (S1) responds that he believes that marriage should be between two people who love each other and not her/politicians belief that marriage should be between man and woman. However, (S2) says that the argument surrounding religious freedom is an empty argument because anyone can argue for his or her beliefs using religion. (S1) says that gay marriage doesn't violate her/politicians rights of heterosexual privilege, entitlement since she is not protecting marriage but her beliefs. (S2) questions how (S1) religious rights are taken away or harmed. (S1) and (S2) discuss about whose religious belief is better for society. (S1) responds that the her/politicians interest is not part of his interests and or society's and does not hurts him in any way.
S1 says that someone, "her", is trying to put their religious beliefs into law that goes against the religious beliefs of S1. The law would bar the family of S1 from legal protections but would not affect her. S1 believes that she isn't protecting marriage but preserving her heterosexual privilege. Her religious belief is that marriage should be between a man and a woman but S1 believes that marriage should be between two people in a loving relationship regardless of sex. If she put her religious belief into law it would hurt S1 but if the religious views of S1 are put into law she would still be able to choose who she wanted and S1 would be able to choose whomever. She is trying to keep gay people out of marriage. S2 understands the perspective of S1 but does not think it is a valid argument. S2 holds that either if either religious belief were put into law that it would infringe on the other's beliefs. S2 claims that S1 has not shown how the religious beliefs would be taken away from such a law. It can be argued in other ways but the idea of not giving gays marriage benefits is imposing on religious freedom is an empty argument. S2 counters S1 by saying that it is impossible to know whether or not she is trying to turn religious belief into law. It is impossible to know or to police someone's real motives.
S1 is talking about a female who is ignoring S1's religious freedom and barring his family from legal protections. S2 questions to how she is impeding S1 from his religious freedom and argues that she isn't trying to take away his right from a religious ceremony. S1 says the woman's religious belief is that marriage should be between a man and a woman, and S1's religious belief is that marriage can be between any two people that love each other regardless of their gender. If the woman's religious belief gets incorporated in to the law, S1 will not be able to marry the person of his choice but she will still be able to marry regardless of the law. The woman is just trying to keep gay people out of marriage to preserve her heterosexual privilege. S2 argues that either one of them can be impeding on the others religious freedom. S1 does not give in and keeps arguing that she is trying to turn her religious belief into law and bar him from getting married.

Figure 3.6: 5 Summaries for dialogue in Figure 3.1.

supported our expectation that listeners would be oriented to 'content' properties that are not strictly propositional. The summaries do express a conversant's stance and the particular aspects on which speakers agree or disagree. In other words, the summaries in addition to pointing out particular beliefs (central propositions) also show the attribution of these beliefs to someone. At

the social level, summarizers were sensitive to the quality of interpersonal interaction of participants, commenting on levels of coordination *(S2 understands the perspective of S1) vs. (S2 is arguing) vs. ((S1) and (S2) discuss whose religious belief is better for society.)* vs. *(S1 and S2 seem to get annoyed with each other by the end of dialog)* and emotional states of participants during the discussion *(S1 is angered)*.

(S1) says that his religious freedom is being ignored by a politician attempting to make her religious beliefs law. This law limits legal protection of gay couples and it is not use to protect marriage but heterosexual privilege. (S2) questions how the politicians view is imposing on (S1) religious freedom. (S1) responds that he believes that marriage should be between two people who love each other and not her/politicians belief that marriage should be between man and woman. However, (S2) says that the argument surrounding religious freedom is an empty argument because anyone can argue for his or her beliefs using religion. (S1) says that gay marriage doesn't violate her/politicians rights of heterosexual privilege, entitlement since she is not protecting marriage but her beliefs. (S2) questions how (S1) religious rights are taken away or harmed. (S1) and (S2) discuss about whose religious belief is better for society. (S1) responds that the her/politicians interest is not part of his interests and or society's and does not hurts him in any way.

Figure 3.7: Sample 'play by play' summary for the dialogue in Figure 3.1

Summarizers also paid attention to the dialogic structure of the discussion in different ways. First, a majority of the summaries included explicit reference to who said what; play by play devices characterizing the utterances in temporal succession. Figure 3.7 shows a play-by-play summary for the dialogue in Figure 3.1. Moreover, summarizers frequently went beyond the text itself to characterize participant emotions, and subjective characterizations of dialogue acts such as S1 is angered, S2 questions, S1 responds.

At the propositional level, many summaries contained explicit statements of the **cen-**

tral propositions on the table (e.g., S1 and S2 are discussing gay marriage in the context of an unknown "she" that is pushing a law to ban gay marriage). They also frequently convey the stance that participants bear towards the propositions that are introduced by their fellow participants (e.g., Her religious belief is that marriage should be between a man and a woman).

3.3 Pyramid Annotation

The pyramid method was developed for content coverage quality assessment for summarization evaluation by analyzing the distribution of content over a pool of human-generated reference summaries (Nenkova and Passonneau, 2004). The annotation of pyramids seeks to uncover the common elements, across several reference summaries. To create a pyramid, annotators identify semantically defined clusters of similar phrases in reference summaries, referred to as Summary Content Units (SCU). An SCU consists of a label, its contributors, and tier rank. The label assigned by the annotator indicates the semantic meaning of the content unit. The contributors retain the actual text snippets to depict the original words used in a particular summary that express the label. The SCU tier rank is based on its frequency in the corpus of reference summaries. It is equal to the number of reference summaries that express that SCU and represents the relative importance of the SCU. The number of contributors per SCU thus ranges from a minimum of one to a maximum equal to the number of reference summaries. The rational behind manual identification of semantic equivalence is to combat human variability in content selection and that multiple maximally informative summaries are possible. An ideal maximally informative summary would express a subset of the most highly weighted SCUs

(Louis and Nenkova, 2013).

Contributor	S1 points to the trend to legalize gay marriage in western countries such as Netherlands, Belgium, and most of Canada.
Contributor	S1 refutes this assertion, citing a number of countries which recognize same-sex marriage.
Contributor	He states the US is more similar to Anglo nations and in many of those gay marriage is legal.
Label	A number of countries recognize same-sex marriage.

Table 3.2: A sample label after removing the attributions from the SCU contributors.

To more systematically index the salient elements of an opinion sharing dialogue, we conducted a pyramid evaluation. We trained linguistic undergraduates to annotate summaries to produce pyramids. The label was subject to revision throughout the annotation process and reflected the semantic meaning shared by all the contributors. Because our aim here was to focus on argument propositional content, the annotators were instructed to keep only the main proposition in the SCU as the label, ignoring any attributions or other types of content. Table 3.2 shows the summary contributors and the corresponding label given by the annotator.

Once the annotation process was complete, final SCUs were partitioned into a pyramid based on their frequency across all of the summaries as indicated by tier rank. Each tier contained all the SCUs with same rank. Since we used annotations from five summaries, our pyramid contained five tiers. SCUs with higher tiers are placed at the top. Table 3.3 shows some of the summary contributors, pyramid labels, and the corresponding tier ranks for Gun Control.

Summary Contributors	**Human Label from Pyramid Annotations**	**Tier Rank**
• S1 says that no one can prove that gun owners are safer than non gun owners. • S1 says no one has been able to prove gun owners are safer than non-gun owners. • S1 points out there is no empirical data suggesting that gun owners are safer than non-gun owners. • S1 states there are no statistics proving owning a gun makes people safer. • S1 believes that there is no proof that gun owners are safer than non-gun owners.	Nobody has been able to prove that gun owners are safer than non-gun owners.	5
• They say that if S2 had a family member die from gun violence it might be more significant to them, • He says if S1 had a personal or family encounter with gun violence, he would feel differently. • That people who have had relatives die from gun violence have a different attitude.	Family encounters with gun violence changes significance.	3
• Pro-gun perspective is: on 9/11, 3000 people died without the ability to defend themselves.	On 9/11, 3000 people died without the ability to defend themselves.	1

Table 3.3: Example summary contributors, pyramid labels and tier rank in Gun Control dialogues

3.4 Central Propositions from Pyramid

The pyramid in Table 3.4 includes data from all the summaries in Figure 3.6. It is evident from the pyramid that higher tier labels express the key arguments that can be used to support a stance for a given topic as compared to lower tiers. SCU labels in Tiers 3-5 give the **central propositions** of the dialogue.

SCU Label	Tier
Law would bar the families from legal protections.	5
Religious view of marriage is based on two people's love for each other.	5
Someone is making religion into law.	5
Religious freedom is not being persecuted.	5
Law affects peoples ability to marry if they are gay.	4
Preservation of heterosexual privilege.	4
Marriage should be between a heterosexual couple.	4
Gay marriage laws would not affect marriage for anyone.	4
Religious freedom is being persecuted.	3
What is believed to be greater for society	3
Points not valid	3
Impeding on the others religious freedom.	2
Not protecting marriage.	2
Religious belief gets incorporated in to the law, some will not be able to marry.	2
Same argument can be made from different perspective.	2
Not trying to take away his right from a religious ceremony.	2
Religious freedom argument could be used in any format to justify any action.	2
This is about gay marriage.	2
Two people annoyed with each other by the end of dialog.	1
Politicians interest is not part of his interests.	1
people against gay marriage.	1
It is impossible to know whether or not she is trying to turn religious belief into law.	1
Not understanding that the persons perspective is being used.	1
The her/politicians interest is not part of societys interests.	1
Focusing on one's own needs.	1
Not giving in.	1
Law that goes against the religious beliefs.	1
Person may not necessarily be that way.	1
Understanding perspectives.	1

Table 3.4: Pyramid for the summaries in Figure 3.6 and dialogue 3.1.

3.5 Pyramid Labels Linked to Dialogue Sentences

Repeated elements of the five summaries end up on higher tiers of the pyramid and indicate the most important content. This results in a ranking of the most important arguments (abstract objects) in a dialogue, but the linguistic representation of these arguments is based on the language used in the summaries themselves. To identify the spans of text in the dialogue itself that correspond to the important arguments, we must link the ranked labels from the summaries back to the dialogue text.

We recruited two graduate and two undergraduate students to back-link each sentence

of the dialogue with the best set of human labels from the pyramids, if exists.

In this task, you will carefully read part of a dialog where two people are discussing the issue of gun control. Several previous workers have each summarized this dialogue, and we have related those summaries by grouping together parts of their summaries that roughly describe the same actions in the dialogue. In this task, you will link these action description groups to sentences in the dialogue. Each dialogue is automatically divided into sentences. Your job is to provide the best action description group for each sentence.

The action description groups are sets of sentences from several summaries that essentially describe the same action in the dialogue in different words. Each group has a unique label and you will select the label that best approximates what is happening in the sentence and select a label using the radio button provided with each sentence.

Please especially note:

- More than one sentence can map to same group. For example, two people may say virtually the same thing multiple times.
- Not all sentences will have a good group, so if you cannot find any similar set for a sentence, then select None of the labels match in the radio button option.
- You are expected to read and comprehend the sentence. Since these come from summaries, the action summaries may use very different words from those used in the dialogues.

Table 3.5: Directions for mapping pyramid labels to sentences.

Table 3.5 shows the HIT design and directions for this task. The SCU and it's label were grouped together and annotators were asked to select the most representative group for a given sentence, if exists. As we were primarily interested in the content that bubbles to the top

S.No	ID	Dialogue sentence	Pyramid label selected by annotators	Tier rank
1	GM	If gays are allowed to marry then any kind of marriage between consenting adults should be allowed.	If homosexuals are allowed to marry then any kind of marriage should be allowed.	5
2	GC	One state can not regulate the activities of another state or city, despite what some mayors may believe to the contrary.	Other states cannot be controlled by the rulings of a different state.	5
3	AB	and while the physical damage reasons are valid , they are apparently not what women consider when making their choice .	Physical damage is not a reason for abortion.	
4	GM	Unless you are a beleiver in a religion, then the whole Earthquake thing is irrelevant and if you are then this is n't the debate for you.	Earthquakes have nothing to do with human behavior.	5
5	GC	In other words your argument is that those who are female, or under 18, or over 46, who were no longer physically able to serve in the militia, were barred from ever possessing a firearm, because their status of being armed wasn't in the interest of the government.	(i) People who cannot serve the state should not own firearms. (ii) The right to keep and bear arms only applies to the militia.	Avg(4,5) 4.5
6	GM	Do you not think that we are a family that deserves all the same protections of heterosexual couples?	The family should be afforded the same rights as a heterosexual couple.	4
7	AB	I know for a fact that my wife 's pregnancy was not " unbelievably hard to go through " – she 's told me otherwise and I believe her.	Not all pregnancies are hard to go through, but some might be.	3
8	GC	You have made this argument a number of times before , suggesting the rest of the United States should adopt California standards as national standards.	Propose that the U.S. adopt California's standards.	3
9	GM	Please tell me who forced you into outsourcing children?	Gay couples who have children through surrogates have outsourced their children.	3
10	GM	Do you need to marry your partner to recieve these benefits?	None	<3
11	AB	These are the reasons women give why they choose to have abortions.	None	<3
12	GC	Being the known apologist and unConstitutional advocate , you also misse Danny Rheberg 's amendment to defund any such gathering of information.	None	<3

Table 3.6: Dialogue sentences mapped to pyramid label by the human annotators (gold standard). The table is sorted by the tier rank. GC=Gun Control, AB=Abortion, GM=Gay Marriage.

across all the dialogues, we restricted the label set to top 3 tiers. The sentences that map either to a lower tier or were not present in the summary would be marked as 'None'. This saved a lot of annotation effort and also gave better quality annotations as it restricted the number of labels to be looked at simultaneously for the back-linking task. Each sentence was annotated by a pair of annotators.

We now had one or more labels for each sentence in a dialogue, but we were primarily interested in the **tier rank** of the sentences. A sentence is given the average rank of the tier labels it was linked to by the two annotators. Table 3.6 shows the example sentences mapped to pyramid labels for each of the three topics of Gun control, Abortion, and Gay Marriage and the corresponding rank of each sentence. The sentence ranks give the ground truth labels for a supervised training task for finding the most important arguments in a dialogue.

Reliability: The assessment of inter-annotator reliability quantifies the degree of agreement among corpus annotators. The goal is to demonstrate consistency as well as repeatability among scores provided by multiple annotators. To calculate reliability of the annotations, we calculated the Cohen's kappa between the annotators based on tier rank selected for a sentence. The Cohen's kappa among the annotators was 0.68 for Gun Control, 0.63 for Abortion, and 0.62 for Gay Marriage, which is respectable based on previous studies in argumentation domain. The final corpus with sentence distribution is given in Table 3.7.

Tier Rank	<3	3-4	>4
Topic	Not Important	Important	
Gun Control	1041	786	224
Gay Marriage	1195	957	354
Abortion	1203	688	161

Table 3.7: Sentence distribution in each domain.

Getting a reliable annotated corpora was a significant step and showed that the approach of extracting central propositions and important arguments from the pyramid structure

is feasible, validating $\mathbf{H_1}$, and the next task of building computational models based on above corpora could be further explored.

3.6 Chapter Summary

In this chapter, we presented the development of an annotated corpus that aims to summarize dialogic arguments from social media debates on ideological topics. Unlike previous works in the field, we employed an objective measure for central proposition detector rather than an annotator's perceived importance. We have demonstrated, for the first time, a central proposition extraction method in an argumentative dialogue based on tier rank labels from pyramid annotation.

In the second part of this chapter, we proposed the first annotation study for extracting important arguments using pyramid labels. We introduced a novel corpus with 161 argumentative dialogues on three popular topics: Abortion, Gay Marriage, and Gun Control from on line debate forums. Each dialogue in the corpus is annotated with 5 summaries, pyramid structure and argumentative rank of every sentence. We showed that human annotators substantially agree when back-linking pyramid labels to argumentative sentences in the dialogue. In particular, we obtained an average kappa of 0.64 for the linking task. In fact, it had been mostly unknown at the beginning of this research whether human annotators would be able to perform the linking task as there may not always be substantial word overlap between a label text and the dialogue sentence. The labels are concise, short, and abstracted from the summary while the original dialogue sentences may be redundant and verbose. Therefore, linking sentences to

pyramid labels required a thorough understanding of the meaning and concepts discussed in the dialogue and all the summaries.

Based on this study, we concluded that the central propositions and important arguments in a dialogue as presented in this thesis could be identified reliably by trained annotators. The next step is to generate a computational model based on the above corpus.

In the next chapter, we describe our end-to-end argument summarization framework and the machine learning set up to extract important arguments from social media dialogue.

Chapter 4

Summarizing Important Arguments in a Dialogue

The previous chapter presented a novel corpus of human-written summaries annotated with the pyramid scheme, and pyramid labels linked to dialogue sentences. The top tier pyramid labels were used to reliably rank argumentative sentences in the dialogue. This set of top-ranked sentences can be used as the gold standard set of important arguments in the original dialogues. This paved the way for testing and evaluating the next hypothesis $\mathbf{H_2}$ that aims to identify the features useful for automatic extraction.

As discussed in Chapter 2, the existing summarization frameworks did not test dialogic argument extraction as the labeled dialogue corpora were not available. This chapter addresses that limitation in the following ways:

1. Review the corpus statistics and properties to formalize the argument extraction as a classification task, Section 4.1.

2. Use the ranked set of sentences to explore extractive methods of summarization with sentences as the basic units of summarization, and evaluate the performance of several existing summarizers as baselines, Section 4.2.1.

3. Present new features based on linguistic knowledge and dialogic context to train two different machine learning models, namely an SVM and a Bidirectional LSTM classifier, Section 4.2.2 and Section 4.2.3.

4. Evaluate the model robustness using standard measures of evaluation (F-measure, which is the harmonic mean of precision and recall) showing that our models are better than the baselines in terms of accuracy, Section 4.2.4.

5. Build ablated models with pairwise testing to prove the hypothesis $\mathbf{H_2}$ about contextual features, Section 4.2.4.

6. We conclude with a discussion of results, Section 4.3

4.1 Problem Formulation

The histogram in Figure 4.1 shows the tier rank-frequency distribution of sentences for each of the three topics, Gun Control, Gay Marriage, and Abortion. The proportion of argumentative sentences ("top tier") units in our corpus amounts to 0.48. These numbers are comparable with previous observations that these conversations are interspersed with irrelevant segments, many messages are simple and informal, and claims made by the users may be unclear, ambiguous, vague, or poorly worded. As a consequence, the probability of find-

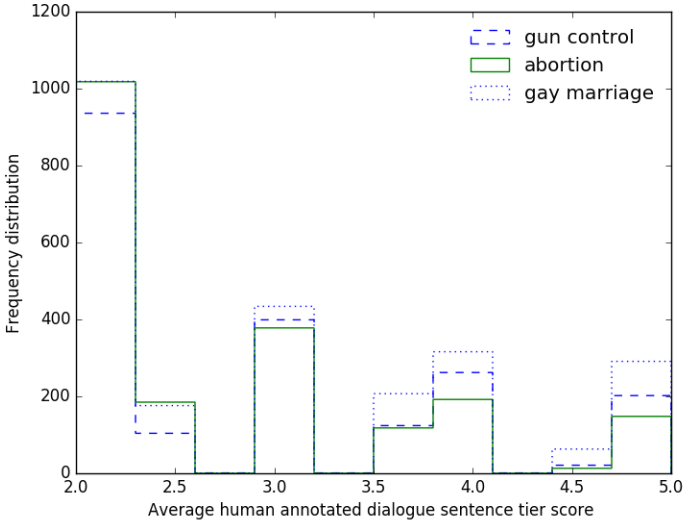


Figure 4.1: Frequency distribution of sentence tier rank distribution.

ing an argumentative sentence in these conversations is low. Indeed, Toni and Torroni (2012) used the term 'bottom-up argumentation' to capture the emergence of arguments and opinions stemming from these types of exchanges. Therefore, it is necessary to first remove these non-argumentative sentences from dialogue, and the problem can be transformed to the selection of the most important argumentative sentences from the dialogue that best represent the issue under discussion. We designated any sentence with an average tier score of 3 or higher as important. Thus, we exploit human summarization and human mapping of labels to sentences to get a well-grounded definition of **important** argument, and the task we address here is binary classification applied to dialogues to select sentences that are important. Table 4.1 shows the resulting number of important sentences for each topic.

