©Copyright 2017
Ethan Roday

Three Cheers For Partisanship:
Lexical Framing and Applause in U.S. Presidential Primary Debates

Ethan Roday

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2017

Committee:

Gina-Anne Levow, Chair

Sayan Pathak

Program Authorized to Offer Degree:
Department of Linguistics

University of Washington

Abstract

Three Cheers For Partisanship:
Lexical Framing and Applause in U.S. Presidential Primary Debates

Ethan Roday

Chair of the Supervisory Committee:
Associate Professor Gina-Anne Levow
Department of Linguistics

Polarization in American politics is at its highest levels in recent history. This polarization can be observed not only in the behaviors of citizens and the politicians who represent them, but also in the rhetoric that politicians use and the reactions of voters to that rhetoric. In this work, I study the language used by candidates in presidential primary debates, and consider the audience's applause (or lack thereof) as a measure of the success of such language. I hypothesize that applause is more likely to occur when the language being used is highly polarized. While previous analyses of voter-directed speech have focused largely on rhetorical structure, this study examines the semantic content of applause-generating language through the automatic discovery of issue-specific lexical framing strategies. Specifically, I present an analysis that 1) models the topics present in a corpus of 104 primary debates, 2) quantifies the party polarization of the language used to discuss those topics, and 3) measures the association between audience applause and topic-specific party polarization. While the relationship is more pronounced for some issues than for others, the results of the analysis lend strong support to the hypothesis that applause is significantly and positively associated with polarity.

TABLE OF CONTENTS

Page

List of Figures iii

List of Tables iv

Chapter 1: Introduction 1

Chapter 2: Background 3
2.1 Rhetoric and Political Speech 3
2.2 Framing 4
2.3 Lexical Framing and Computational Political Science 6
2.4 Sentiment Analysis 7
2.5 Topic Modeling 9
2.5.1 Latent Dirichlet Allocation 11
2.5.2 Evaluation of Topic Models 13

Chapter 3: Methods 17
3.1 Dataset 17
3.1.1 Data Acquisition and Preprocessing 18
3.1.2 Composition of the Dataset 21
3.1.3 Distributional Statistics Over the Parsed Dataset 22
3.2 Phase I: Topic Modeling 24
3.2.1 Defining Documents 25
3.2.2 Document Preprocessing and LDA Implementation 26
3.2.3 Distributional Statistics Over the Processed Dataset 28
3.2.4 Notation for the Processed Dataset 29
3.2.5 Model Selection: The Healthcare Problem 30
3.2.6 Model Selection: Detecting Bad Topic Pairs 33

3.2.7 Model Selection: The Final Model . . . 41
3.3 Phase II: Measuring Polarity . . . 45
3.3.1 Log Odds Ratio . . . 46
3.3.2 Variance-Weighted Log Odds Ratio . . . 49
3.3.3 Variance-Weighted Log Odds Ratio With An Informative Bayesian Prior 54
3.3.4 Document-Level Polarity . . . 55
3.4 Phase III: Associating Applause and Polarity . . . 57
3.4.1 Attributing Applause . . . 57
3.4.2 Measuring Applause . . . 58
3.4.3 Testing the Hypothesis . . . 59

Chapter 4: Results and Discussion . . . 62
4.1 Results . . . 62
4.2 Interpretation of the Results . . . 63
4.2.1 Statistical Significance . . . 63
4.2.2 Goodness of Fit . . . 63
4.2.3 Ideological Unity . . . 66
4.2.4 Topic Ownership . . . 69
4.3 Limitations . . . 71
4.3.1 Data Limitations . . . 71
4.3.2 Modeling Limitations . . . 72
4.3.3 Analysis Limitations . . . 73
4.4 Future Directions . . . 73