We reserved 13 random dialogues in each topic for our test set, using the rest as training. Sentences were automatically split. This led to several sentences consisting essentially

Topic	Important	Not Important
Gun Control	1010	1041
Gay Marriage	1311	1195
Abortion	849	1203

Table 4.1: Sentence distribution in each domain.

of punctuation or numerals, and were without any content. Such sentences were removed by filtering for sentences without a verb and at least three dictionary words. For learning, we created a balanced training and test set by randomly selecting an equal number of sentences for each class, giving the following combinations: 1236 train and 462 test sentences for Abortion, 1578 training and 534 test for Gay Marriage , and 1352 training and 476 test for Gun Control.

4.2 Experimental Method

We first define several baselines using off-the-shelf summarizers such as LexRank and SumBasic (Erkan and Radev, 2004; Nenkova and Vanderwende, 2005). Our experiments explore the effectiveness of combining traditional linguistic and contextual features with Word2Vec in both SVMs and Bidirectional LSTMs. We show that applying coreference, and representing the context improves performance. Performance is overall better for the Bidirectional LSTM, but both the models perform better when linguistic features and contextual features are combined with word embeddings. We achieve a best F-measure of 0.74 for Gun Control, 0.71 for Gay Marriage, and 0.67 for Abortion.

4.2.1 Baseline Performance

We used several off-the-shelf extractive summarization engines (frequency, probability distribution, and graph-based) from the python package *sumy* [1] to provide a baseline for comparison with our models. To enable direct comparison, we defined a sentence as **important** if it appears in the top n sentences in the output of the baseline summarizer, where n is the number of **important** sentences for the dialogue as defined by our method.

SumBasic: Nenkova and Vanderwende (2005) showed that content units and words that are often repeated are likely to be mentioned in a human summary, and that frequency is a powerful predictor of human choices in content selection for summarization. SumBasic uses a greedy search approximation with a frequency-based sentence selection component, and an additional component to re-weight the word probabilities in order to minimize redundancy. Each sentence is assigned a weight equal to the average probability (n/N) of the content words in the sentence, where n is the number of times the word appeared in the input, and N is the total number of content word tokens in the input. SumBasic then selects the best scoring sentence that contains the word with the highest probability. In the next step, the probability of each word that appears in the chosen sentence is reduced to the square of the original probability to reflect the idea that a summary is more likely to contain distinct words. This procedure is repeated until the desired summary length is reached.

KL-SUM: KL divergence is a measure that quantifies the difference between two probability distributions. Using KL divergence, important sentences are ranked based on the 'similarity' between probability distribution of the documents and their summaries. The idea is that the KL

[1] https://pypi.python.org/pypi/sumy

divergence of the original input and a good summary will be low. KL-SUM greedily adds a sentence to a summary as long as it decreases the KL Divergence (Kullback and Leibler, 1951; Haghighi and Vanderwende, 2009; Lin et al., 2011).

LexRank: It uses a ranking approach similar to Page Rank. It is based on the notion that sentences endorse other similar sentences. The importance of a sentence is determined by the importance of the sentences endorsing it. Thus, if a sentence has a high similarity score to other sentences, it is probably of greater importance. It is a graph-based method, where every sentence represents a vertex in the graph, and the edges represent the intra-sentence cosine similarity of TF-IDF vectors. The salience of each vertex in the graph is computed based on the weights of the edges that are connected to it after applying a threshold to the cosine values (Erkan and Radev, 2004).

ID	**Classifier**	**Gun Control**	**Gay Marriage**	**Abortion**
1	KL-SUM (KL)	0.51	0.52	0.47
2	SumBasic (SB)	0.53	0.57	0.49
3	Lex-Rank (**LR**)	**0.58**	**0.58**	**0.59**

Table 4.2: Baselines performance, best model in **bold.**

The evaluation results of each baseline model are shown in Table 4.2. The results indicate that graph-based centrality outperformed the frequency based methods for argumentative dialogue. In an attempt to better understand the error sources, we compared our gold standard summary and the summary sentences selected by Lex Rank as shown in Figure 4.2. LexRank identifies many of the important sentences, but it also includes a number of sentences which

Summary Sentences selected by human annotators
Nobody has been able to prove that gun owners are safer than non-gun owners.
You can play around with numbers to make the problem seem insignificant.
I suppose you could also say that only 3,000 people died in 9/11 and use your logic to say that it 's only a small problem.
Perhaps if somebody in your family had died of gun violence you would have a different attitude.
Nobody has been able to prove that non-gun owners are safer than gun owners.
So if you can not prove things one way or the other why try to infringe on my rights?
I did n't say that it ca n't be proven one way or the other.
I just said you ca n't prove that gun owners are safer.
Using illogic , skewed statistics , revisionist history all in an attempt to violate my constitutional rights , that would be you and other gun grabbers who are trying to infringe on law abiding citizens rights.
Show me in the Constitution where it says that making an illogical argument is a violation of somebody 's rights.
You and your ilk are doing everything in your power to implement your " victim disamament " program in " violation " of my civil rights.
No different than " jim crow " laws and other unconstitutional drivel.

(a) Human selected summary sentences.

Summary sentences selected by LexRank
Show me in the Constitution where it says that making an illogical argument is a violation of somebody 's rights.
Nobody has been able to prove that gun owners are safer than non-gun owners.
I just said you ca n't prove that gun owners are safer.
Wow that is easy.
At least have the courage to say it ... .
Witch hunt.
No different than " jim crow " laws and other unconstitutional drivel.
So if you can not prove things one way or the other why try to infringe on my rights?
Oh, stop your witch hunt.
You can play around with numbers to make the problem seem insignificant.
Using illogic, skewed statistics, revisionist history all in an attempt to violate my constitutional rights, that would be you and other gun grabbers who are trying to infringe on law abiding citizens rights.
I suppose you could also say that only 3,000 people died in 9/11 and use your logic to say that it 's only a small problem.

(b) Lex Rank selected sentences.

Figure 4.2: A comparison of Human Annotator and Lex Rank.

cannot be used to construct a summary such as *'Wow that is easy'*.

Recently there has been a surge in data-driven approaches to summarization based on neural networks and continuous sentence features. An encoder-decoder architecture is the main framework used in these types of models. However, one major bottleneck to applying neural network models to extractive summarization is that the generation systems need a tremendous amount of training data, i.e., documents with sentences labeled as summary-worthy. (Nallapati et al., 2016; Rush et al., 2015; See et al., 2017) used models trained on the annotated version of the Gigaword corpus and paired the first sentence of each article with its headline to form sentence-summary pairs. Iyer et al. (2016) trained an end-to-end neural attention model using LSTMs to summarize source code from online programming websites. Pairing the post title with the source code snippet from accepted answers gives a significant amount of training data that can be used to generate summaries. Such newswire models did not work well here; the neural summarization model from OpenNMT framework (Klein et al., 2017) very often generated <UNK> tokens for our dialogue data. Neural machine translation frameworks use restricted vocabularies. Word embeddings may not capture words that are not frequent in training data. Unknown words are usually replaced with the same symbol such as <UNK>. This may be especially more common with the social media text where words are sometimes misspelled, hence generating these <UNK> tokens.

The baseline outputs, in general, suggest that frequency or graph similarity alone leave room for improvement when predicting important sentences in user-generated argumentative dialogue.

4.2.2 Features

Most formal models of argumentation have focused on carefully crafted debates or face-to-face exchanges. However, the 'bottom-up' argumentative dialogues in online social networks are far less logical (Gabbriellini and Torroni, 2013; Toni and Torroni, 2012), and the serendipity of the interactions yields less rule-governed conversational turns, ones that violate even the rules of naturalistically grounded argument models (Walton and Krabbe, 1995). This makes it difficult to construct useful theoretically-grounded features. In place of that enterprise, we exploit more conventional summarization, sentiment, word class, and sentence complexity features.

We also construct features sensitive to dialogic context. The theoretical literature discusses the ways in which dialogic argumentation shows different speech act uses than in less argumentative genres (Budzynska and Reed, 2011; Jacobs and Jackson, 1992), including the fact that arguments in these conversations are frequently smuggled in via non-assertive speech acts (e.g., hostile questions). Levy et al. (2014) showed that contextual information is crucial to the performance of state-of-the-art argumentation mining tools. Inspired by this, we implemented three basic methods for dialogic context: we extracted the dialogue act tag and some word class class information from the previous sentence; we extracted a rough-grained measure of a sentence's position within a turn; and we used coreference chains to resolve anaphora in a sentence to acquire a (hopefully) more contentful antecedent. Below, we describe these features in more detail.

Word2Vec embeddings: Unsupervised word embeddings are used to learn distributed word

representations and have been successfully used in numerous NLP tasks in recent years. The representation is learned based on the distributional hypothesis: words that occur in the same contexts tend to have similar meaning (Harris, 1954). This enables words that share similar meanings to have similar representations, thus capturing the semantic relationship between words. Mikolov et al. (2013a,b) proposed two efficient techniques, the continuous bag-of-words (CBOW) and Skipgram, to induce unsupervised learning of word representations from large-scale text corpora. Skipgram predicts context words around an input word, while the objective in CBOW is to predict a word given the context words surrounding it across a window of size k. A neural network language model trained on a large corpus captures the semantic and syntactic characteristics of words.

We used the pre-trained word vectors from Google Word2Vec, which are 300 dimensional vectors trained on Google News dataset (about 100 billion words)[2]. Previous work on argument mining has developed methods using Word2Vec that are effective for argument recognition (Habernal and Gurevych, 2015). We created a 300-dimensional vector by filtering stopwords and punctuation and then averaging the word embeddings from Google's Word2Vec model for the remaining words. This vector is then directly used as a feature vector.

Readability Grades: Readability provides quantitative measures to determine quality and text complexity (i.e., vocabulary and syntax). It measures how easily a text is understood by readers of a certain level using surface characteristics. Standard measures were developed that could be easily calculated using factors such as word frequency and length. These could be considered as rough approximations to the linguistic factors that determine text quality (Pitler and Nenkova,

[2] https://code.google.com/archive/p/word2vec/

2008). We hypothesized that contentful sentences were more likely to be complex. To measure that, we used readability grades, which calculate a series of linear regression measures based on the number of words, syllables, and sentences. The readability[3] measures used were: Flesch-Kincaid Score, Automated Readability Index, Coleman-Liau Index, SMOG Index, Gunning Fog Index, Flesch Reading Ease, LIX, and RIX.

LIWC: The Linguistics Inquiry Word Count (LIWC) (Pennebaker et al., 2001) tool has been useful in previous work on stance detection, and we suspected it would help to distinguish personal conversation from substantive analysis. It classifies words into different categories based on thought processes, emotional states, intentions, and motivations. For each LIWC category, we computed an aggregate frequency score for a sentence. Using these categories, we aim to capture both the style and the content types in the argument. Style words are linked to measures of people's social and psychological worlds while content words are generally nouns and regular verbs that convey the content of a communication. To capture additional contextual information, we computed the LIWC score of the previous sentence.

Sentiment: Sentiment features have shown to be useful for argumentative quality assessment and here too we suspected that name-calling and the like could be flagged by sentiment features. We used the Stanford sentiment analyzer from Socher et al. (2013) to compute five sentiment categories (very negative to very positive) per sentence.

Dialogue Act of Previous Sentence (DAC): We hypothesized that **important** sentences may be more likely in response to particular dialogue acts, like questions, e.g., a question might be followed by an explanation or an answer. To identify if a previous sentence was a question,

[3] https://pypi.python.org/pypi/readability

we combined the tags into two categories indicating whether the previous sentence was a question type or not. We implemented a binary PreviousSentAct feature which used Dialogue Act Classification from NLTK (Loper and Bird, 2002).

Sentence position: Position of a sentence in a document has been used as an indicator of sentence importance for sentiment and email summarization (Beineke et al., 2004; Rambow et al., 2004). Hence, it follows that relative sentence position could benefit the argument extraction process as well. We included features that encode the absolute position of the sentence within the document. We divide a turn into thirds and create an integral feature based on which third a sentence is located in the turn.

Coref: Interpretation of many language expressions in a conversation depends upon entities in context, i.e., previous parts of the text. In the hope that coreference resolution would help ground utterance semantics, we replaced anaphoric words with their most representative mention obtained using Stanford coreference chain resolution (Manning et al., 2014).

4.2.3 Machine Learning Models

Numerous machine learning algorithms have been applied to prediction or classification task of identifying argumentative segments in the text. These include, but are not limited to, Naive Bayes, Logistic Regression, Support Vector Machines. Feature engineering plays a vital role in determining the performance of these systems and hence a significant effort is invested in deriving features that represent the text.

Recently, neural network based approaches have lead to major breakthroughs and successes in tasks related to text classification. One of the key advantages of these deep net-

works is end-to-end learning manner and the ability to automatically learn and train itself from the information provided, and optimize its weights, thus alleviating the need for hand-crafted features. The two most common deep learning frameworks are Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN).

A (CNN) is a feed-forward neural network with convolution layers interleaved with pooling layers. Pooling reduces the output dimensionality but keeps the most salient information, analogous to feature detection. Kim (2014) experimented with CNN trained on top of pre-trained word vectors for sentence-level classification tasks. The model achieved remarkable results on multiple benchmarks, improved upon the state of the art on 4 out of 7 tasks, including sentiment analysis and question classification.

RNNs have been proven to be efficient in constructing sentence representations and perform well for a broad range of tasks including sentiment classification, relation classification, and paraphrase detection (Socher et al., 2011a,b; Yin et al., 2017). RNN is sensitive to word order as compared to CNN, which is invariant to location. RNNs can preserve historical information and are better at handling variable-length sequences. However, RNNs suffer from the problem that later words make more influence on the final text representation than former words. We used two machine learning models, the first is a traditional SVM, while the other is a combination Bidirectional LSTM based on deep networks.

SVM. We used Support Vector Machines with a linear Kernel from Scikit-learn (Pedregosa et al., 2011) with our manually engineered features based on linguistic knowledge and contextual information, which are speculated to be useful for representing an argument. SVM is highly sensitive to the choice of the regularization parameter C. Therefore, in order to find the optimal

parameters, we performed a grid search using cross-validation set. The reason why we choose linear SVM for this purpose is that they are robust against noise and outliers, considered one of the state-of-the-art algorithms for text classification systems, and are very fast to train.

CNN + BiLSTM with Glove embeddings. CNN can learn spatial features from a local group of neighbors but lacks the history sensitivity required to learn sequential correlations; on the other hand, RNN is specialized for sequential modeling but incapable of extracting features in a parallel (Zhou et al., 2015). A combination of a CNN and an RNN has been used for sentence representations (Wang et al., 2016) where CNN is able to learn the local features from words or phrases in the text, and the RNN learns long-term dependencies. The model utilizes CNN to extract a sequence of higher-level phrase representations, which are fed into a long short-term memory RNN (LSTM). Using this as a motivation, we include a convolutional layer and max pooling layer before the input is fed into an RNN.

GloVe Embeddings: Word embeddings are often used as the first data processing layer in a deep learning model. A GloVe is an unsupervised algorithm for obtaining vector representations for words. It is based on global word-word co-occurrence statistics. The co-occurrence matrix of words and contexts are factorized into a lower dimensional matrix to extract representations for each word in the vocabulary (Pennington et al., 2014). We used GloVe embeddings to initialize our Long Short-Term Memory (LSTM) models as GloVe embeddings have been trained on web data, and in some cases work better than Word2Vec (Stojanovski et al., 2015).

Figure 4.3 shows the deep learning model architecture. It is based on a combination of Convolutional Neural Network and Bidirectional LSTM. The convolution layer takes as input the GloVe embeddings. These pre-trained word embeddings are 100-dimensional vectors, and

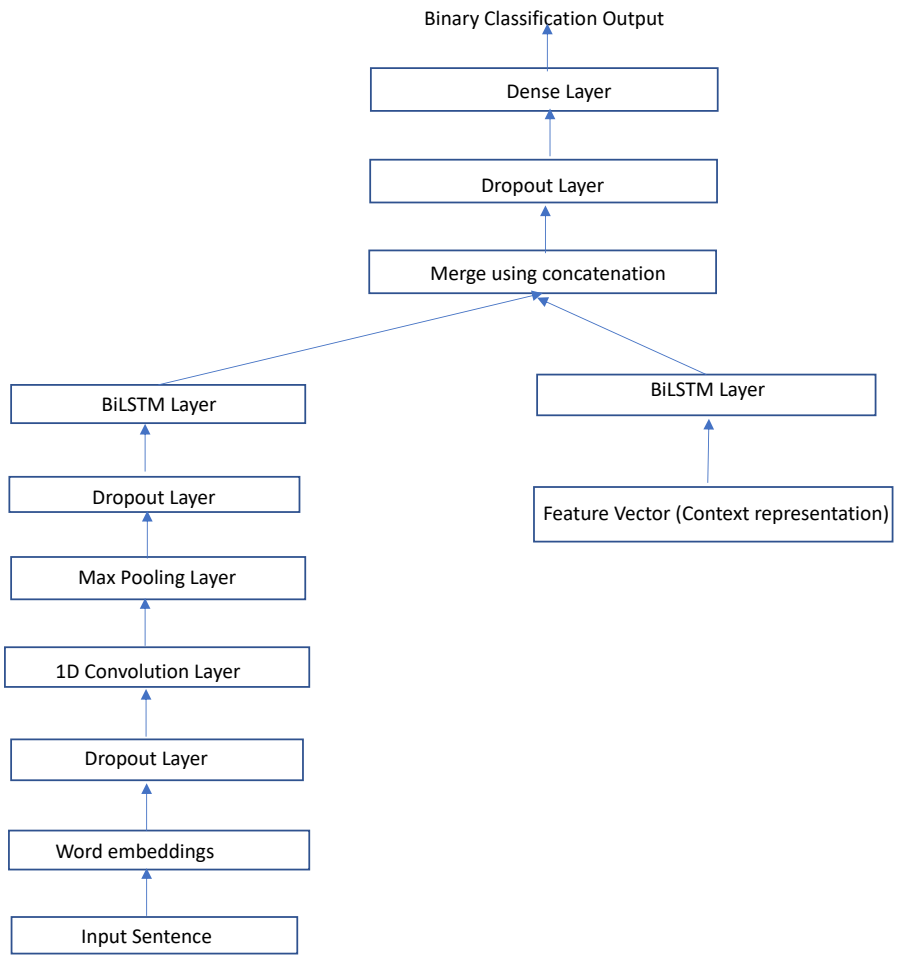


Figure 4.3: CNN-BiLSTM Architecture

each sentence is represented as a concatenation of word vectors. To prevent overfitting we use a dropout layer with a dropout rate of 0.2. The model consists of a 1D convolution layer of window size 3 and 32 different filters. Padding is set to a value same so that size of output is same as that of input. A 1-max pooling is performed i.e., the largest number from each feature map is recorded giving a fixed-sized feature vector. A bidirectional LSTM layer is stacked on the convolution layer and then concatenated with another layer of bidirectional LSTM to

encode the features and feature combinations. The merged outputs are fed through a sigmoid layer for binary classification. LSTM creates a validation set by a 4 to 1 random selection on the training set. The model is optimized using the Adam (Kingma and Ba, 2014) optimizer. The deep network was implemented using the Keras package (Chollet, 2015).