Chapter 5: Conclusion . . . 76

Bibliography . . . 78

Appendix A: List of Stop Words . . . 87

LIST OF FIGURES

Figure Number Page

3.1 Histogram of $ITC \cdot CPD$ values over all most-related topic pairs. 42
3.2 Distribution of model scores from all models split by k. 42
3.3 Log-likelihood per iteration of collapsed Gibbs sampling in the final topic model. 45
3.4 Distribution of $Z(LOR_w^{\mathcal{D}-\mathcal{R}})$ scores for each term on a log scale. 56
3.5 Distribution of $Z(LOR_w^{\mathcal{N}-\mathcal{A}})_{reg}$ scores for each term on a log scale. 60

4.1 Results for logistic and linear regression on the IMMIGRATION topic. 67
4.2 Results for logistic and linear regression on the HEALTH topic. 68
4.3 Results for logistic and linear regression on the RACE topic. 70

LIST OF TABLES

Table Number Page

3.1 Number of debates by political party and election year. 21
3.2 Number of debates by political party and election year after removing debates with no applause transcribed. 22
3.3 Distributional statistics over the dataset for utterances and tokens. 23
3.4 Distributional statistics over the dataset for applause. 24
3.5 Distributional statistics over the processed dataset. 29
3.6 Two healthcare topics in a 40-topic LDA model with $\alpha = 0.1$ and $\eta = 0.1$. . 31
3.7 Moderator-specific topics in a 20-topic LDA model trained using only moderators' utterances, with $\alpha = 0.01$ and $\eta = 0.01$. 33
3.8 Values of inter-topic coherence over all pairs of topics in a 40-topic LDA model with $\alpha = 0.1$ and $\eta = 0.1$. 37
3.9 Values of cross-party distance over all pairs of topics in a 40-topic LDA model with $\alpha = 0.1$ and $\eta = 0.1$. 40
3.10 Parameter values for LDA grid search over k, α, and η. 43
3.11 Top terms for each topic the final topic model. 44
3.12 Top 20 terms for each party, weighted by $LOR_w^{\mathcal{D}-\mathcal{R}}$. 47
3.13 Most polar terms from progressively improved polarity measurements for the Abortion topic. 50
3.14 Most polar terms from progressively improved polarity measurements for the Immigration topic. 51
3.15 Most polar terms from progressively improved polarity measurements for the Health topic. 52

4.1 Results of the logistic regression tests for each topic and party. 64
4.2 Results of the linear regression tests for each topic and party. 65

DEDICATION

To Leon Roday, who made me wonder
why fat chance and slim chance mean the same thing.

And to Lisa Roday, obviously.

ACKNOWLEDGMENTS

Many thanks to everyone whose help and advice contributed to the success of this document, including Gina-Anne Levow, Sayan Pathak, Laura Panfili, Harrison Roday, Lucy Vanderwende, Philip Resnick, Viet-An Nguyen, Jordan Boyd-Graber, Richard Wright, Gagan Chopra, and many others.

Chapter 1

INTRODUCTION

The United States presidential election of 2016 was widely viewed as a uniquely polarized affair; on election day, *The New York Times* called it "one of the most divisive...campaign seasons in memory" (Robertson, 2016). But while the vitriol of this election may have been distressing to many, it should not, arguably, have been surprising. The 2016 race fits well into a long-standing pattern of increasing polarization in American politics.

In 1994, 64% of adults who identified as Republicans held more conservative views than the median Democrat. Among Democrats, 70% were more liberal than the median Republican (Pew Research Center, 2014). Twenty years later, these numbers had become substantially more extreme. In 2014, 92% of Republicans held views to the right of the median Democrat. Likewise, 94% of Democrats fell to the left of the median Republican. Among politically engaged citizens, the 2014 split was even more extreme: 99% for Republicans and 98% for Democrats.