4.2.4 Evaluation

We use standard classification evaluation measures based on Precision/Recall and F-score. Performance evaluation uses weighted average F-score on test set. We first evaluate simple models based on a single feature.

Simple Ablation Models. Table 4.3, Row 1 show the results for Lex Rank, the best baseline system across all topics. Overall the F-score is low, indicating that summarizers aimed at newswire or monologic data do not work on argumentative dialogue. Row 2 shows that Dialogue Act Classification works better than the random baseline for Gun Control and Gay Marriage, but not for Abortion. Row 3 shows that Word2Vec improves over the baseline, but this did not work as well as it did in previous research (Habernal and Gurevych, 2015). One reason could be that averaged Word2Vec embeddings for each word lose too much information in long sentences. Interestingly, Row 6 shows that sentiment by itself beats LexRank across all topics, suggesting a relationship of sentiment to argument that could be further explored. Each Row has an additional column for each topic indicating what happens when we first run Stanford Coreference replacing each pronoun with its most representative mention. The results show that coreference improves the F-score for both Gun Control and Abortion. LIWC categories and Readability perform well across topics.

			Gun Control		Gay Marriage		Abortion	
ID	**Classifier**	**Features**	**F-score.**	**F-score Coref**	**F-score**	**F-score Coref**	**F-score**	**F-score Coref**
1	Baseline	Lex-Rank (**LR**)	0.58		0.58		0.59	
2	SVM	Dialog Act (**DAC**)	0.61	0.60	0.58	0.58	0.42	0.41
3	SVM	Word2Vec	0.65	0.65	0.63	0.56	0.58	0.58
4	SVM	Readability (**R**)	0.64	0.67	0.68	0.68	0.63	0.64
5	SVM	LIWC current sentence (**LC**)	0.72	**0.74**	**0.69**	0.66	0.64	0.63
6	SVM	Sentiment (**SNT**)	0.66		0.62		0.61	
7	SVM	Sentence Turn (**ST**)	0.61	0.61	0.40	0.40	0.33	0.33
8	Bi LSTM		0.68	0.69	0.63	0.58	0.64	**0.65**
		Feature Combinations						
9	SVM	LIWC current + previous (**LCP**)	0.73	0.72	0.66	0.67	0.61	0.61
10	SVM	LCP + R	0.73	0.73	0.70	0.68	0.61	0.60
11	SVM	R+DAC	0.65	0.66	0.68	0.68	0.63	0.63
12	SVM	LCP + DAC + R	0.72	0.73	0.69	0.68	0.61	0.61
13	Bi LSTM	DAC	0.67	0.68	0.69	0.65	0.65	0.66
14	Bi LSTM	ST	0.66	0.66	0.61	0.67	0.64	0.52
15	Bi LSTM	LCP	0.70	0.68	0.52	0.52	0.65	**0.67**
16	Bi LSTM	R	0.70	0.70	0.59	0.63	0.65	0.66
17	Bi-LSTM	LCP+ DAC	0.70	0.71	0.69	0.68	0.61	0.62
18	Bi-LSTM	R+ DAC	0.70	0.68	0.63	0.62	0.60	0.64
19	Bi-LSTM	R+ LCP	0.69	0.68	**0.71**	0.67	0.64	0.66
20	Bi-LSTM	LCP+R +DAC	0.73	**0.74**	0.70	0.69	0.62	0.63

Table 4.3: Results for classification on test set for each topic. Best performing model in **bold.**

Feature Combination Models. We first evaluate SVM with different feature combinations, with details on results in Table 4.3. For the Gun Control topic, LIWC categories on the current sentence give an F-score of 0.72. Adding LIWC from the previous sentence improves it to 0.73 (rows 5 and 9, without coref column). In contrast, just doing a coref replacement improves LIWC current sentence score to 0.74 (row 5 for Gun Control, with and without coref columns). A paired t-test on the result vectors shows that coref replacement provides a statistically significant improvement at ($\mathbf{p < 0.04}$). For the Abortion topic, the overall performance is low as compared to the other two topics suggesting that arguments used for Abortion are harder to identify. Both DAC, Word2Vec scores are quite low, but Readability and LIWC do better.

The LSTM models on their own do not perform better than SVM across topics, but adding features to the LSTM models improves them beyond the SVM results. We paired only LSTM (row 8) separately with the best performing model in bold for each topic in Table 4.3 to evaluate if the combination is significant. Paired t-tests on the result vectors show that the differences in F-score are statistically significant when we compare LSTM to LSTM with features for each topic ($\mathbf{p < 0.01}$) for all topics, indicating that adding contextual features makes a significant improvement. Adding LIWC categories from current and previous utterances to LSTM also improves performance for Gun Control and Abortion. For the Gay Marriage topic, LSTM combined with LIWC and readability works better than LSTM alone. This proves the hypothesis $\mathbf{H_2}$ that adding contextual information from neighboring sentences and linguistic knowledge derived from LIWC categories improves argument extraction.

4.3 Analysis and Discussion

To qualitatively gain some insight into the limitations of some of the systems, we examined random predictions from different models. One reason that a Graph-based system such as LexRank performs well on DUC datasets might be that DUC data sets are clustered into related documents by human assessors. To observe the behavior of the method on noisy data, the authors of LexRank added random documents to each cluster to show that LexRank is insensitive to some limited noise in the data. However, topic changes are more frequent in these debate dialogues and they contain content that is not necessarily related to the argumentative purpose of the dialogue.

Many previous summarization algorithms such as LexRank have been developed based on the idea that sentences reinforce other similar sentences. Thus, if something is written twice, it is likely to be more important. This may be true for well-edited documents such as newswire domain but not for dialogue, which is naturally more redundant. For example, based on lexical overlap, LexRank highly ranked the following two irrelevant sentences. These sentences do not provide any useful information about the topic.

(i) *Well it's not going to work.*

(ii) *Get to work!*

One reason that SVM with sentiment features performs well is that positive sentiment predicts the not-important class. Phatic communication and conversational fillers are often used in everyday dialogue exchanges and represent a social function of language rather than any contentful information. Sentiment analyzers predict phatic features of dialogues as positive whereas negative sentiment identified the more argumentative class, as shown by the following

examples:

Positive sentiment

- *Here ya go.*
- *Yes it does.*
- *Sounds right to you?*

Negative sentiment

- *Otherwise you are discriminating against the person wanting to enter into the marriage , however distasteful it appears to you.*
- *The 10th Amendment says that those rights not specifically given to the Feds belong to the State or individuals.*
- *If we want to talk about the world the over 90% of the world doesn't recognize homosexuality.*

The results show that LIWC performs well and that LIWC used to represent context performs even better, ($\mathbf{H_2}$). To understand which LIWC features were important, we performed chi-square feature selection over LIWC features on the training set. Content categories were highly ranked across topics, suggesting that the LIWC features were being exploited as a form of within-topic topic detection; this suggests that more general topic modeling could help results.

Table 4.4 shows the top 5 LIWC categories for each topic based on chi-square based feature selection on the training set for all the three topics. Unsurprisingly, across all topics, the LIWC marker of complexity (Words Per Sentence) appears. In addition, many other topics link commonsense with important facets of these debates – the opposition to Abortion between questions of the sanctity of life (biological processes), and the health of individuals involved.

Similarly, with Gay Marriage, we see sides of the debate between personal relationships (family, affiliation) and questions of sexual practice (sexual, drives). The case of Gun Control is somewhat surprising, since one might expect to see LIWC categories relating to life and safety. Instead, we see the money category coming from discussions about gun buy back and gun prices.

To understand better why coreference resolution was helping, we also examined cases where coreference matters. Coreference resolution can also interact with different features such as LIWC, where it can change the frequency distribution of categories in the text. Performing a corefence resolution moves a word from the pronoun to some other category. For example, replacing *'it'* by *'Government'* decreases Impersonal Pronouns and Total Pronouns, while increasing Six Letter Words. In several cases, these replacements produce correct predictions, e.g.

Only if it is legal to sell it.

Topic	LIWC Categories
Abortion	*Biological Processes, Health, Second Person, Sexual, Words Per Sentence.*
Gun Control	*First Person Singular, Money, Second Person, Third Person Plural, Words Per Sentence.*
Gay Marriage	*Family, Sexual, Words Per Sentence, Affiliation, Drives.*

Table 4.4: Top 5 LIWC categories by chi-square for each topic

4.4 Chapter Summary

The annotated argument dialogue corpus introduced in Chapter 3 enabled us to test the validity of our proposed models using linguistic and contextual features, and compare our results to existing summarization frameworks. In this chapter, we presented a novel method for argument summarization of dialogue exchanges from social media debates with our results significantly beating the traditional summarization baselines. The core idea is based on using information from both current and neighboring sentences in the dialogue for argument extraction and importance rather than traditional document-oriented frequency assessment.

The argument extraction task was modeled as a binary classification task, and evaluated using the standard measure of weighted F-score. We tested two different models, an SVM and a BiLSTM, and conducted experiments with different feature combinations to infer which feature combinations are the most predictive. Both the models perform better than the baseline approaches, indicating the robustness of the proposed method. Feature ablation experiments showed that categories from LIWC were most effective for extracting the most important arguments from a dialogue, across all topics. Some generic features as captured by readability and sentiment were also useful. This direction can be further explored in future work to develop cross-domain models. Finally, in order to achieve the best performance, contextual information obtained through features derived from previous sentences is a necessary supplemental to linguistic dimensions of the current sentence obtained from LIWC, thus validating our hypothesis $\mathbf{H_2}$.

Given the central propositions, and the important arguments in a dialogue; we move to

the next research question about the actual issues and facets discussed in these debates. The next chapter describes our approach to finding these facets that often manifest as similar repeated arguments across these discussions.

Chapter 5

Argument Facet Similarity

In this chapter, we address the research question on abstract issues and objects in a discussion about social or political topics. Particularly, we determine the specific arguments and counter arguments that emerge from and reflect the dynamic nature of argumentative dialogue. This view of argumentation goes beyond the examination of arguments in isolation and analyzes them in relation to other arguments addressed in all the dialogues in a discussion.

The main motivation is that all the conversants in a discussion make some contribution towards the topic, and summarizing the entire range of perspectives and ideas on both sides gives a complete view of the topic. When people converse about social or political issues, similar arguments are often paraphrased by different speakers, across many different conversations. For example, consider the dialogue excerpts in Figure 5.1 from the 89K sentences about gun control in the IAC 2.0 corpus of online dialogues (Abbott et al., 2016). Each of the sentences **S1** to **S6** provide different linguistic realizations of the same proposition namely that *Criminals will have guns even if gun ownership is illegal*.

S1: To inact a law that makes a crime of illegal gun ownership has no effect on criminal ownership of guns.
S2: Gun free zones are zones where criminals will have guns because criminals will not obey the laws about gun free zones.
S3: Gun control laws do not stop criminals from getting guns.
S4: Gun control laws will not work because criminals do not obey gun control laws!
S5: Gun control laws only control the guns in the hands of people who follow laws.
S6: Gun laws and bans are put in place that only affect good law abiding free citizens.

Figure 5.1: Paraphrases of the *Criminals will have guns* facet from multiple conversations.

Debate websites, such as Idebate and ProCon produce curated summaries of arguments on the Gun Control topic, as well as many other topics.[1] [2] These summaries typically consist of lists, e.g., Figure 5.2 lists eight different aspects of the Gun Control argument from Idebate. Such manually curated summaries identify different linguistic realizations of the same argument to induce a set of common, repeated, aspects of arguments, what we call **argument facets**. For example, a curator might identify sentences **S1** to **S6** in Figure 5.1 with a label to represent the facet that *Criminals will have guns even if gun ownership is illegal*.

Figure 5.3 illustrates a different type of summary for the Death Penalty topic from ProCon, where the argument facets are called out as the "Top Ten Pros and Cons" and given labels such as `Morality`, `Constitutionality`, and `Race`. See the top of Figure 5.3. The bottom of Figure 5.3 shows how each facet is then elaborated by a paragraph for both its Pro

[1] http://debatepedia.idebate.org/en/index.php/Debate:_Gun_control
[2] http://gun-control.procon.org/

Pro Arguments
A1: The only function of a gun is to kill.
A2: The legal ownership of guns by ordinary citizens inevitably leads to many accidental deaths.
A3: Sports shooting desensitizes people to the lethal nature of firearms.
A4: Gun ownership increases the risk of suicide.
Con Arguments
A5: Gun ownership is an integral facet of the right to self defense.
A6: Gun ownership increases national security within democratic states.
A7: Sports shooting is a safe activity.
A8: Effective gun control is not achievable in democratic states with a tradition of civilian gun owership.

Figure 5.2: The eight facets for Gun Control on Idebate, a curated debate site.

and Con side, showing the summary for the `Morality` facet here.

These summaries are curated, thus one would not expect that different sites would call out the exact same facets, or even that the same type of labels would be used for a specific facet. As we can see, ProCon (Figure 5.3) uses one word or phrasal labels, while Idebate (Figure 5.2) describes each facet with a sentence. Moreover, these curated summaries are not produced for a particular topic once-and-for-all: the curators often re-organize their summaries, drawing out different facets, or combining previously distinct facets under a single new heading. We hypothesize that this happens because new facets arise over time. For example, it is plausible that for the Gay Marriage topic, the facet that *Gay marriage is a civil rights issue* came to the

Figure 5.3: Facets of the Death Penalty debate as curated on ProCon.org

fore only in the last ten years.

Our aim is to produce summaries similar to these curated summaries, but automatically, and over time, so that as new argument facets arise for a particular topic, we can identify them. Recognizing the **facets** of an argument automatically entails at least two subtasks.

- **Task1: Argument Extraction**: How can we extract sentences from a dialogue that

clearly express a particular argument facet?

- **Task2: Argument Facet Similarity (AFS)**: How can we recognize that two sentential arguments are semantically similar, i.e., that they are different linguistic realizations of the same facet of the argument?

PostID:Turn
S1:1 Agreed She is ignoring my religious freedom and trying to **institute her religion into law. The law that will bar my family from legal protections. It won't protect her marriage but will bar me and my people from from being full citizens**. She isn't protecting marriage but perserving her heterosexual privledge.
S2:1 How on earth is she **impeding on you religious freedom?** She isn't trying to take away your right to any religious ceremony. With such a wide-open standard of what constitutes religious freedom that you seem to have, any legislation could be construed as imposing on religious freedom.
S1:2 Because it is her religious belief that **marriage is between a man and a woman.** My religious belief is that **marriage is between two people that love each other regardless of sex**. She is tying to place her religious belif into law over mine. Who gets hurt here? If my religious belief is put into law she can still marry the person of her choice. If her religious belief gets put into law she can still marry the person of her choice but I do not get to. So I and my people are hurt by codifing her religious belief into law. She is trying to keep gay people out of marriage and thus preserve her heterosexual privledge.
S2:2 But by that definition, either one could be viewed as impeding on religious freedom, including your view impeding on hers ! We don't define imposing on religious freedom on the basis of having different ideals. **It doesn't effect your religion or religious freedom if you don't get benefits under gay marriages**. You can argue in other ways, on other basis, but the idea that not giving gays marriage benefits is imposing on religious freedom is an empty " argument ".

Figure 5.4: Excerpt from Gay Marriage (Dialogue-1 in Figure 3.1).

Task1 is needed because social media dialogues consist of many sentences that either do not express an argument, or cannot be understood out of context. Thus, sentences that are useful for inducing argument facets must first be automatically identified. Hence, there must be a system that can extract the most essential arguments for a given discussion, or **central**

PostID:Turn
S1:1 Certainly not yours. You should know that I am for **no marriage in government.** It should be left to a religious institution where it will actually mean something. The states should then go back to doing something that actually makes sense and doesn't reward people like Britney Spears for being white trash.
S2:1 That is all well and good, but it is not the religious ceremony and sanction that gays are looking for. They already have that; there are churches that perform same-sex marriages. **It is the civil benefits that are at issue. Are you saying you would be in favor of foregoing ALL the legal rights and benefits you are afforded by marriage?** For example: *Assumption of Spouse's Pension *Automatic Inheritance *Automatic Housing Lease Transfer *Bereavement Leave.... What do you say?
S1:2 yeah I know. I'm saying that there should be a better system. For example, if you had a best friend who you are roommates with... both hetero for the sake of argument... and never wish to get married then **could they get some of the benefits you described**?

Figure 5.5: Excerpt from Gay Marriage (Dialogue-2 in Figure 3.5).

propositions of a conversation. Important arguments in Figures 5.4 and 5.5 are provided in bold.

Task2: There must be another system, the **Argument Facet** inducer, that relates these conversation-specific arguments to each other in terms of facets, e.g. that identifies the two specific central propositions in Figures 5.4 and 5.5 about legal protections and civil benefits as the same (abstract) **facet**, namely that same-sex marriage is about getting the civil rights benefits of marriage.

Related work on argument mining (discussed in more detail in Chapter 2.2.2) defines a finite set of facets for each topic, similar to those from Idebate in Figure 5.2 or from ProCon in Figure 5.3. Previous work then labels posts or sentences using these facets, and trains a classifier to return a facet label (Conrad et al., 2012; Hasan and Ng, 2014; Boltuzic and Šnajder, 2014; Naderi and Hirst, 2016), *inter alia*. However, this simplification may not work in the long

term, both because the sentential realizations of argument facets are propositional, and hence graded, and because facets evolve over time, and hence cannot be represented by a finite list. This chapter addresses these limitations.

We created two different datasets for training the AFS. The first dataset comprised of human-generated pyramid label pairs from summaries of dialogic arguments for a single topic of Gay Marriage. These were concise, grammatical, and well structured. It served as a starting point for the AFS, especially since the task was novel and the process of argumentation itself is challenging. AFS annotations were collected using Mechanical turk. The average score across all the annotators correlated at 0.7 with our gold standard annotation indicating that the task was clear and accurately specified. Considering the high reliability of the overall approach, we decided to construct and test AFS on another, much harder dataset, where we automatically extracted arguments of high quality from social media dialogues. In what follows, we provide a description of both the AFS datasets constructed using the above two approaches, and the experiments performed. Specifically, the overall procedure of finding these facets consists of the following steps:

1. Problem formulation that takes pairs of sentences from Task1 and then learns a regressor that can predict Argument Facet Similarity **(AFS)**, Section 5.1.