This growing ideological divide among the American public is starkly reflected in the politicians who represent it. In 2014, a longitudinal study of voting behavior in the U.S. House of Representatives found that party polarization in Congress "has been increasing exponentially for over 60 years" (Andris et al., 2015).[1]

A wealth of previous research in political science has attempted to determine the demographic, cultural, and sociopolitical factors underlying the high degree of polarization in American politics. Regardless of its root, my goal is to understand what role this polarization plays in the way politicians communicate with the electorate, and the way voters respond to

[1]To clarify, this is not an exaggeration. The researchers discovered that Congressional polarization over time actually fits an exponential growth curve to a statistically significant degree.

varying levels of polarity in political speech. Using a corpus of 104 Democratic and Republican presidential primary debates, I present a novel method for measuring voters' reactions to polarized rhetoric at scale. Specifically, I use audience applause as an indicator of positive response and measure the association between these applause events and the polarization of the language preceding them.

It is well-documented that ideological polarization is associated with increased cross-party animosity. Extreme liberals tend to view extreme conservatives negatively *as people*, and vice versa (Pew Research Center, 2014). Similarly, those who are more ideologically polarized are more likely to have an unfavorable view of the opposing party overall. However, it is not my intention to study this polarization of sentiment. Though it would no doubt be interesting, I will focus instead on polarization with respect to *issue-specific language*. Given a particular topic, Republicans and Democrats differ not only in their opinions about that topic, but also in how they frame the discussion of the issue. My aim is to understand how voters react to varying levels of polarization in these topic-specific framing strategies.

I hypothesize that the more polarized a candidate's language is on a given topic, the more likely the audience is to react positively in the form of applause.

In order to test this hypothesis, I first divide the corpus into topics using automatic topic modeling methods. I then present a method for quantifying issue-specific polarization through the lens of lexical framing. Finally, I measure the association between applause and polarity using standard statistical techniques. The results, though preliminary, lend strong support to the hypothesis that applause is associated with polarity.

In the next chapter, I will review previous work in a number of relevant fields – including sociology, political science, and linguistics – in preparation for the analysis.

Chapter 2

BACKGROUND

An analysis of language in politics naturally draws from a number of interconnected subfields of computational linguistics, as well as from related efforts in the burgeoning field of computational social science. In this chapter I will review some of the relevant literature and discuss its applications to the present analysis.

2.1 Rhetoric and Political Speech

The Ancient Greeks, and the Ancient Romans after them, were among the first to systematically study and describe the phenomenon of rhetoric, and their works continue to have a lasting impact. Rhetorical strategies first codified by the Sophists of Ancient Greece some seven centuries ago are still cited in modern analyses of political discourse. Previous efforts to study the relationship between language and applause in political contexts have largely focused on the impact of these rhetorical devices. In a 1986 study of British political party conferences, researchers manually labeled instances of rhetorical strategies from several hours of speeches (Heritage and Greatbatch, 1986). They found that seven common strategies accounted for about two-thirds of the applause in their dataset, although how exactly they arrived at this particular set of devices or how they labeled their uses is not entirely clear.

More recently, several studies have focused on audience reactions to the semantic content of political speech, rather than to its rhetorical structure. Most of these have relied on expensive methods for manually annotating, or directly measuring, audience response. In one recent study, viewers of a 2012 presidential debate were provided with a smartphone app on which they could register their agreement or disagreement with what was being said at any given time (Boydstun et al., 2014). These signals were aggregated to measure the

reactions of various audience cohorts at particular moments in the debate. Other studies have relied on more traditional methods, such as post-debate surveys (Cano-Basave and He, 2016).

Separately from the methods used to measure audience reactions, all of these studies employed various linguistic methods to characterize the semantic content of political speech. These methods largely fall under the umbrella of *framing*, a discipline at the crossroads of psychology and linguistics.