2. Compile two novel datasets for the AFS task. The first dataset input is created using human labels from pyramid annotations (**central propositions** as described in Chapter 3) as important arguments while the second contains actual high quality arguments from social media posts. Getting reliable annotations is an important goal here to ensure that

the AFS task is clear to the annotators, and can be repeated, Sections 5.2.1 and 5.3.2 .

3. Describe two different regression models, one for each dataset. Propose new features better suited to the AFS task including features based on Word2Vec, LIWC and dependency trees. Performance evaluation using correlation coefficient statistic and root mean squared error. Perform experimental comparisons on both the models that demonstrate that the proposed set of features significantly outperforms the STS baseline, validating hypothesis $\mathbf{H_3}$, Sections 5.2.2 and 5.3.3

5.1 Problem Formulation

We hypothesized that we can use important arguments to induce facets of a topic across a set of dialogues. We treat a cluster of **important arguments** as a facet label, just as a synset concept in WordNet is labeled by its members. Calculating similarity is a central aspect of AFS. At this point, we could treat a facet as a cluster of 'paraphrased important arguments' or 'more or less similar but graded arguments'. Our corpus analysis showed that many a times these arguments are similar but not exactly same. A similar observation was also reported in a previous study by Boltuzic and Šnajder (2015) who performed clustering based on similarity and found that it is difficult to draw clear-cut boundaries between arguments based only on textual similarity (STS). Hence, we framed the semantic equivalence as a continuous value rather than a binary one. The goal is to define a similarity metric and train a regression model that takes as input two important arguments and returns a scalar value that predicts their similarity (AFS). The model must reflect the fact that similarity is graded, e.g. the same argument facet

may be repeated with different levels of explicit detail. For example, sentence A1 in Figure 5.2 is similar to the more complete argument, *Given the fact that guns are weapons—things designed to kill—they should not be in the hands of the public*, which expresses both the premise and conclusion. Sentence A1 leaves it up to the reader to infer the (obvious) conclusion.

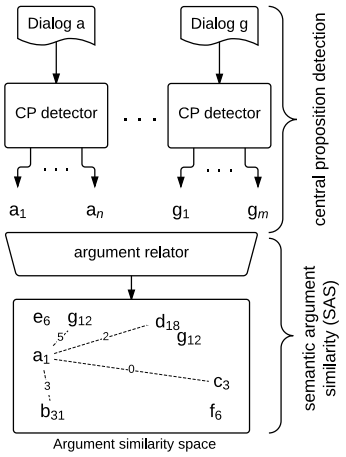
5.2 AFS with Pyramid Labels

The process of automatic dialogue extraction, human summarization, and human SCU generation (graphically depicted in Figure 5.6b) is how we arrive at the **central propositions or important arguments** for the **Argument Facet Similarity** task. The pyramid structure directly reflects the content that the annotators deem most important in the original dialogue. We are interested in the content that bubbles to the top across all the dialogues and take the Tier 3 and above SCUs as our **central propositions**.

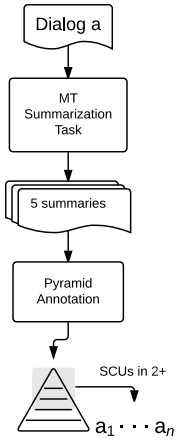
5.2.1 AFS Corpus with Pyramid Labels

For our first set of experiments, we used 45 Gay Marriage dialogue pyramids and extracted the labels of those SCUs that had a tier rank 3 and above. This gave a total of 329 SCU labels. As described below (and sketched in Figure 5.6c), we used Mechanical Turk to provide similarity scores between pairs of SCU central propositions. Although, in principle, we could have asked about all possible pairs of the 329 **central propositions**, most pairs are likely to be unrelated, and so we used an initial clustering algorithm to help reduce the work and cost.

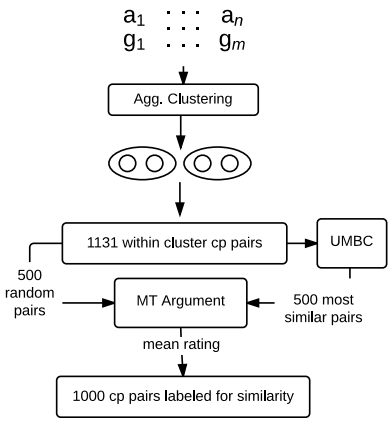
Agglomerative Clustering: To group similar arguments, we performed clustering across our



(a) Basic engineering approach for extracting central propositions and clustering them into argument **facets** across several dialogues;



(b) Workflow for detecting central propositions via pyramid evaluation of multiple summaries;



(c) Workflow for obtaining gold-standard labels for AFS task.

Figure 5.6: The overall engineering architecture of our approach.

329 labels. We performed Agglomerative Clustering using Scikit-learn (Pedregosa et al., 2011), (Agg Clustering in Figure 5.6c). It recursively merges the pair of clusters that minimally increases a given linkage distance. We used cosine similarity as the distance measure with average linkage criteria. The linkage criterion is used to determine the distance to use between sets of observations. To focus on topic-specific cues, the clustering was performed using only nouns, verbs and adjectives. After generating all pairwise combinations within a cluster, this approach yielded 1131 argument pairs used in the Mechanical Turk AFS task. See Figure 5.6c.

Instructions
These HITS are part of an ongoing research project that is attempting to analyze the arguments presented in online debates. Each pair of phrases below is taken from an online, public debate forum. The two phrases may be from author's communicating directly with each other, or they may be from completely different discussions. We would like you to classify each of the following sets of pairs based on your perception of how SIMILAR the arguments are, on the following scale, examples follow.
(5) Completely equivalent, mean pretty much exactly the same thing, using different words.
(4) Mostly equivalent, but some unimportant details differ. One argument may be more specific than another or include a relatively unimportant extra fact.
(3) Roughly equivalent, but some important information differs or is missing. This includes cases where the argument is about the same FACET but the authors have different stances on that facet.
(2) Not equivalent, but share some details. For example, talking about the same entities but making different arguments (different facets)
(1) Not equivalent, but are on same topic
(0) On a different topic

Facet: A facet is a low level issue that often reoccurs in many arguments in support of the author's stance or in attacking the other author's position. There are many ways to argue for your stance on a topic. For example, in a discussion about the death penalty you may argue in favor of it by claiming that it deters crime. Alternatively, you may argue in favor of the death penalty because it gives victims of the crimes closure. On the other hand you may argue against the death penalty because some innocent people will be wrongfully executed or because it is a cruel and unusual punishment. Each of these specific points is a facet.
For two utterances to be about the same facet, it is not necessary that the authors have the same belief toward the facet. For example, one author may believe that the death penalty is a cruel and unusual punishment while the other one attacks that position. However, in order to attack that position they must be discussing the same facet.

Figure 5.7: Instructions for AFS MT HIT.

MT Argument Facet Similarity HIT: Figure 5.7 shows the instructions defining AFS for the

MT HIT. Inspired by the scale used for Semantic Textual Similarity (STS) (Agirre et al., 2012), we collected annotations on a 6 point scale. One crucial difference in our formulation was a desire to capture similarity in **facet** and **argument** simultaneously. The use of the value 3 for 'same facet, contradictory stance' was a well-thought decision in the definition of AFS. Just as two words can only be antonyms if they are in the same semantic field, two arguments can only be contradictory if they are about the same facet. Thus, we instruct annotators to give a score of 3 to opposing arguments on the same facet. A score of 3 or more than 3 represents argument pairs on the same facet. Since clustering is not perfect, it is possible to have arguments from different facets within the same cluster. Such pairs get a score of less than 3.

The task was put on Mechanical Turk using two separate batches. For the first batch, we randomly selected 500 pairs from our pairs dataset of 1131 pairs. However, our subsequent impression was that the clustering had not filtered out enough of the unrelated pairs (score 0-1). For the second batch, we applied an off-the-shelf system that calculates STS as a filter on pairs of central propositions in an attempt to bias the distribution of the training set for our system to have a much larger set of more similar pairs. However, the AFS task is clearly different than STS, partly because the data is dialogic and partly because it is argumentative. We selected the top 500 pairs according to the UMBC similarity score (Han et al., 2013).

This gave us a final pair dataset of 1000 pairs. Since AFS is a novel and subjective task, workers took a qualification test. Then each pair was annotated by 5 workers, and one of the authors provided gold standard labels. The HIT allowed 5 AFS judgements per hit, thus the number of pairs annotated by a worker varied from 5 to 1000. To increase reliability, we removed the annotations from those workers who had attempted less than 4 hits (20 pairs) and

had the lowest pairwise correlations with our gold standard annotation. Our final AFS score was the average score across all the annotators. The final AFS score correlated at 0.7 with our gold standard annotation, showing that the AFS similarity task is well-defined, and understandable by minimally trained annotators on MT. We discuss AFS values and features in the Section 5.2.2 below.

5.2.2 Experiments

Given the data collected above, we defined a supervised machine learning experiment with AFS as our dependent variable and different collections of features inspired from previous work as our independent variables.

5.2.2.1 Features

NGRAM overlap. This is our primary baseline. For each argument, we extracted all the unigrams, bigrams and trigrams, and then counted how many were in overlap across the two arguments. For unigrams, we did not include stop words. Stemmed n-grams were used to get better overlap.

UMBC. This is our secondary baseline. This feature is the Semantic Textual Similarity obtained using UMBC Semantic Similarity tool (Han et al., 2013).

DISCO Distributionally Similar Category. We used the distributional similarity tool DISCO with the pre-computed English Wikipedia word space (Kolb, 2008). We extracted the top 5 distributionally similar nouns, verbs, and adjectives for each argument. For each argument pair, three vector pairs (over nouns, verbs, and adjectives) were created with this extended

vocabulary. Stemming was performed, and cosine similarity between these vector pairs was calculated.

LIWC Category. This feature set is based on the Linguistics Inquiry Word Count tool (Pennebaker et al., 2001). To tune these features, we first used a set of Gay Marriage posts from websites such as CreateDebate and ConvinceMe to extract relevant LIWC categories. We supplemented this data with Gay Marriage posts from 4forums, but excluded the discussion threads in our dialogue corpus. From this data, we extracted the LIWC categories for the most frequent nouns, verbs, and adjectives. For the verbs category, we excluded the verbs present in the NLTK stop word list. We retained only semantically rich categories such as Biological Processes, Causation, Cognitive Processes, Humans, Negative Emotion, Positive Emotion, Religion, Sexual, and Social Processes. The score for this set was the LIWC category overlap count across pairs for each category.

ROUGE Scores. ROUGE is a family of metrics used to determine the quality of a summary by comparing it to other ideal summaries (Lin, 2004). It is based on a number of overlapping units such as n-gram, word sequences, and word pairs. This feature includes all of the rouge f-scores available via the package `https://pypi.python.org/pypi/pyrouge/0.1.0`.

5.2.2.2 Results

Given the **central proposition pairs** from the CP detector, we needed to train an argument **Facet** inducer. We define AFS as a regression problem and evaluate support vector regression and linear regression for 10-fold cross validation using the Weka machine learning toolkit (Hall et al., 2005).

Classifier	**RMS**	**MAE**	**R**
SMO	1.0208	0.8019	0.532
Linear Regression	0.9996	0.8003	0.540

Table 5.1: Support Vector and Linear Regression.

RMS: Root Mean Squared Error, MAE: Mean Absolute Error, R: Correlation Coefficient.

Table 5.1 shows that the results for support vector regression are worse than the linear regression model using our proposed features combined with UMBC, hence we focus hereon on linear regression. Table 5.2 provides the correlations, MAE, and RMS values for models produced using various sets of features. We considered two baselines, simple Ngram overlap and the off-the-shelf UMBC STS metric (Han et al., 2013). In general, we found that Ngram overlap (Row 1) performed best alone of our features, but falls short of the UMBC baseline (Row 2). It is interesting that Ngram alone out-performs distributional measures (which Conrad et al. (2012) found most helpful for argument tagging task) as well as Rouge (which contains metrics insensitive to linear adjacency).

Table 5.2, Row 15, shows that the best correlation that is achievable without UMBC is the combination of Ngram, LIWC, ROUGE, and DISCO (NLRD). This combination significantly improves over the UMBC baseline of 0.46 to 0.50 (paired t-test, $p < .05$).

We then tested combinations of features to determine which feature sets are complementary. LIWC + NGRAM is significantly different than NGRAM alone ($p < 0.01$), and ROUGE + NGRAM is significantly different than NGRAM alone ($p = 0.03$), but DISCO does not add anything ($p = 0.2$). This shows that LIWC and ROUGE features complement Ngram

Row	Feature Set	R	MAE	RMS
1	**NGRAM (N)**	0.39	0.90	1.09
2	**UMBC (U)**	0.46	0.86	1.06
3	**LIWC (L)**	0.32	0.92	1.13
4	**DISCO (D)**	0.33	0.93	1.12
5	**ROUGE (R)**	0.34	0.91	1.12
6	**N-U**	0.47	0.85	1.05
7	**N-L**	0.45	0.86	1.06
8	**N-R**	0.42	0.88	1.08
9	**N-D**	0.41	0.89	1.08
10	**U-R**	0.48	0.84	1.04
11	**U-L**	0.51	0.83	1.02
12	**U-D**	0.45	0.86	1.06
13	**N-L-R**	0.48	0.84	1.04
14	**U-L-R**	0.53	0.81	1.00
15	**N-L-R-D**	0.50	0.83	1.03
16	**N-L-R-U**	0.54	0.80	1.00
17	**N-L-R-D-U**	0.54	0.80	1.00

Table 5.2: Results for Different Individual Features and Feature Combinations.

features. Other combinations of interest are NGRAM + LIWC (Row 7) which amazingly performs as well as UMBC while UMBC includes sentence alignment, a model of negation, and distributional measures (Han et al., 2013). This suggests that AFS is a different task than STS.

Additionally, we also combined our proposed set of features with UMBC. A comparison of Row 15 (our feature set) with Rows 16 and 17 of Table 5.2 where we combine our features with UMBC shows that this improves the correlation further, from the UMBC baseline of 0.46 to 0.54 ($p < 0.01$).

It is also interesting to examine the differences in model scores for particular argument pairs as shown in Table 5.3. The best performing model for each row is in **bold** in Table 5.3. The table is sorted by the AFS score (gold standard). The argument pairs shown in **bold** are cases where UMBC by itself beats our proposed model. As described in the HIT instructions in Figure 5.7, values of AFS near 0 (Row 1) indicate different topics and no similarity. Values near 1 indicate same topic but different arguments (Rows 2,3). Values of 3 and above indicate same **facet** (Rows 7,8), and values near 5 are same **facet** and very similar argument (Rows 12 and 13). Both Arg1 and Arg2 in Row 10 make the same argument, but Arg1 includes additional argumentation. In Row 12, there is very low n-gram overlap, but strong AFS and NLRD performs better than the other models, and LIWC performs well by itself.

In Row 1, UMBC performs the best with a predicted score of 0.37 as opposed to an AFS score of 0.00. The top performance of NLRD in Row 5 without UMBC perhaps arises from the semantic information that extermination and holocaust are somehow related. NGRAM overlap does the best in Row 13 despite the fact that the phrase *No one argues the point that* does not participate in the NGRAM overlap.

Row	N	L	U	NLRD	NLRDU	MT	Arg1	Arg2
1	1.38	1.50	**0.37**	1.31	0.40	0.00	**Everyone has the freedom of speech.**	**Service in the military.**
2	2.00	2.02	**1.55**	2.33	1.86	1.14	**Gay people should be able to marry a person of their choice and get equal rights.**	**Referring to namecalling and violence from the original post that was opposing gay rights.**
3	2.00	1.29	2.52	**1.37**	1.54	1.33	Constitutional right to be opposed to gay marriage as well as gay people themselves.	Arguing about marriage benefits between single people and married.
4	2.00	1.70	2.74	**1.77**	1.98	1.80	People should not pick and choose what they want equal rights on.	People did not want gay marriage.
5	1.38	1.92	0.88	**1.94**	1.64	2.50	The Republicans creating another Holocaust.	No republican in leadership would call for the extermination of gays.
6	1.69	2.02	**2.58**	1.89	2.49	2.60	**Homosexuals have all the same rights as heterosexuals.**	**Opposition to equal rights for gay couples.**
7	1.83	2.40	1.46	**2.81**	2.51	3.00	There was prejudice against gays in 1909 just as there is now.	It is prejudice as opposed to religious or moral beliefs which fuel the anti-gay agenda;
8	2.00	1.70	**3.16**	1.73	2.41	3.40	**Homosexual relationships should not compare to heterosexual marriages because only heterosexuals are legally allowed to marry.**	**Marriage should be between a heterosexual couple.**
9	2.00	2.70	2.09	2.83	**3.03**	3.50	It is prejudice as opposed to religious or moral beliefs which fuel the anti-gay agenda;	When people claim religion in doing prejudice they are actually abandoning their morals.
10	**2.94**	2.02	2.93	2.18	2.70	3.50	**Gay people should be able to marry a person of their choice and get equal rights.**	**Gay couples are unable to get any benefits that married people do.**
11	2.14	1.50	**2.91**	2.08	2.62	3.60	**Paul Cameron is the voice of the Republicans.**	**Conversation about Paul Cameron.**
12	2.63	3.63	2.60	**3.75**	3.57	4.17	In opening this opportunity for gay marriage, the definition of marriage will change.	Opponents of homosexual marriage tend to argue that a change to marriage law would make it too open ended.
13	**4.23**	2.72	2.26	4.82	4.12	4.50	AIDs was initially spread in the United States primarily by homosexuals.	No one argues the point that AIDs was spread in the United States by homosexuals.

Table 5.3: Predicted Scores for each model and the Mechanical Turk AFS gold standard.

KEY: Feature sets model. N = NGRAM, U = UMBC STS tool, L = LIWC; R = Rouge, D = DISCO, AFS= Mean of Mechanical Turker AFS scores, our gold standard.

5.3 AFS from Actual Social Media Arguments

A potential limitation to the central proposition based AFS approach was that it was tested on human authored abstractive labels derived from the pyramid method. A better generalization would be to test it on an independent argument dataset. To have the same granularity of argument as the previous dataset, we focus on **Sentential Arguments**, single sentences that clearly express a particular argument facet in dialogue. The eventual goal is to use **Sentential Arguments** to produce extractive summaries of online dialogues about current social and political topics.

5.3.1 Argument Quality Data

Previous work has shown that probability of finding a high-quality argument in randomly selected sentences from social media debates is low. A good quality argument is a prerequisite for the AFS task.