2.2 Framing

The study of framing can be traced to the sociologist Erving Goffman and his 1974 book *Frame analysis: An essay on the organization of experience*. In this sweeping work, Goffman developed the discipline of *frame analysis* as the study of the psychological "schema" through which we organize our experiences. In doing so, he drew on such disparate fields as psychology, evolution, and dramaturgy. The central notion of framing, picked up by later cognitive scientists such as George Lakoff (Lakoff and Johnson, 1980), is the idea that there are fundamental frames of mind through which we perceive the world around us, and that those frames have a strong effect on how we shape our views and interpret our experiences. From a linguistic perspective, framing can be viewed in at least two ways. Speakers' own psychological frames affect the way they choose to word a particular message. Likewise, listeners' frames impact the way they react to a particular wording of that message.

The implications of framing in politics and political speech are powerful (Chong and Druckman, 2007). If a politician wants to convey a message that will resonate with her base, she should express that message in such a way as to engage the psychological frames that she expects will create a strong response with voters. Which framing strategies to use depends heavily on the issue being discussed and the ideology of the target audience. For an example of framing strategies in action, consider the regulation of firearms, a long-standing and contentious issue in U.S. politics. As with any topic on which substantial disagreement exists, there are numerous and partially overlapping viewpoints but, for the purposes of this

example, I will assume there are only two: those in favor of stricter gun regulations and those against. This example is particularly interesting, since both sides can present their case from the perspective of safety. Those in favor of increased regulation argue that more firearms in circulation heightens the risk of deadly violence, whereas those against argue that restricting people's ability to own guns would leave them unable to defend themselves in dangerous situations.

The two sides differ significantly, though, in the framing strategies they use to convey these safety-related arguments. Proponents of gun regulation often make emotional appeals, as former Maryland Governor Martin O'Malley did in this excerpt from a 2016 Democratic presidential primary debate (Peters and Woolley, 2016a):[1]

> I'm the one candidate on this stage that actually brought people together to pass comprehensive gun safety legislation. [$\cdots$] I will never forget one occasion visiting a little boy in Johns' Hopkins Hospital, he was getting a birthday haircut, the age of three when drug dealers turned that barbershop into a shooting gallery and that boy's head was pierced with a bullet. [$\cdots$] I remember visiting him and his mother in Johns Hopkins Hospital.

On the other hand, opponents of regulation make appeals to freedom, individual liberties, and self-protection, as Florida Senator Marco Rubio did in this quote from a 2016 Republican presidential primary debate (Peters and Woolley, 2016b):

> The Second Amendment, as I've said before, is not a suggestion. It is the constitutional right of every American to protect themselves and their families. [$\cdots$] It is right after the defense of the freedom of speech for a reason, for clearly the founders of our nation understood and the framers of the Constitution understood that you cannot have life and you cannot have liberty and cannot pursue happiness if you are not safe.

In these two examples, the candidates are doing more than just stating their views on gun control. Each one is orienting the discussion through a certain lens that he hopes will

[1] Throughout this document, I indicate truncations of debate excerpts with "[$\cdots$]" so as to distinguish them from the actual ellipses that appear in the transcripts.

resonate with the audience. In other words, the candidates are invoking very specific – and very different – frames to talk about the same issue.

Often, a particular framing strategy becomes strongly associated with a certain position on a certain topic. In the United States, for example, politicians are described as being "pro-life" if they are against abortion, or "pro-choice" if they are in favor. Note, crucially, that these phrases are not actually semantic descriptions of the positions they refer to. Rather, they are framing strategies used to present each view in a positive light and – perhaps more importantly – the opposing view in a negative one. After all, who could possibly defend being "anti-life" or "anti-choice" as a good thing?

By extension, when positions are strongly associated with political parties, framing strategies emerge as differentiating factors in each party's discussion of an issue. It is these kinds of polarized associations that I will focus on in this analysis. I intend to automatically discover topic-specific framing strategies at scale, and use the party polarization of these strategies to test the hypothesis that applause is associated with polarity.

Framing as a whole is undoubtedly a complex phenomenon, and even restricting one's focus to the linguistic aspects of framing does little to simplify the situation. Framing strategies can be analyzed at all levels of linguistic representation, from syntax to discourse structure. Performing these kinds of analyses at scale would be immensely complicated. Fortunately, recent work has made substantial progress in analyzing framing on a level that is computationally tractable: the level of individual words.