Swanson et al. (2015) created a corpus of argumentative dialogues on the topics Gay Marriage (GM, 22425 posts), Gun Control (GC, 38102 posts), Death Penalty (DP, 5283 posts) and Evolution (EV, 43624), by combining the Internet Argument Corpus (IAC) (Walker et al., 2012c), with dialogues from `http://www.createdebate.com/`, and trained a regressor model to predict the quality of extracted arguments. The Argument Quality (**AQ**) regressor gives a score to each sentence, which is intended to reflect how easily the speaker's argument can be understood from the sentence without any context. Easily understandable sentences are assumed to be prime candidates for producing extractive summaries. In Swanson et al. (2015),

the annotators rated AQ using a continuous slider ranging from hard (0.0) to easy to interpret (1.0).

Topic	Original	Rescored	Sampled	AQ #N (%)
Gun Control	89,722	63,025	2140	1887 (88%)
Death Penalty	17,904	11,435	1986	1520 (77%)
Gay Marriage	51,543	40,306	2062	1745 (85%)

Table 5.4: Sentence count in each domain. Sampled bin range > 0.55 and number of sentential arguments (high AQ) after annotation.

We expected to apply Swanson's AQ regressor to our sample completely "out of the box". However, we first discovered that many sentences given high AQ scores were very similar, while we need a sample that realizes many **diverse** facets. We then discovered that some extracted sentential arguments were not actually high quality. We hypothesized that the diversity issue arose primarily because Swanson's dataset was filtered using high PMI n-grams. We also hypothesized that the quality issue had not surfaced because Swanson's sample was primarily selected from sentences marked with the discourse connectives *but*, *first*, *if*, and *so*. Our sample (Original column of Table 5.4) is much larger and was not similarly filtered. We refined the Mechanical Turk task to elicit new training data for AQ as described below.

Figure 5.8 plots the distribution of word counts for sentences from our sample that were given an AQ score > 0.91 by Swanson's trained AQ regressor. The first bin shows that many sentences with less than 10 words are predicted to be high quality, but many of these sentences in our data consisted of only a few elongated words (e.g. HAHAHAHA...). The

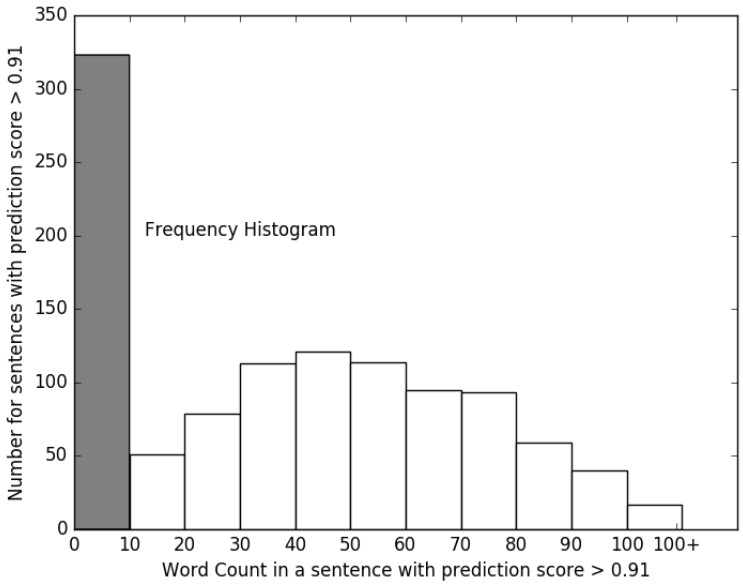


Figure 5.8: Word count distribution for argument quality prediction scores > 0.91 for Swanson's original model.

upper part of the distribution shows a large number of sentences with more than 70 words with a predicted AQ > 0.91. We discovered that most of these long sentences are multiple sentences without punctuation. We thus refined the AQ model by removing duplicate sentences, and rescoring sentences without a verb and with less than 4 dictionary words to AQ $= 0$. We then restricted our sampling to sentences between 10 and 40 tokens, to eliminate run-on sentences and sentences without much propositional content. We did not retrain the regressor, rather we resampled and rescored the corpus. See the Rescored column of Table 5.4. After removing the two tails in Figure 5.8, the distribution of word counts is almost uniform across bins of sentences from length 10 to 40.

As noted above, the sample in Swanson et al. (2015) was filtered using PMI, and PMI

contributes to AQ. Thus, to end up with a diverse set of sentences representing many facets of each topic, we decided to sample sentences with lower AQ scores than Swanson had used. We binned the sentences based on predicted AQ score and extracted random samples across bins ranging from 0.55–1.0, in increments of 0.10. Then we extracted a smaller sample and collected new AQ annotations for Gay Marriage and Death Penalty on Mechanical Turk, using the definitions in Figure 5.9.

Score	**Scoring Criteria**
3	The phrase is clearly interpretable AND either expresses an argument, or a premise or a conclusion that can be used in an argument about a facet or a sub-issue for the topic of gay marriage.
2	The phrase is clearly interpretable BUT does not seem to be a part of an argument about a facet or a sub-issue for the topic of gay marriage.
1	The phrase cannot be interpreted as an argument.

Figure 5.9: Argument Quality HIT as instantiated for the topic Gay Marriage.

We pre-selected three annotators using a qualifier that included detailed instructions and sample annotations. A score of 3 was mapped to a *yes* and scores of 1 or 2 mapped to a *no*. We simplified the task slightly in the HIT for Gun Control, where five annotators were instructed to select a yes label if the sentence clearly expressed an argument (score 3), or a no label otherwise (score 1 or 2).

We then calculated the probability that the sentences in each bin were high quality arguments using the resulting AQ gold standard labels, and found that a threshold of predicted

AQ > 0.55 maintained both diversity and quality. Table 5.4 summarizes the results of each stage of the process of producing the new AQ corpus of 6188 sentences (sampled and then annotated). The last column of Table 5.4 shows that gold standard labels agree with the rescored AQ regressor between 77% and 88% of the time.

5.3.2 AFS Corpus with Sentential Arguments

S7: Since there are gun deaths in countries that have banned guns, the gun bans did not work.
S8: It is legal to own weapons in this country, they are just tightly controlled, and as a result we have far less gun crime (particularly where it's not related to organised crime).
S9: My point was that the theory that more gun control leaves people defenseless does not explain the lower murder rates in other developed nations.

Figure 5.10: Paraphrases of the *Gun ownership does not lead to higher crime* facet of the Gun Control topic across different conversations.

Our approach draws strongly on recent work on semantic textual similarity (STS) (Agirre et al., 2013). STS measures the degree of semantic similarity between a pair of sentences with values that range from 0 to 5. However, we distinguish AFS from STS because: (1) our data is so different: STS data consists of descriptive sentences whereas our sentences are argumentative excerpts from dialogues; and (2) our definition of facet allows for sentences that express opposite stance to be realizations of the same facet (AFS $= 3$) in Figure 5.7. Related work had primarily used entailment or semantic equivalence to define argument similarity (Habernal and Gurevych, 2015; Boltuzic and Šnajder, 2015). We believe the definition of AFS

given in Figure 5.7 will be more useful in the long run than semantic equivalence or entailment, because two arguments can only be contradictory if they are about the same facet. For example, consider that sentential argument **S7** in Figure 5.10 is anti gun-control, while sentences **S8** and **S9** are pro gun-control. Our annotation guidelines label them with the same facet, in a similar way to how the curated summaries on ProCon provides both a Pro and Con side for each facet.

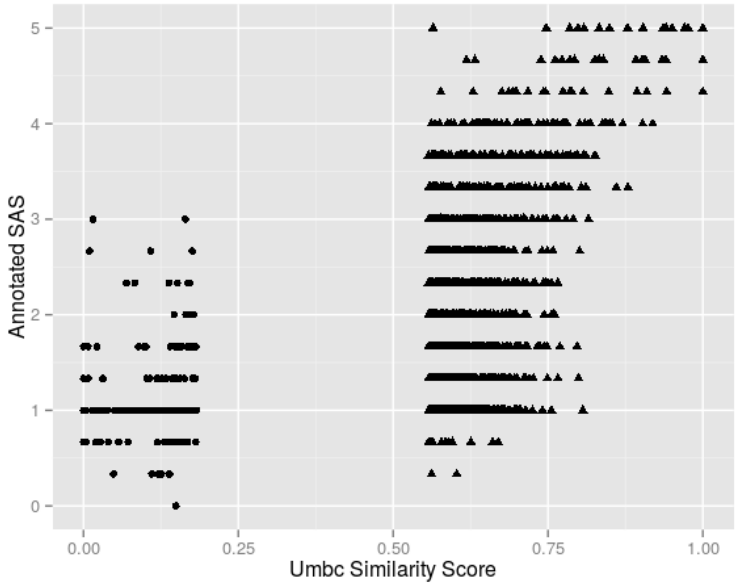


Figure 5.11: The distribution of AFS scores as a function of UMBC STS scores for Gun Control sentences.

In order to efficiently collect annotations for AFS, we want to produce training data pairs that are more likely than chance to be the same facet (scores 3 and above as defined in Figure 5.7). Similar arguments are rare with an all-pairs matching protocol, e.g. in ComArg, approximately 67% of the annotations are "not a match" (Boltuzic and Šnajder, 2014). Also, we found that Turkers are confused when asked to annotate similarity and then given a set of

sentence pairs that are almost all highly dissimilar. Annotations also cost money. We, therefore, used UMBC STS (Han et al., 2013) to score all potential pairs, an off-the-shelf STS tool from University of Maryland Baltimore County [3]. To foreshadow, the plot in Figure 5.11 shows that this pre-scoring works: (1) the lower quadrant of the plot shows that STS $< .20$ corresponds to the lower range of scores for AFS; and (2) the lower half of the left hand side shows that we still get many arguments that are low AFS (values below 3) in our training data.

We selected 2000 pairs in each topic, based on their UMBC similarity scores, which resulted in lowest UMBC scores of 0.58 for GM, 0.56 for GC, and 0.58 for DP. To ensure a pool of diverse arguments, a particular sentence can appear in at most ten pairs. MT workers took a qualification test with definitions and instructions as shown in Figure 5.7. Sentential arguments with sample AFS annotations were part of the qualifier. The 6000 pairs were made available to our three most reliable pre-qualified workers. The last row of Table 5.6 reports the human topline for the task, i.e. the average pairwise r across all three workers. Interestingly, the Gay Marriage topic ($r = 0.60$) is more difficult for human annotators than either Death Penalty ($r = 0.74$) or Gun Control ($r = 0.69$).

5.3.3 Experiments

Given the data collected above, we defined a supervised machine learning experiment with AFS as our dependent variable. We developed a number of baselines using off the shelf tools. Features are grouped into sets and discussed in detail below.

[3] http://swoogle.umbc.edu/SimService/

5.3.3.1 Features

NGRAM cosine. Our primary baseline is an n-gram overlap feature. For each argument, we extracted the unigrams, bigrams, and trigrams, and then calculated the cosine similarity between two texts represented as vectors of their n-gram counts.

Rouge. Rouge is a family of metrics for comparing the similarity of two summaries (Lin, 2004), which measures overlapping units such as continuous and skip n-grams, common subsequences, and word pairs. We used all the rouge f-scores from the pyrouge package. Our analysis showed that rouge_s*_f_score correlates most highly with AFS.[4]

UMBC STS. We consider STS, a measure of the semantic similarity of two texts (Agirre et al., 2012), as another baseline, using the UMBC STS tool. Figure 5.11 illustrates that in general, STS is a rough approximation of AFS. It is possible that our selection of data for pairs for annotation using UMBC STS either improves or reduces its performance.

Google Word2Vec. Word embeddings from Word2Vec (Mikolov et al., 2013b) are popular for expressing semantic relationships between words, but using word embeddings to express entire sentences often requires some compromises. In particular, averaging Word2Vec embeddings for each word may lose too much information in long sentences. Previous work on argument mining has developed methods using Word2Vec that are effective for clustering similar arguments (Habernal and Gurevych, 2015; Boltuzic and Šnajder, 2015). Other research creates embeddings at the sentence level using more advanced techniques such as Paragraph Vectors (Le and Mikolov, 2014).

We took a more direct approach in which we used the word embeddings directly as

[4] https://pypi.python.org/pypi/pyrouge/

features. For each sentential argument in the pair, we created a 300-dimensional vector by filtering for stopwords and punctuation and then averaging the word embeddings from Google's Word2Vec model for the remaining words.[5] Each dimension of the 600 dimensional concatenated averaged vector was used directly as a feature. In our experiments, this concatenation method significantly outperformed cosine similarity (Table 5.5, Table 5.6).

Custom Word2Vec. We also created our own 300-dimensional embeddings for our dialogic domain using the Gensim library (Řehůřek and Sojka, 2010), with default settings, and a very large corpus of user-generated dialogic content. This included the corpus described in Section 5.3.2 (929, 206 forum posts), an internal corpus of 1, 688, 639 tweets on various topics, and a corpus of 53, 851, 542 posts from Reddit.[6]

LIWC category and Dependency Overlap. Both dependency structures and the Linguistics Inquiry Word Count (LIWC) tool have been useful in previous work (Pennebaker et al., 2001; Somasundaran and Wiebe, 2009; Hasan and Ng, 2013a). We developed a novel feature set that combines LIWC category and dependency overlap, aiming to capture a generalized notion of concept overlap between two arguments, i.e., to capture the hypothesis that classes of content words such as affective processes or emotion types are indicative of a shared facet across pairs of arguments.

We created partially generalized LIWC dependency features and counted overlap normalized by sentence length across pairs, building on previous work (Joshi and Penstein-Rosé, 2009). Stanford dependency features (Manning et al., 2014) are generalized by leaving one

[5] https://code.google.com/archive/p/word2vec/

[6] One month sample https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment

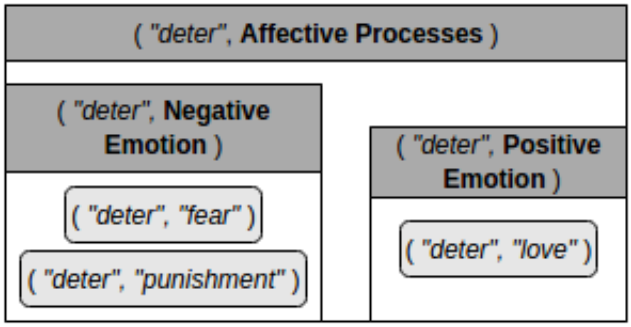


Figure 5.12: LIWC Generalized Dep. tuples

dependency element lexicalized, replacing the other word in the dependency relation with its LIWC category and by removing the actual dependency type (nsubj, dobj, etc.) from the triple. This created a tuple of (*"governor token", "LIWC category of the dependent token"*). We call these simplified LIWC dependencies.

Figure 5.12 illustrates the generalization process for three LIWC simplified dependencies, *("deter", "fear")*, *("deter", "punishment")*, and *("deter", "love")*. Because LIWC is a hierarchical lexicon, two dependencies may share many generalizations or only a few. Here, the tuples with dependent tokens, *fear* and *punishment* are more closely related because their shared generalization include both *Negative Emotion* and *Affective Processes*, but the tuples with dependent tokens *fear* and *love* have a less similar relationship because they only share the *Affective Processes* generalization.

5.3.3.2 Results

We randomly selected 90% of our annotated pairs to use for nested 10-fold cross-validation, setting aside 10% for qualitative analysis of predicted vs. gold-standard scores. We

		Gun Control				Gay Marriage				Death Penalty			
		RR		SVR		RR		SVR		RR		SVR	
ID	**Features**	r	**RMSE**	r	**RMSE**	r	**RMSE**	r	**RMSE**	r	**RMSE**	r	**RMSE**
1	UMBC	0.49	0.90	0.50	0.90	0.16	0.90	0.21	0.90	0.21	1.16	0.20	1.20
2	Ngram	0.46	0.91	0.46	0.92	0.24	0.88	0.24	0.91	0.23	1.16	0.24	1.18
3	Rouge	0.52	0.88	0.57	0.86	0.22	0.89	0.26	0.90	0.39	1.10	**0.40**	**1.11**
4	LIWC dependencies	0.50	0.89	**0.59**	**0.85**	**0.27**	**0.88**	0.26	0.90	0.34	1.12	0.40	1.12
5	Custom W2Vec Cosine	0.47	0.91	0.52	0.89	0.22	0.89	0.25	0.90	0.29	1.14	0.30	1.16
6	Google W2Vec Cosine	0.40	0.94	0.47	0.93	0.16	0.90	0.20	0.92	0.29	1.14	0.30	1.16

Table 5.5: Results for predicting AFS with individual features using Ridge Regression (RR) and Support Vector Regression (SVR) with 10-fold Cross-Validation on the 1800 training items for each topic.

used Ridge Regression (RR) with l2-norm regularization and Support Vector Regression (SVR) with an RBF kernel from scikit-learn (Pedregosa et al., 2011). Performance evaluation uses two standard measures, Correlation Coefficient (r) and Root Mean Squared Error (RMSE). A separate inner cross-validation within each fold of the outer cross-validation is used to perform a grid search to determine the hyperparameters for that outer fold. The outer cross-validation reports the scoring metrics.

Simple Ablation Models. We first evaluate simple models based on a single feature using both RR and SVR. Table 5.5, Rows 1, 2, and 3 show the baseline results: UMBC Semantic Textual Similarity (STS), Ngram Cosine, and Rouge. Surprisingly, the UMBC STS measure does not

perform as well as Ngram Cosine for Death Penalty and Gay Marriage. LIWC dependencies (Row 4) perform similarly to Rouge (Row 3) across topics. Cosine similarity for the custom Word2Vec model (Row 5) performs about as well or better than n-grams across topics, but cosine similarity using the Google model (Row 6) performs worse than n-grams for all topics except Death Penalty. Interestingly our custom Word2Vec models perform significantly better than the Google Word2Vec models for Gun Control and Gay Marriage, with both much higher r and lower RMSE, while performing identically for Death Penalty.

Feature Combination Models. Table 5.6 shows the results of testing feature combinations to learn which ones are complementary. Since SVR consistently performs better than RR, we used SVR only in combinations. Significance was calculated using paired t-tests between the RMSE values across folds. We paired Ngrams separately with LIWC and ROUGE to evaluate if the combination is significant. Ngram+Rouge (Row 1) is significantly better than Ngram for Gun Control and Death Penalty ($p < .01$), and Gay Marriage ($p = .03$). Ngram+LIWC (Row 2) is significantly better than Ngram for Gun Control, and Death Penalty ($p < .01$). Thus both Rouge and LIWC provide complementary information to Ngrams.

Our best result using our hand-engineered features was a combination of LIWC, Rouge, and Ngrams (Row 3). Interestingly, adding UMBC STS (Row 4) gave a small but significant improvement ($p < 0.01$ for Gun Control; $p = 0.07$ for Gay Marriage). Thus we took Ngrams, LIWC, Rouge, and UMBC STS (Row 4) as our best hand-engineered model across all topics with a correlation of 0.65 for Gun Control, 0.50 for Death Penalty and 0.40 for Gay Marriage. This combination was significantly better than the baselines for Ngram baseline ($p < .01$), UMBC STS ($p <= .02$) and Rouge ($p < .01$) for all three topics.

ID	Feature Combinations with SVR	Gun Control		Gay Marriage		Death Penalty	
		r	**RMSE**	r	**RMSE**	r	**RMSE**
1	Ngram- Rouge	0.59	0.85	0.29	0.89	0.40	1.11
2	Ngram- LIWC dependencies	0.61	0.83	0.34	0.88	0.43	1.10
3	Ngram- LIWC dependencies- Rouge	0.64	0.80	0.38	0.86	0.49	1.05
4	Ngram- LIWC dependencies- Rouge- UMBC	0.65	0.79	0.40	0.86	0.50	1.05
5	CustomW2Vec Concatenated vectors	0.71	0.72	0.48	0.80	0.56	0.99
6	GoogleW2Vec Concatenated vectors	0.71	0.72	0.50	0.79	0.57	0.98
7	Ngram- LIWC dependencies- Rouge- UMBC- CustomW2Vec Concatenated vectors	0.73	0.70	0.54	0.77	0.62	0.93
8	Ngram- LIWC dependencies- Rouge- UMBC- GoogleW2Vec Concatenated vectors	**0.73**	**0.70**	**0.54**	**0.77**	**0.63**	**0.92**
9	**HUMAN TOPLINE**	0.69		0.60		0.74	

Table 5.6: Results for feature combinations for predicting AFS, using Support Vector Regression (SVR) with 10-fold Cross-Validation on the 1800 training items for each topic.

We then further combined the hand-engineered features (Row 4) with the Google Word2Vec features (Row 6), creating the model in Row 8. A paired t-test between RMSE values from each cross-validation fold for each model (Row 4 vs. Row 8 and Row 6 vs. Row 8) showed that the our hand-engineered features are complementary to Word2Vec, and their

combination yields a model significantly better than either model alone ($p < .01$).

Although it is common to translate word embeddings into single features or reduced feature sets for similarity through the use of clustering (Habernal and Gurevych, 2015) or cosine similarity (Boltuzic and Šnajder, 2015), we showed that it is possible to improve results by directly combining word embeddings with hand-engineered features. In our task, sentences were limited to a maximum of 40 tokens in order to encourage single-facet sentences, but this may have provided an additional benefit by allowing us to concatenate word embeddings while still preserving useful signal. The custom Word2Vec concatenation features perform worse for Death Penalty and Gay Marriage as compared to Google Word2Vec model. Gun Control concatenation vectors perform similarly for both the models (Row 5 and Row 6 in Table 5.6). This may be due to the topic distribution in IAC. It has a lot more data for Gun Control (89,722 sentences). The low performance of Gay Marriage and Death Penalty for concatenation vectors may be due to relatively few sentences for them in the IAC [(51,543(GM); 17,904(DP)].

Our results also demonstrated that using concatenation for learning similarity with vector representations works much better than the common practice of reducing a pair of vectors to a single score using cosine similarity. Previous work (Li et al., 2015; Pennington et al., 2014) also showed that all dimensions are not equally useful predictors for a specific task. For sentiment classification, Li et al. (2015) found that "too large a dimensionality leads many dimensions to be non-functional, causing two sentences of opposite sentiment to differ only in a few dimensions". This may also be the situation for the 300-dimensional embeddings used for AFS. Hence, when using concatenation, single dimensions can be weighted to adjust for non-functional dimensions, but using cosine makes this per-dimension weighting impossible. This

ID	Argument 1	Argument 2	STS	Ngram	Rouge	LIWC dep	W2Vec	AFS	MT AFS
GC1	You say that gun control must not be effective because the study's conclusions about gun control were inconclusive.	You're right that gun control isn't about guns, however, but 'control' is a secondary matter, a means to an end.	1.82	2.56	2.22	1.53	**1.40**	1.5	**1**
DP2	I don't feel as strongly about the death penalty as I feel about the abortion rights debate since I can relate to the desire for vengeance that people feel.	Well I, as creator of this debate, think that there should not be a death penalty.	1.82	2.38	2.07	**1.29**	1.44	1.24	**1.33**
GC3	They do not have the expressed, enumerated power to pass any law regarding guns in the constitution.	Which passed the law requireing "smart guns", if they ever become available (right now they do not exist).	1.74	1.83	2.67	1.50	1.82	**1.88**	**2.0**
GM4	Technically though marriage is not discrimination, because gays are still allowed to marry the opposite sex.	Everyone has the right to marry someone of the opposite sex, and with gay marriage, everyone will have the right to marry someone of the same AND opposite sex.	1.76	2.09	1.68	2.00	**2.23**	2.06	**2.33**
GM5	If the state wants to offer legal protections and benefits to straight married couples, it cannot constitutionally refuse equal protections to gay ones.	Same-sex couples are denied over 1,000 benefits, rights, and protections that federal law affords to married, heterosexual couples, as well as hundreds of such protections at the state level.	1.77	1.91	1.77	2.66	**3.56**	3.72	**3.33**
DP6	In addition, it is evident that the death penalty does not deter murder rates.	BUT it is not apparent that death penalty lower crime rate.	2.03	2.31	3.71	2.21	3.84	**3.95**	**4.0**
DP7	Living in jail for life costs less money then the death penalty.	Morality aside, no evidence of deterrence aside, the death penalty costs more than life imprisonment.	1.84	2.43	2.56	**3.23**	2.90	2.90	**4.33**

Table 5.7: Illustrative Argument pairs, along with the predicted scores from individual feature sets, predicted(**AFS**) and the Mechanical Turk human topline (**MT AFS**). The best performing feature set is shown in bold. GC=Gun Control, DP=Death Penalty, GM=Gay Marriage.

might explain why our custom Word2Vec model outperformed the Google model when using cosine as compared to concatenation, i.e. more dimensions are informative in the custom model but overall, the Google model provides more complementary information when non-functional dimensions are accounted for. More analysis is needed to fully support this claim.

To qualitatively illustrate some of the differences between our final AFS regressor model (Row 8 of Table 5.6) and several baselines, we applied the model to a set-aside 200 pairs per topic. Table 5.7 shows examples selected to highlight the strengths of AFS prediction for

different models as compared to the AFS gold standard scores.

MT AFS values near 1 indicate same topic but no similarity. Rows GC1 and DP2 talk about totally different facets and only share the same topic (AFS = 1). Rouge and Ngram features based on word overlap predict scores that are too high. In contrast, LIWC dependencies and Word2Vec based on concept and semantic overlap are more accurate. MT values near 3 indicate same facet but somewhat different arguments. Arguments in row GM4 talk about marriage rights to all, and there is some overlap in these arguments beyond merely being the same topic, however, the speakers are on opposite stance sides. Both of the arguments in row GM5 (MT AFS of 3.3) reference the same facet of the financial and legal benefits available to married couples, but Arg2 is more specific. Both Word2Vec and our trained AFS model can recognize the similarity in the concepts in the two arguments and make good predictions.

MT values above 4 indicate two arguments that are the same facet and very similar. Row DP6 gets a high Rouge overlap score and Word2Vec relates 'lower crime rate' as semantically similar to 'deter murder rates' thus yielding an accurately high AFS score. DP7 is an example where LIWC dependencies perform better as compared to other features, because it focuses in on the dependency between the death penalty and cost, but none of the models do well at predicting the MT AFS score. One issue here may be that, despite our attempts to sample pairs with more representatives of high AFS, there is just less training data available for this part of the distribution. Hence all the regressors will be conservative at predicting the highest values.

5.4 Chapter Summary

This chapter introduced a new task of **Argument Facet Similarity** (**AFS**) aimed at identifying facets across opinionated dialogues. We employed a novel application of facet identification, using two different strategies.

The first method employed a pyramid summarization scheme to derive facets from **central propositions** that rise to the top of the pyramid across summarizers, and then (via AFS) across many dialogues on a topic. We used clustering in combination with measures of semantic similarity to group the central propositions into the important **facets** of an argument across many different dialogues. Importantly, we do not attempt to enumerate the possible **facets** for an argument in advance, believing that bottom-up discovery of **facets** is a better fit to the problem. Our method instead explicitly models graded similarity of arguments. Because of the novelty of the task, an initial step was to investigate if human annotators could reliably label the argument pairs based on AFS definition for a single topic of Gay Marriage. Mechanical Turk workers achieved an average pairwise correlation of 0.7 with gold standard indicating that the task is feasible for humans to annotate manually. The result of this annotation study was a novel annotated corpus of argument pairs annotated with AFS scores. We then addressed the automatic assessment of AFS, where we experimented with different regression models and multiple feature groups. The effectiveness of various features is investigated by experimenting with different feature combinations, feature ablation tests, and significance tests. Our results showed that we could identify AFS with a correlation of 0.54 using features based on LIWC, Rouge, Ngram, and UMBC as opposed to a baseline of 0.46 provided by the STS system designed for

a similar task.

A limitation of the Pyramid tool was that it required the annotator to provide a human readable label for a collection of contributors that realize the same propositional content. To overcome this limitation, the next step was to develop and test an AFS regressor on automatically extracted raw sentences from social media dialogues. We analyzed and improved the argument extractor from Swanson et al. (2015), by testing it on a much larger dataset, and developed a larger gold standard corpus for Argument Quality (AQ). We created a new corpus of sentential arguments with gold-standard labels for AFS. We developed a regressor that can predict AFS on extracted sentential arguments using a combination of hand-engineered and unsupervised features with a correlation averaging 0.63 compared to a human top line averaging 0.68, for three debate topics namely Gay Marriage, Death Penalty, and Gun Control. Finally, the evaluation results from both pyramid labels and sentential argument experiments clearly demonstrated that proposed features yield better performance on AFS than those used for STS, thus validating the hypothesis $\mathbf{H_3}$. In summary, we have presented two novel techniques for modeling argument facet similarity in social media domain along with a reliable corpus, evaluated our ability to identify facets against several reasonable baselines, and improved on STS systems for the AFS task.

Chapter 6

Conclusion and Future Work

Online debate platforms and social networks provide a two way communication platform with a massive amount of opinion and argument rich information. Such enormous feedback from society offers a unique opportunity to the government and policy makers to get useful insights and public sentiment on issues under discussion. However, this information is distributed and unstructured. To identify and integrate arguments across these discussions requires computational methods that can facilitate users to systematically search, analyze and summarize arguments as well as reason about the relationship among arguments, on a given topic of interest.

In this thesis, we presented a systematic approach to leverage the pyramid based summarization framework to rank and select arguments in social media dialogue, which is a novel method for ranking arguments in conversational data. We developed techniques to recognize abstract objects and issues under discussion that are central to a speakers stance by introducing a new task of Argument Facet Similarity (AFS). A bottom-up approach was used to induce

facets by identifying similar repeated arguments across many discussions of a topic. A graded argument similarity model was defined that takes as input two sentential arguments and returns a scalar value that predicts their similarity (AFS). Section 6.1 summarizes the major contributions and findings of the thesis. Section 6.2 discusses the limitations. Finally, we conclude with a discussion of possible avenues for future work that stem from this research in Section 6.3.

6.1 Contributions and Conclusion

In this section, we revisit and discuss how far the hypothesis presented in the introduction have been validated, and the underlying research contributions.

6.1.1 Central Propositions and Important Arguments of a Social Media Dialogue

The first contribution of the thesis is to identify the debate propositions most central to a particular stance, which are used to summarize the abstract issues in an ideological online debate dialogue.

Obtaining reliable data for central propositions was a first step in this direction. In Chapter 3, we described a systematic collection of naturally occurring dialogue corpus from political debate on internet forums, with 61 dialogues on the topic of Gay Marriage, 50 on Abortion, and 50 on Gun Control. Annotating text spans in the dialogue that correspond to central propositions is a cognitive process and can be biased. We increased the objectivity of the task by using saliency as a measure for determining central propositions. We elicit salience using summarization by many human summarizers. The common elements across the summaries

extracted from the pyramid structure, also known as high tier pyramid labels, reflected the important content in the dialogue. These pyramid labels then identified the issues under discussion in debate dialogues, giving the central propositions as demonstrated in Section 3.4. This work is amongst the first to apply the pyramid evaluation scheme to argumentative dialogue. Our first hypothesis is:

H_1: *It is possible to map the ranked labels from the pyramid annotation back to the dialogue with good reliability.*.

In Section 3.5, we showed that our annotation scheme could be successfully applied to link pyramid labels to dialogue sentences with substantial agreement. In particular, we obtained a Cohen's Kappa of 0.68 for Gun Control, 0.63 for Abortion, and 0.62 for Gay Marriage. The step by step approach using pyramid labels linked to dialogue sentences identified important arguments in a dialogue in a verifiable and precise manner rather than based on purely subjective judgments of the annotators. As a result of this study, we created an Argument Dialogue Corpus, and contributed to the current state of knowledge in the following ways: (1) we collected summaries of spontaneously-produced written dialogue of high social and political importance (2) summaries reliably annotated with pyramid structure, and dialogues with well-defined important arguments. This annotated corpora fostered the development of computational methods to validate the next hypothesis in this thesis.

H_2: *Adding linguistic and contextual features improves argument extraction.*

Our next contribution is the development of an automatic summarization framework for extracting important arguments in social media dialogue, as described in Chapter 4. Statistical corpus analysis in Section 4.1 showed that a considerable number of sentences were marked

as not important and therefore non-argumentative. As a result, we modeled argument extraction as a binary classification task to select important sentences from social media dialogue. Subsequently, we used an SVM and a Bidirectional LSTM to train and predict. Section 4.2 described the core feature extraction module based on linguistic and contextual features. Moreover, our analysis gave an in-depth insight into the LIWC categories that provided the topical information useful for identifying sentences with substantive issues. Our system obtained a higher F-score (**0.74** for Gun Control, **0.71** for Gay Marriage, and **0.67** for Abortion) significantly beating the traditional summarization baselines, for all the three debate topics. Experimental results showed that our model makes use of both topical features (LIWC) and general features (Readability). High performance of our model was also due to our new contextual features. Feature combination and ablation tests in results Table 4.3 demonstrated that leveraging contextual information from Stanford Coreference module, and information from previous sentences significantly improved the performance of the classifier.

6.1.2 Argument Facet Similarity

In Chapter 5, we introduced the notion of an Argument Facet, defined a novel task of Argument Facet Similarity, and designed the experiments to support our last hypothesis:

H_3: *The proposed features yield better performance on AFS than those used for STS.*

Issues and concerns which manifest in a large number of arguments represent important topics of conversation and debate. Social media argumentation is a dynamic activity, and many of the interesting and topical arguments are developed in dialogical situations. People use arguments to support their beliefs and claims, as well as attack the opponents. In Chapter 5, we

defined an argument **Facet** as a low level issue that often reoccurs in many arguments in support of the author's stance or in attacking the other author's position. Another major contribution of the thesis is the development of Argument Facet Similarity framework as discussed in Sections 5.2 and 5.3 , where we developed a new corpus and computational models aimed at identifying various facets across opinionated dialogs. These facets can give an insight into an average person's viewpoint on the current issues.

Other work in this area categorizes sentences or posts using topic-specific argument labels, which are functionally similar to our facets as discussed above (Conrad et al., 2012; Hasan and Ng, 2014; Boltuzic and Šnajder, 2014; Naderi and Hirst, 2016). Boltuzic and Šnajder (2015) used a predefined list to label posts. They applied unsupervised clustering using a semantic textual similarity tool, but evaluate clusters using their hand-labeled argument tags. This scheme is very limited and can identify explicit aspects as present in this predefined list. However, the issues in a debate are not fixed, and argumentation is an ongoing process where the new aspects emerge all the time. In contrast, our approach as described in 5.2 and 5.3 induced facets in a bottom-up manner by explicitly modeling graded similarity of sentential arguments, and does not map arguments to any predefined list. Moreover, these earlier approaches assume the labels are dependent on a particular stance towards an issue, whereas our facets are deliberately designed to unify across stance disagreement to capture similarity in facet and argument simultaneously. Chapter 5 illustrated the design of two different experiments for AFS. In Section 5.2 we developed an AFS regressor for predicting the similarity of human-generated labels for summaries of dialogic argument while in section 5.3 the AFS regressor operated on pairs of high quality arguments extracted from social media conversations. Section 5.3.1 improved

the argument quality extractor from Swanson et al. (2015) and created a large dataset for Argument Quality (AQ) for three debate topics, namely Gun Control, Gay Marriage and Death Penalty. We then developed another corpus of sentential argument pairs annotated with AFS scores with high reliability. Establishing clear guidelines for the AFS task was crucial so that homogeneous and consistent annotations could be obtained, which was required to successfully apply machine learning algorithms that learn to predict AFS on new unseen data. To this end, we achieved an average pairwise correlation of 0.7 for the first set and 0.68 for the second set of annotations using Mechanical Turk. We proposed novel feature sets consisting of Ngrams, LIWC dependencies, and Rouge for the AFS task, and compared predictions using feature ablation studies across different models. The results from both the set of experiments showed that AFS with our proposed features significantly outperformed the STS baselines, validating $\mathbf{H_3}$.

In summary, we introduced a methodical approach to discover argument facets across multiple dialogues on topics of social and political importance. Our end-to- end argument extraction approach allows summarizing important arguments from a dialogue, whereas the two approaches on argument facet similarity enable the recognition of semantically similar arguments, i.e., arguments about the same facet. Together, these approaches represent significant steps towards processing real-world arguments in natural conversations, provide an output that would enable users to get an overview of issues in a discussion, and systematically reason about the arguments on both sides of an issue.

6.2 Limitations

Summarizing important arguments and argument facet induction is a first step towards bottom-up approach of facet induction in social media dialogue. The reported work has limitations, which we address in this section.

6.2.1 Manual Pyramid Annotation

Recognizing the central propositions in a dialogue required human intervention at two levels. The first step was to collect multiple summaries of a given dialogue, followed by the pyramid annotation of summaries. The summary collection was quick and easy as summarization did not require formal training, and was accomplished using Mechanical Turk, a crowdsourcing platform. Though pyramid annotations have proven to be highly reliable (Nenkova and McKeown, 2011), they were costly and time-consuming. The annotators were undergraduates with linguistic background and had to be trained for several hours on how to use the pyramid software. Even after the training phase is complete, pyramid generation is a slow process as the annotators iterate over the process until they are satisfied with the semantic content units, and the summary contributor assignments. The lack of automation is a major hindrance to the more widespread use of pyramid method for summarization. Yang et al. (2016) proposed Pyramid Evaluation via Automated Knowledge Extraction (PEAK), a first attempt to generate pyramids automatically. It uses open information extraction to identify subject-predicate-object triples as SCUs, and graphs constructed from the triples to identify and assign weights to SCUs. Semantic similarity of triples is analyzed to identify the contributors for a SCU. The method achieved

high correlation between scores based on an automated pyramid and scores based on a manual pyramid. The original text to be summarized was an elementary physics text. This data is quite different from social media dialogues which are informal, unstructured, and often noisy. Therefore, it could be useful to evaluate in future studies the performance of PEAK method on our dialogic social media data.