2.3 Lexical Framing and Computational Political Science

The study of lexical framing is based on the observation that one can often accurately identify people's positions on an issue not through the detection of explicit statements such as, "I am against the death penalty," but by looking at the words they use when they talk about the issue. Returning to the example of abortion, we might expect to be able to identify supporters and opponents by how much they use the words *choice* and *life*, respectively. Recently, this observation has been exploited by computational linguists and political scientists attempting

to automatically detect and characterize political ideologies.

Before discussing these efforts, I should clarify what is meant by "political ideology." In the United States, political views are often modeled as a points on a spectrum between two extremes: liberal on the left and conservative on the right. Another variant replaces political ideology with party affiliation, placing Democrat on the left extreme and Republican on the right. While these may be overly simplistic representations of reality, they have been very successful in the realm of ideology detection, and I will adopt them here.

Lexical framing approaches to ideology detection lie at the intersection of computational linguistics and computational social science, and previous work in this area covers a wide array of research contexts. Some efforts have focused on discovering and describing the polarization of lexical frames through the analysis of political texts. Grimmer (2010), for example, used topic modeling and lexical framing to measure political agenda setting in U.S. Senate press releases. Monroe et al. (2009) discovered polarized lexical frames from Congressional floor debates, then used that information to analyze a number of interesting patterns, including comparing the language of male and female members of Congress, assessing regional variations in debates about public lands, and tracking the party ownership of certain words over time.

Other studies have been more predictive in nature, using lexical framing information to build machine learning models like binary classifiers. In Gerrish and Blei (2011), for example, researchers were able to predict a politician's vote on a given bill from the words that he or she had used in floor debates on the issue. Djemili et al. (2014) used lexical analysis to characterize and predict the ideological alignment of voters on Twitter.

These previous efforts in lexical frame analysis often involved developing novel computational models and measures, some of which I will draw on throughout this study.

2.4 Sentiment Analysis

Another linguistic discipline relevant to any study of political discourse is sentiment analysis. Indeed, ideology detection can in some sense be considered a special case of sentiment

analysis. Broadly, this area of research focuses on automatically discovering opinions in text, and efforts in this area fall into two main categories (Pang and Lee, 2008):

1. *Subjectivity detection* focuses on detecting whether a piece of text contains an opinion relevant to the phenomenon being studied.

2. *Sentiment detection*, given a piece of opinionated text, attempts to identify the opinion being expressed.

Sentiment detection usually focuses on classifying polar opinions. In other words, given two sides of a spectrum – whether it be yes vs. no or left vs. right – the task is to place a given piece of text into one of those two buckets. A traditional binary classification approach is well-suited to this particular framing of the task and, as such, it is very popular in the sentiment detection literature (see Pang and Lee (2008), in which numerous classification-based studies are reviewed). Efforts in this vein generally pattern similarly. They begin by collecting and/or presenting labeled data for the phenomenon of interest, and proceed by describing a supervised machine learning algorithm suited to the task, training it, and using it to classify a held-out test set. Finally, the models are evaluated using accuracy, precision, recall, and related measures.

This approach has several advantages. The most obvious is the plethora of robust and mature binary classification algorithms, many of which are available in efficient, off-the-shelf implementations.[2] Using labeled data means that ground truth is readily accessible, and widely agreed upon evaluation metrics exist for supervised machine learning techniques. In many areas of research, gathering labeled data is often painstaking and expensive, but it can be a less onerous task for political science problems, where texts are often already associated with a label in the form of a particular political party. Indeed, the present analysis is one of them.

---

[2] `scikit-learn` for Python (Pedregosa et al., 2011) is just one example of a software package that includes ready-to-use implementations for a number of classification algorithms. There are many others.