6.2.2 Domain Specific AFS

The system we have trained for AFS is domain and topic specific. Since it was a novel task and we developed the framework from scratch, we could not test it for generalization to other domains. To overcome this limitation, semi-supervised approaches such as clustering and topic modeling as well as topic independent features could be explored. Previous work shows that metrics used for evaluating machine translation quality perform well on paraphrase recognition tasks (Madnani et al., 2012). In our experiments, ROUGE performed very well, suggesting that other machine translation metrics such as Terp and Meteor may be useful (Snover et al., 2009; Lavie and Denkowski, 2009). Recently, Yin et al. (2016) used general Attention Based Convolutional Neural Network for modeling a pair of sentences. Interdependent sentence pair representations were shown to be more powerful than isolated sentence representations for answer selection (AS), paraphrase identification (PI) and textual entailment (TE). The architecture could be exploited further for developing a generalized topic independent framework for finding argument similarity.

6.2.3 Back-Linking Task of Pyramid Labels to Sentences

We only considered Tier 3 and above labels for the linking task. The lower tiers may have important information but we define the upper tiers as the most important content as per the pyramid method. The Mechanical Turk summaries were abstractive and hence sometimes may contain individual views and opinions such as "Author 1 and Author 2 are annoyed with each other". These types of statements that reflect personalized views of the summarizers often end up on lower tiers of the pyramid and may not map to any actual dialogue utterance in the back-linking task.

6.3 Future Work

6.3.1 Argument Quality Scores for Summary Construction

We used an off the shelf AQ tool from Swanson et al. (2015). It gives a score to each sentence indicating how clearly it expresses a particular argument facet. The tool was refined to improve the predicted scores for the sentences. Dialogue Sentences can be also ranked based on their relative AQ predicted scores. This ranking would produce another argument summary for a given dialogue. However, we need to retrain and rebuild the model from scratch to get the ranking for new dialogue sentences. It would be interesting to compare this type of argument summary with that obtained from pyramid linking task and may serve as another baseline for the argument dialogue summarization framework described in Chapter 4.

6.3.2 Argument Quality Pairs

The refined AQ extractor in Section 5.3.1 treated all posts on a topic equally, operating on a set of concatenated posts. As a consequence, the AQ extractor may eliminate arguments made by less articulate citizens. Every conversation has different dynamics and a single preference order across all discussions may not exist. An approach to get more diverse arguments could be based on argument quality labels between argument pairs within a dialogue. Sentences in the Annotated Dialogue Corpus with a score of 3 or greater than 3 are argumentative, and represent the relative importance of the content. This can provide an alternate definition of argument quality (AQ) score. The AQ score of each argument reflects the relative preference within the dialogue, giving a more diverse set of arguments where every dialogue in a debate contributes. For a given dialogue, the arguments with an average AQ score of 3 or 3.5 may be considered in the lower quality set while those with a score of 4.5 or 5 are the more preferred ones. Pairs can be generated by taking the cross product of these two sets, giving an argument preference corpus. We performed a preliminary investigation to determine what these pairs look like. Table 6.1 shows example pairs from the corpus, and argument comparison looks a promising direction for future work. Habernal and Gurevych (2016b) address a closely related task of argument comparison to measure qualitative properties of Web arguments based on their convincingness. However, they model pairs by concatenating the isolated argument representations. For the computational model, we plan to model the argument pair using Siamese architecture (Bromley et al., 1993). These networks have been used in the past to model sentence pairs for comparison (Yin et al., 2016), and thus we expect these models to be useful for argument

comparison tasks as well.

Topic	Argument 1	Argument2
Gun Control	**The original intent and purpose of the Second Amendment was to preserve and guarantee ,not grant the pre-existing right of individuals, to keep and bear arms.**	Criminals do not break either law , they become criminals after they broke them.
Gun Control	**Would you feel secure knowing people around you had guns.**	Surely its obvious to you that this poses far more threats than securities.
Gay Marriage	**Why is it in the interests of society to prevent government involvement in marriage?**	Why should government get out of the marriage business.
Gay Marriage	**I personally define marriage as a union of two people who love one another and wish to make a commitment to one another**	The cultural definition comes from observation.
Abortion	**People respected life unlike today.**	No, it was a different era.
Abortion	**Abortion is controversial for the simple fact that many people consider the procedure murder.**	Pain in the medical experiments I referred to was not intentional but a consequence of the experiment.

Table 6.1: Argument pairs with more preferred arguments in bold.

6.3.3 Explore Lower Tier Dialogue Sentences

The machine learning model did not distinguish between lower tiers, i.e. sentences that have not appeared in the summary to those that map to tiers 1 or 2. It is possible that some of the dialogues had important content in lower tiers, thus in future work it would be useful to refine the model further and analyze the misclassifications in lower tiers. For instance, consider the sample lower tier labels in Table 6.2 In future, one would get annotations for the lower tier mappings and see if the model can distinguish between the statements that map to a pyramid

S.No	Pyramid Label	Tier Rank
1	Religious freedom argument could be used in any format to justify any action.	2
2	Two people annoyed with each other by the end of dialogue.	1

Table 6.2: Lower tier labels.

label as compared to statements not present in the summary at all. It would also be beneficial to see which pyramid labels do not map to any sentence. For example, sentence 2 in the Table 6.2 may not map to any sentence in the dialogue as it represents an abstractive general viewpoint of a summarizer.

6.3.4 Fuzzy Argument Clustering

In future, we plan to use our AFS regressor to cluster and group similar arguments and produce argument facet summaries as a final output of our pipeline. However, since AFS is based on the notion of graded similarity, argument propositions in a facet overlap with varying levels of similarity. Therefore, a fuzzy clustering approach where arguments are allowed to be present in multiple clusters with certain probabilities would be a better fit to the problem. A fuzzy clustering algorithm that could be explored using AFS similarity pairs is given by Skabar and Abdalgader (2013). Here the authors proposed a fuzzy relational sentence level clustering algorithm that can be applied to any domain in which the relationship between objects is expressed in terms of pair wise similarities. Additionally, since the dialogue corpus comes

with stance annotations, we will also sample by stance-side, so that summaries can be organized using Pro and Con, as in curated summaries.

Our final goal is to combine quality-based argument extraction, our AFS model, stance, post and author level information, so that our summaries represent the wide diversity of views and opinions expressed on a topic in a debate dialogue.

Bibliography

(2012). *SemEval '12: Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Abbott, R., Ecker, B., Anand, P., and Walker, M. (2016). Internet argument corpus 2.0: An sql schema for dialogic social media and the corpora to go with it. In *Language Resources and Evaluation Conference, LREC2016*.

Abbott, R., Walker, M., Jean E. Fox Tree, Anand, P., Bowmani, R., and King, J. (2011). How can you say such things?!?: Recognizing Disagreement in Informal Political Argument. In *Proc. of the ACL Workshop on Language and Social Media*.

Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., and Guo, W. (2013). Sem 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity. In *In* SEM 2013: The Second Joint Conference on Lexical and Computational Semantics*.

Agirre, E., Diab, M., Cer, D., and Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot

on semantic textual similarity. In *Proc. of the First Joint Conference on Lexical and Computational Semantics*, volume 1, pages 385–393.

Agrawal, R., Rajagopalan, S., Srikant, R., and Xu, Y. (2003). Mining newsgroups using networks arising from social behavior. In *Proc. of the 12th international conference on World Wide Web*, pages 529–535.

Aharoni, E., Polnarov, A., Lavee, T., Hershcovich, D., Levy, R., Rinott, R., Gutfreund, D., and Slonim, N. (2014). A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the First Workshop on Argument Mining, hosted by the 52nd Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2014, June 26, 2014, Baltimore, Maryland, USA*, pages 64–68. The Association for Computer Linguistics.

Allen, K., Carenini, G., and Ng, R. T. (2014). Detecting disagreement in conversations using pseudo-monologic rhetorical structure. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*.

Anand, P., Walker, M., Abbott, R., Jean E. Fox Tree, Bowmani, R., and Minor, M. (2011). Cats Rule and Dogs Drool: Classifying Stance in Online Debate. In *Proc. of the ACL Workshop on Sentiment and Subjectivity*.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *in Proc. of LREC*.

Beineke, P., Hastie, T., Manning, C., and Vaithyanathan, S. (2004). An exploration of sentiment

summarization. Technical report, In AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications (AAAI.

Berg-Kirkpatrick, T., Gillick, D., and Klein, D. (2011). Jointly learning to extract and compress. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 481–490, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bex, F. and Walton, D. (2016). Combining explanation and argumentation in dialogue. *Argument & Computation*, 7(1):55–68.

Biran, O. and Rambow, O. (2011). Identifying justifications in written dialogs. In *2011 Fifth IEEE International Conference on Semantic Computing (ICSC)*, pages 162–168.

Blair, J. A. (2004). Argument and its uses. *Informal Logic*, 24(2):137–151.

Boltuzic, F. and Šnajder, J. (2014). Back up your stance: Recognizing arguments in online discussions. In *Proc. of the First Workshop on Argumentation Mining*, pages 49–58.

Boltuzic, F. and Šnajder, J. (2015). Identifying prominent arguments in online debates using semantic textual similarity. In *Proc. of the Second Workshop on Argumentation Mining*.

Bromley, J., Bentz, J. W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., and Shah, R. (1993). Signature verification using A "siamese" time delay neural network. *IJPRAI*, 7(4):669–688.

Budzynska, K. and Reed, C. (2011). Speech acts of argumentation: Inference anchors and

peripheral cues in dialogue. In *Proceedings of the 10th AAAI Conference on Computational Models of Natural Argument*, AAAIWS'11-10, pages 3–10. AAAI Press.

Cabrio, E., Hirst, G., Villata, S., and Wyner, A. (2016). Natural Language Argumentation: Mining, Processing, and Reasoning over Textual Arguments (Dagstuhl Seminar 16161). *Dagstuhl Reports*, 6(4):80–109.

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

Carbonell, J. and Goldstein, J. (1998). The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 335–336, New York, NY, USA. ACM.

Chen, W.-F. and Ku, L.-W. (2016). Utcnn: a deep learning model of stance classification on social media text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1635–1645, Osaka, Japan. The COLING 2016 Organizing Committee.

Cheng, J. and Lapata, M. (2016). Neural summarization by extracting sentences and words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Chollet, F. (2015). Keras.

Christensen, J., Mausam, Soderland, S., and Etzioni, O. (2013). Towards coherent multi-document summarization. In Vanderwende, L., III, H. D., and Kirchhoff, K., editors, *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA*, pages 1163–1173. The Association for Computational Linguistics.

Conrad, A., Wiebe, J., et al. (2012). Recognizing arguing subjectivity and argument tags. In *Proc. of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 80–88.

Dung, P. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games* 1. *Artificial intelligence*, 77(2):321–357.

Erkan, G. and Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Int. Res.*, 22(1):457–479.

Feng, V. W. and Hirst, G. (2011). Classifying arguments by scheme. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 987–996.

Filippova, K., Alfonseca, E., Colmenares, C. A., Kaiser, L., and Vinyals, O. (2015). Sentence compression by deletion with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 360–368.

Florou, E., Konstantopoulos, S., Koukourikos, A., and Karampiperis, P. (2013). Argument

extraction for supporting public policy formulation. In Lendvai, P. and Zervanou, K., editors, *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities, LaTeCH@ACL 2013, August 8, 2013, Sofia, Bulgaria*, pages 49–54. The Association for Computer Linguistics.

Freeman, J. (1991). *Dialectics and the Macrostructure of Arguments: A Theory of Argument Structure*. Pragmatics and Discourse Analysis Series. Foris Publications.

Gabbriellini, S. and Torroni, P. (2013). Ms dialogues: Persuading and getting persuaded. a model of social network debates that reconciles arguments and trust. *Proc. 10th ArgMAS*.

Galley, M., McKeown, K., Hirschberg, J., and Shriberg, E. (2004). Identifying agreement and disagreement in conversational speech: Use of bayesian networks to model pragmatic dependencies. In *Proc. of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 669–es.

Garcia-Mila, M., Gilabert, S., Erduran, S., and Felton, M. (2013). The effect of argumentative task goal on the quality of argumentative discourse. *Science Education*, 97(4):497–523.

Germesin, S. and Wilson, T. (2009). Agreement detection in multiparty conversation. In *Proceedings of the 2009 International Conference on Multimodal Interfaces*, ICMI-MLMI '09, pages 7–14, New York, NY, USA. ACM.

Ghosh, D., Muresan, S., Wacholder, N., Aakhus, M., and Mitsui, M. (2014). Analyzing argumentative discourse units in online interactions. In *ArgMining@ACL*, pages 39–48. The Association for Computer Linguistics.

Gilbert, M. A. (1997). Coalescent argumentation.

Gillick, D. and Favre, B. (2009). A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, ILP '09, pages 10–18, Stroudsburg, PA, USA. Association for Computational Linguistics.

Goudas, T., Louizos, C., Petasis, G., and Karkaletsis, V. (2014). Argument extraction from news, blogs, and social media. In *Artificial Intelligence: Methods and Applications*, pages 287–299. Springer.

Groarke, L. (2017). Informal logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition.

Groza, T., Handschuh, S., and Breslin, J. G. (2008). Adding provenance and evolution information to modularized argumentation models. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE.

Gurevych, I. and Strube, M. (2004). Semantic similarity applied to spoken dialogue summarization. In *Proc. of the 20th international conference on Computational Linguistics*, pages 764–771. ACL.

Habernal, I., Eckle-Kohler, J., and Gurevych, I. (2014). Argumentation mining on the web from information seeking perspective. In *Proc. of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 26–39.

Habernal, I. and Gurevych, I. (2015). Exploiting debate portals for semi-supervised argumenta-

tion mining in user-generated web discourse. In *Proc. of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2127–2137.

Habernal, I. and Gurevych, I. (2016a). What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In Su, J., Carreras, X., and Duh, K., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1214–1223. The Association for Computational Linguistics.

Habernal, I. and Gurevych, I. (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, ACL*. The Association for Computational Linguistics.

Habernal, I. and Gurevych, I. (2017). Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179.

Haghighi, A. and Vanderwende, L. (2009). Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL '09, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hahn, S., Ladner, R., and Ostendorf, M. (2006). Agreement/disagreement classification: exploiting unlabeled data using contrast classifiers. In *Proc. of the Human Language Technol-*

ogy Conference of the NAACL, Companion Volume: Short Papers, NAACL-Short '06, pages 53–56. Association for Computational Linguistics.

Hall, M., Eibe, F., Holms, G., Pfahringer, B., Reutemann, P., and Witten, I. (2005). The weka data mining software: An update. *SIGKDD Explorations*, 11(1).

Han, L., Kashyap, A., Finin, T., Mayfield, J., and Weese, J. (2013). Umbc ebiquity-core: Semantic textual similarity systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, pages 44–52.

Harman, D. and Over, P. (2002). The duc summarization evaluations. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 44–51, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Harris, Z. S. (1954). Distributional structure. *WORD*, 10(2-3):146–162.

Hasan, K. S. and Ng, V. (2013a). Frame semantics for stance classification. In *CoNLL*, pages 124–132.

Hasan, K. S. and Ng, V. (2013b). Stance classification of ideological debates: Data, models, features, and constraints. In *International Joint Conference on Natural Language Processing*, pages 1348–1356.

Hasan, K. S. and Ng, V. (2014). Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*.

Hazen, T. J. (2011). Latent topic modeling for audio corpus summarization. In *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, pages 913–916. ISCA.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA. MIT Press.

Hillard, D., Ostendorf, M., and Shriberg, E. (2003). Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2*, NAACL-Short '03, pages 34–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

Iyer, S., Konstas, I., Cheung, A., and Zettlemoyer, L. (2016). Summarizing source code using a neural attention model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Jackson, S. and Jacobs, S. (1980). Structure of conversational argument: Pragmatic bases for the enthymeme. *Quarterly Journal of Speech*, 66(3):251–265.

Jacobs, S. (1989). Speech acts and arguments. *Argumentation*, 3(4):345–365.

Jacobs, S. and Jackson, S. (1982). Conversational argument: A discourse analytic approach. *Advances in argumentation theory and research*, pages 205–237.

Jacobs, S. and Jackson, S. (1992). Relevance and digressions in argumentative discussion: A pragmatic approach. *Argumentation*.

Johnson, R. and Blair, J. (2006). *Logical Self-defense*. Key titles in rhetoric, argumentation, and debate series. International Debate Education Association.

Joshi, M. and Penstein-Rosé, C. (2009). Generalizing dependency features for opinion mining. In *Proc. of the ACL-IJCNLP 2009 Conference Short Papers*, pages 313–316.

Kim, H. D., Ganesan, K., Sondhi, P., and Zhai, C. (2011). Comprehensive review of opinion summarization.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *EMNLP*, pages 1746–1751. ACL.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Kobayashi, H., Noguchi, M., and Yatsuka, T. (2015). Summarization based on embedding distributions. In Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., and Marton, Y., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1984–1989. The Association for Computational Linguistics.

Kolb, P. (2008). Disco: A multilingual database of distributionally similar words. In *Proc. of KONVENS-2008*.

Kouylekov, M. and Negri, M. (2010). An open-source package for recognizing textual entailment. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden, System Demonstrations*, pages 42–47. The Association for Computer Linguistics.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lavie, A. and Denkowski, M. J. (2009). The meteor metric for automatic evaluation of machine translation. *Machine Translation*, 23(2-3):105–115.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In Jebara, T. and Xing, E. P., editors, *Proc. of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

Levy, R., Bilu, Y., Hershcovich, D., Aharoni, E., and Slonim, N. (2014). Context dependent claim detection. In Hajic, J. and Tsujii, J., editors, *COLING 2014, 25th International Con-*

ference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland, pages 1489–1500. ACL.

Li, J., Chen, X., Hovy, E., and Jurafsky, D. (2015). Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*.

Li, J. J. and Nenkova, A. (2015). Fast and accurate prediction of sentence specificity. In *Proceedings of the Twenty-Ninth Conference on Artificial Intelligence (AAAI)*, pages 2281–2287.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries rouge: A package for automatic evaluation of summaries. In *Proc. of the Workshop on Text Summarization Branches Out (WAS 2004)*.

Lin, H. and Bilmes, J. (2011). A class of submodular functions for document summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 510–520, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lin, S.-H., Yeh, Y.-M., and Chen, B. (2011). Leveraging kullback-leibler divergence measures and information-rich cues for speech summarization. *Trans. Audio, Speech and Lang. Proc.*, 19(4):871–882.

Lippi, M. and Torroni, P. (2015a). Argument mining: A machine learning perspective. In Black, E., Modgil, S., and Oren, N., editors, *Theory and Applications of Formal Argumentation*, pages 163–176, Cham. Springer International Publishing.

Lippi, M. and Torroni, P. (2015b). Context-independent claim detection for argument mining. In Yang, Q. and Wooldridge, M., editors, *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 185–191. AAAI Press.

Liu, B. (2015). Opinion summarization and search. In *Sentiment Analysis*, pages 218–230. Cambridge University Press.

Loper, E. and Bird, S. (2002). NLTK: The natural language toolkit. In *Proc. of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics.

Louis, A. and Nenkova, A. (2013). Automatically assessing machine summary content without a gold standard. *Computational Linguistics*, 39(2):267–300.

Lukin, S. and Walker, M. (2013). Really? well. apparently bootstrapping improves the performance of sarcasm and nastiness classifiers for online dialogue. In *Proceedings of the Workshop on Language Analysis in Social Media*, pages 30–40, Atlanta, Georgia. Association for Computational Linguistics.

Lukin, S. M., Walker, M. A., Anand, P., and Whittaker, S. (2017). Argument strength is in the eye of the beholder: Audience effects in persuasion. In *EACL (1)*, pages 742–753. Association for Computational Linguistics.

Madnani, N., Tetreault, J., and Chodorow, M. (2012). Re-examining machine translation met-

rics for paraphrase identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 182–190, Stroudsburg, PA, USA. Association for Computational Linguistics.

Majumder, G., Pakray, P., Gelbukh, A., and Pinto, D. (2016). Semantic textual similarity methods, tools, and applications: A survey. *Computación y Sistemas*, 20(4).

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory. Toward a functional theory of text organization. *Text*, 8(3):243–281.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proc. of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Marcu, D. (1999). Discourse trees are good indicators of importance in text. *Advances in automatic text summarization*, pages 123–136.

Maskey, S. and Hirschberg, J. (2005). Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, pages 621–624. ISCA.

McDonald, R. (2007). A study of global inference algorithms in multi-document summarization. In *Proceedings of the 29th European Conference on IR Research*, ECIR'07, pages 557–564, Berlin, Heidelberg. Springer-Verlag.

McKeown, K., Hirschberg, J., Galley, M., and Maskey, S. (2005). From text summarization to speech summarization. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP05), Special session on Human Language Technology: Applications and Challenges for Speech Processing.*

Mei, Q., Guo, J., and Radev, D. (2010). Divrank: The interplay of prestige and diversity in information networks. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, pages 1009–1018, New York, NY, USA. ACM.

Mihalcea, R. and Tarau, P. (2004). Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing , EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 404–411. ACL.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Misra, A., Anand, P., Jean E. Fox Tree, and Walker, M. (2015). Using summarization to discover argument facets in dialog. In *Proc. of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Misra, A., Ecker, B., and Walker, M. (2016). Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the SIGDIAL 2016 Conference*. Association for Computational Linguistics, Association for Computational Linguistics.

Misra, A., Tandon, S., Ts, S., Anand, P., and Walker, M. (2017). Summarizing dialogic arguments from social media. In *SEMDIAL 2017 (SaarDial) Workshop on the Semantics and Pragmatics of Dialogue*. ISCA.

Misra, A. and Walker, M. A. (2013). Topic independent identification of agreement and disagreement in social media dialogue. In *Proc. of the SIGDIAL 2013 Conference: The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*.

Mochales, R. and Ieven, A. (2009). Creating an argumentation corpus: Do theories apply to real arguments?: A case study on the legal argumentation of the echr. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 21–30, New York, NY, USA. ACM.

Moens, M.-F. (2013). Argumentation mining: Where are we now, where do we want to be and how do we get there? In *Proceedings of the 5th 2013 Forum on Information Retrieval Evaluation - FIRE13*.

Moens, M.-F., Boiy, E., Palau, R. M., and Reed, C. (2007). Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, ICAIL '07, pages 225–230, New York, NY, USA. ACM.

Morita, H., Sasano, R., Takamura, H., and Okumura, M. (2013). Subtree extractive summa-

rization via submodular maximization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 1023–1032. The Association for Computer Linguistics.

Murakami, A. and Raymond, R. (2010). Support or oppose?: classifying positions in online debates from reply activities and opinion expressions. In *Proc. of the 23rd International Conference on Computational Linguistics: Posters*, pages 869–875. Association for Computational Linguistics.

Murray, G., Renals, S., and Carletta, J. (2005). Extractive summarization of meeting recordings. In *in Proceedings of the 9th European Conference on Speech Communication and Technology*, pages 593–596.

Murray, G., Renals, S., Carletta, J., and Moore, J. (2006). Incorporating speaker and discourse features into speech summarization. In *Proc. of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 367–374. Association for Computational Linguistics.

Naderi, N. and Hirst, G. (2016). Argumentation mining in parliamentary discourse. In *Principles and Practice of Multi-Agent Systems*, pages 16–25. Springer International Publishing.

Nallapati, R., Zhou, B., and Zhou, B. (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. In *CoNLL*. The Association for Computational Linguistics.

Nenkova, A. and McKeown, K. (2011). Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233.

Nenkova, A. and Passonneau, R. (2004). Evaluating content selection in summarization: The pyramid method. In *Proc. of HLT-NAACL*, volume 2004.

Nenkova, A., Passonneau, R., and McKeown, K. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4–es.

Nenkova, A. and Vanderwende, L. (2005). The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005*.

Nesselhauf, N. (2004). Learner corpora and their potential for language teaching. In *How to Use Corpora in Language Teaching*, pages 125–152. John Benjamins Publishing Company.

Nguyen, H. and Litman, D. J. (2015). Extracting argument and domain words for identifying argument components in texts. In *ArgMining@HLT-NAACL*, pages 22–28. The Association for Computational Linguistics.

Oraby, S., Harrison, V., Reed, L., Hernandez, E., Riloff, E., and Walker, M. (2016). Creating and characterizing a diverse corpus of sarcasm in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics.

Oraby, S., Reed, L., Compton, R., Riloff, E., Walker, M. A., and Whittaker, S. (2015). And that's A fact: Distinguishing factual and emotional argumentation in online dialogue. In *ArgMining@HLT-NAACL*, pages 116–126. The Association for Computational Linguistics.

Ouyang, Y., Li, W., Wei, F., and Lu, Q. (2009). Learning similarity functions in graph-based

document summarization. In *Proceedings of the 22Nd International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*, ICCPOL '09, pages 189–200, Berlin, Heidelberg. Springer-Verlag.

Oya, T. and Carenini, G. (2014). Extractive summarization and dialogue act modeling on email threads: An integrated probabilistic approach. In *Proceedings of the SIGDIAL 2014 Conference, The 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 18-20 June 2014, Philadelphia, PA, USA*, pages 133–140.

Palau, R. M. and Moens, M. (2008). Study on the structure of argumentation in case law. In *Legal Knowledge and Information Systems - JURIX 2008: The Twenty-First Annual Conference on Legal Knowledge and Information Systems, Florence, Italy, 10-13 December 2008*, pages 11–20.

Palau, R. M. and Moens, M.-F. (2009). Argumentation mining: The detection, classification and structure of arguments in text. In *Proc. of the 12th International Conference on Artificial Intelligence and Law*, ICAIL '09, pages 98–107, New York, NY, USA. ACM.

Park, J. and Cardie, C. (2014). Identifying appropriate support for propositions in online user comments. In *ArgMining@ACL*, pages 29–38. The Association for Computer Linguistics.

Paulus, R., Xiong, C., and Socher, R. (2017). A deep reinforced model for abstractive summarization. *CoRR*, abs/1705.04304.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher,

M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peldszus, A. and Stede, M. (2013). From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 7(1):1–31.

Peldszus, A. and Stede, M. (2015). Joint prediction in mst-style discourse parsing for argumentation mining. In Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., and Marton, Y., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 938–948. The Association for Computational Linguistics.

Peldszus, A. and Stede, M. (2016). An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015 / Vol. 2*, pages 801–815, London. College Publications.

Pennebaker, J. W., Francis, M. E., and Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP*.

Perelman, C. and Olbrechts-Tyteca, L. (1969). *The New Rhetoric: A Treatise on Argumentation*. University of Notre Dame Press.

Persing, I. and Ng, V. (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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 543–552. The Association for Computer Linguistics.

Pitler, E. and Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *EMNLP*, pages 186–195. ACL.

Prabhakaran, V., Rambow, O., and Diab, M. T. (2010). Automatic committed belief tagging. In Huang, C. and Jurafsky, D., editors, *COLING 2010, 23rd International Conference on Computational Linguistics, Posters Volume, 23-27 August 2010, Beijing, China*, pages 1014–1022. Chinese Information Processing Society of China.

Radev, D. R., Jing, H., Stys, M., and Tam, D. (2004). Centroid-based summarization of multiple documents. *Inf. Process. Manage.*, 40(6):919–938.

Rajendran, P., Bollegala, D., and Parsons, S. (2016). Contextual stance classification of opinions: A step towards enthymeme reconstruction in online reviews. In *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*. The Association for Computer Linguistics.

Rambow, O., Shrestha, L., Chen, J., and Lauridsen, C. (2004). Summarizing email threads. In

Proceedings of HLT-NAACL 2004: Short Papers, HLT-NAACL-Short '04, pages 105–108, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ranade, S., Gupta, J., Varma, V., and Mamidi, R. (2013a). Online debate summarization using topic directed sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, WISDOM '13, New York, NY, USA. ACM.

Ranade, S., Sangal, R., and Mamidi, R. (2013b). Stance classification in online debates by recognizing users' intentions. In *SIGDIAL Conference*, pages 61–69. The Association for Computer Linguistics.

Reed, C. and Rowe, G. (2004). Araucaria: Software for argument analysis, diagramming and representation. *International Journal on Artificial Intelligence Tools*, 13(04):961–979.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proc. of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50.

Rinott, R., Dankin, L., Perez, C. A., Khapra, M. M., Aharoni, E., and Slonim, N. (2015). Show me your evidence - an automatic method for context dependent evidence detection. In Màrquez, L., Callison-Burch, C., Su, J., Pighin, D., and Marton, Y., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 440–450. The Association for Computational Linguistics.

Roitman, H., Hummel, S., Rabinovich, E., Sznajder, B., Slonim, N., and Aharoni, E. (2016). On the retrieval of wikipedia articles containing claims on controversial topics. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, pages 991–996, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Rooney, N., Wang, H., and Browne, F. (2012). Applying kernel methods to argumentation mining. In *Florida AI Conference*.

Rosenthal, S. and McKeown, K. (2012). Detecting opinionated claims in online discussions. In *Sixth IEEE International Conference on Semantic Computing, ICSC 2012, Palermo, Italy, September 19-21, 2012*, pages 30–37. IEEE Computer Society.

Rosenthal, S. and McKeown, K. (2015). I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 2-4 September 2015, Prague, Czech Republic*, pages 168–177.

Rush, A. M., Chopra, S., and Weston, J. (2015). A neural attention model for abstractive sentence summarization. In *EMNLP*. The Association for Computational Linguistics.

Salmon, M. H. and Zeitz, C. M. (1995). Analyzing conversational reasoning. *Informal Logic*, 17(1).

Sardianos, C., Katakis, I. M., Petasis, G., and Karkaletsis, V. (2015). Argument extraction from news. In *Proceedings of the 2nd Workshop on Argumentation Mining, ArgMining@HLT-*

NAACL 2015, June 4, 2015, Denver, Colorado, USA, pages 56–66. The Association for Computational Linguistics.

Searle, J. R. (1975). Indirect speech acts. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics III: Speech Acts*, pages 59–82. Academic Press, New York.

See, A., Manning, C., and Liu, P. (2017). Get to the point: Summarization with pointer-generator networks. In *Association for Computational Linguistics*.

Serban, I. V., Lowe, R., Henderson, P., Charlin, L., and Pineau, J. (2015). A survey of available corpora for building data-driven dialogue systems. *CoRR*, abs/1512.05742.

Siegel, H. (2015). *Argumentative Norms: How Contextual Can They Be? A Cautionary Tale*, pages 205–215. Springer International Publishing, Cham.

Sipos, R., Shivaswamy, P., and Joachims, T. (2012). Large-margin learning of submodular summarization models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 224–233, Stroudsburg, PA, USA. Association for Computational Linguistics.

Skabar, A. and Abdalgader, K. (2013). Clustering sentence-level text using a novel fuzzy relational clustering algorithm. *IEEE Trans. on Knowl. and Data Eng.*, 25(1):62–75.

Skeppstedt, M., Sahlgren, M., Paradis, C., and Kerren, A. (2016). Unshared task: (dis)agreement in online debates. In *Proceedings of the Third Workshop on Argument Mining, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2016, August 12, Berlin, Germany*.

Snover, M., Madnani, N., Dorr, B., and Schwartz, R. (2009). Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 259–268, Athens, Greece. Association for Computational Linguistics.

Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D. (2011a). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. In *NIPS*, pages 801–809.

Socher, R., Pennington, J., Huang, E. H., Ng, A. Y., and Manning, C. D. (2011b). Semi-supervised recursive autoencoders for predicting sentiment distributions. In *EMNLP*, pages 151–161. ACL.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, Washington, USA. Association for Computational Linguistics.

Somasundaran, S. and Wiebe, J. (2009). Recognizing stances in online debates. In *Proc. of the 47th Annual Meeting of the ACL*, pages 226–234.

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

Song, Y., Heilman, M., Klebanov, B. B., and Deane, P. (2014). Applying argumentation

schemes for essay scoring. In *Proceedings of the First Workshop on Argument Mining, hosted by the 52nd Annual Meeting of the Association for Computational Linguistics, ArgMining@ACL 2014, June 26, 2014, Baltimore, Maryland, USA*, pages 69–78. The Association for Computer Linguistics.

Sridhar, D., Foulds, J., Huang, B., Getoor, L., and Walker, M. (2015). Joint models of disagreement and stance in online debate. In *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Stab, C. and Gurevych, I. (2014a). Annotating argument components and relations in persuasive essays. In *COLING*, pages 1501–1510. ACL.

Stab, C. and Gurevych, I. (2014b). Identifying argumentative discourse structures in persuasive essays. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 46–56. ACL.

Stegmann, K., Wecker, C., Weinberger, A., and Fischer, F. (2012). Collaborative argumentation and cognitive elaboration in a computer-supported collaborative learning environment. *Instructional Science*, 40(2):297–323.

Stojanovski, D., Strezoski, G., Madjarov, G., and Dimitrovski, I. (2015). Twitter sentiment analysis using deep convolutional neural network. In *Hybrid Artificial Intelligent Systems -*

10th International Conference, HAIS 2015, Bilbao, Spain, June 22-24, 2015, Proceedings, pages 726–737.

Swanson, R., Ecker, B., and Walker, M. (2015). Argument mining: Extracting arguments from online dialogue. In *Proc. of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 217–226.

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., and Lee, L. (2016). Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 613–624, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Teufel, S., Siddharthan, A., and Batchelor, C. R. (2009). Towards domain-independent argumentative zoning: Evidence from chemistry and computational linguistics. In *EMNLP*, pages 1493–1502. ACL.

Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proc. of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics.

Toni, F. and Torroni, P. (2012). Bottom-up argumentation. In *Theorie and Applications of Formal Argumentation*, pages 249–262. Springer Berlin Heidelberg.

Toulmin, S. E. (1958). *The Uses of Argument*. Cambridge University Press.

Van Eemeren, F. (1995). *Proceedings of the Third ISSA Conference on Argumentation: Perspectives and approaches*. Perspectives and approaches. Sic Sat.

Van Eemeren, F. and Grootendorst, R. (2004). *A Systematic Theory of Argumentation: The Pragma-dialectical Approach*. A Systematic Theory of Argumentation: The Pragma-dialectical Approach. Cambridge University Press.

Van Eemeren, F., Krabbe, E., and Henkemans, A. (2014). *Handbook of Argumentation Theory*. Springer Verlag.

Villata, S., Cabrio, E., Jraidi, I., Benlamine, S., Chaouachi, M., Frasson, C., and Gandon, F. (2017). Emotions and personality traits in argumentation: An empirical evaluation1. *Argument & Computation*, 8(1):61–87.

Wachsmuth, H., Stein, B., and Ajjour, Y. (2017). Pagerank for argument relevance. In Lapata, M., Blunsom, P., and Koller, A., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 1117–1127. Association for Computational Linguistics.

Walker, M. A., Anand, P., Abbott, R., and Grant, R. (2012a). Stance classification using dialogic properties of persuasion. In *Proc. of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 592–596.

Walker, M. A., Anand, P., Abbott, R., Tree, J. E. F., Martell, C., and King, J. (2012b). That

is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729.

Walker, M. A., Tree, J. E. F., Anand, P., Abbott, R., and King, J. (2012c). A corpus for research on deliberation and debate. In *LREC*, pages 812–817.

Walton, D. (1989). *Informal Logic: A Handbook for Critical Argument*. Bibliografía e índice. Cambridge University Press.

Walton, D. (1992). *Plausible Argument in Everyday Conversation*. SUNY Series, Educational Leadership. State University of New York Press.

Walton, D. and Krabbe, E. (1995). *Commitment in Dialogue: Basic concept of interpersonal reasoning*. State University of New York Press.

Walton, D., Reed, C., and Macagno, F. (2008). *Argumentation Schemes*. Cambridge University Press.

Wang, L. and Cardie, C. (2011). Summarizing decisions in spoken meetings. In *Proceedings of the Workshop on Automatic Summarization for Different Genres, Media, and Languages*, WASDGML '11, pages 16–24, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wang, L. and Ling, W. (2016). Neural network-based abstract generation for opinions and arguments. In *HLT-NAACL*. The Association for Computational Linguistics.

Wang, W., Yaman, S., Precoda, K., Richey, C., and Raymond, G. (2011). Detection of agreement and disagreement in broadcast conversations. In *The 49th Annual Meeting of the Asso-*

ciation for Computational Linguistics: Human Language Technologies, Proc. of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers, pages 374–378. The Association for Computer Linguistics.

Wang, X., Jiang, W., and Luo, Z. (2016). Combination of convolutional and recurrent neural network for sentiment analysis of short texts. In *COLING*.

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

Woodsend, K. and Lapata, M. (2012). Multiple aspect summarization using integer linear programming. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 233–243, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yang, Q., Passonneau, R. J., and de Melo, G. (2016). PEAK: pyramid evaluation via automated knowledge extraction. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.*, pages 2673–2680.

Yasunaga, M., Zhang, R., Meelu, K., Pareek, A., Srinivasan, K., and Radev, D. R. (2017). Graph-based neural multi-document summarization. In Levy, R. and Specia, L., editors, *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL*

2017), Vancouver, Canada, August 3-4, 2017, pages 452–462. Association for Computational Linguistics.

Yin, J., Thomas, P., Narang, N., and Paris, C. (2012). Unifying local and global agreement and disagreement classification in online debates. In *Proc. of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 61–69.

Yin, W., Kann, K., Yu, M., and Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *CoRR*, abs/1702.01923.

Yin, W. and Pei, Y. (2015). Optimizing sentence modeling and selection for document summarization. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1383–1389. AAAI Press.

Yin, W., Schütze, H., Xiang, B., and Zhou, B. (2016). ABCNN: attention-based convolutional neural network for modeling sentence pairs. *TACL*, 4:259–272.

Zechner, K. (2001). Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *Proc. of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 199–207. ACM.

Zechner, K. (2002). Automatic summarization of open-domain multiparty dialogues in diverse genres. *Comput. Linguist.*, 28(4):447–485.

Zhou, C., Sun, C., Liu, Z., and Lau, F. C. M. (2015). A C-LSTM neural network for text classification. *CoRR*, abs/1511.08630.

Zhu, X. and Penn, G. (2006). Summarization of spontaneous conversations. In *Interspeech*.