Despite these advantages, however, a classification approach is not suitable for detecting the polarity of topic-specific lexical frames. The primary reason is that I am interested in polarity *quantification* rather than polarity *classification*. If the goal of this analysis were to distinguish between Republican and Democratic utterances, training a binary classifier would be a natural course to pursue. But my goal is not simply to know which bucket a given lexical frame belongs to. In order to test the hypothesis that applause is associated with polarity, I need to know where on the spectrum it lies. And while it is true that most supervised algorithms output a confidence associated with each decision, there is no reason to believe that this confidence is indicative of polarity. Additionally, even if I were to interpret the classifier's confidence as some sort of polarity score, I would still not be able to use standard evaluation techniques to measure its effectiveness. Evaluating the accuracy of a classifier on binary labels is not the same as measuring how well calibrated its confidence scores are.

Nonetheless, the literature on sentiment classification is extensive and, while I cannot adopt its overall strategy, insights from it will be useful on the road ahead. In particular, some popular methods for feature selection will be used as a starting point for arriving at a quantifiable definition of polarity.

2.5 Topic Modeling

Any study of framing requires associating raw text with topics, since framing strategies are inherently topic-specific. Previous studies of framing in politics (such as Iyyer et al. (2014); Nguyen et al. (2015)) have largely relied on texts whose topics have been predetermined. Sometimes these topic labels are inherent to the dataset. Monroe et al. (2009), for instance, studied framing in the context of U.S. Congressional floor debates, which are by definition about a specific bill. Unfortunately, the dataset of presidential primary debates used for the present analysis does not have that advantage. It is certainly true that specific presidential primary debates are often focused on one or a few general topics (e.g. "the economy"). However, beyond the challenge of identifying those debate-specific topics in the first place, it

is by no means a valid assumption that all speech in the debate is confined to those topics.

In other studies, as in Tsur et al. (2015), topic labels were assigned manually by annotators, followed by a second post-processing phase in which the researchers combined related topics to arrive at a final representation. Given the scale of my dataset, I consider this approach to be prohibitive from the perspective of both time and cost.

Aside from the fact that debates are not globally confined to a predefined set of topics, debate speech displays a considerable amount of local variability as well. Candidates may touch on a multitude of topics even within a single response. Consider the following excerpt from a 2016 Republican primary debate, in which Governor Scott Walker of Wisconsin addresses at least five separate topics – education, taxes, job creation, energy policy, and healthcare – in response to a question about the minimum wage (Peters and Woolley, 2015).

> **TAPPER:** Governor Walker, I want to go to you. Dr. Carson wants to raise the Federal Minimum Wage, you have called it a lame idea. Why is raising the Federal Minimum Wage lame?
> **WALKER:** So, the best way to help people see their wages go up is to get them the education, the skill they need [$\cdots$] I've cut income taxes, I've cut property taxes. In fact, property taxes are lower today in my state than they were before we took office. The real issues about jobs. [$\cdots$] All the things we should be talking about tonight are about how do we create jobs, helping people get the skills and the education qualifications they need to succeed. [$\cdots$] It's part of our large plan to reform the tax code, to cut taxes, to put in place an education system that gives people the skills and education that they need. [$\cdots$] To put in place all the above energy policy, but you start on day one with repealing Obamacare. I'm the only one on this stage that's actually got a plan, introduced an actual plan to repeal Obamacare on day one. I'll send a bill up to Congress, and to make sure enact it...[$\cdots$]

This local topic variability is another important feature that distinguishes the present analysis from much of the existing literature on political framing. Often, a given document is assumed to be about a single topic. This single topic assumption has been used successfully in analyses of floor debates and Senate press releases (Thomas et al., 2006; Grimmer, 2010). In these contexts, it is probably valid to assume that the documents under study are about a single topic. However, the assumption does not hold for presidential primary debates, so an alternative representation of a document's topics is needed.

To summarize the constraints just outlined, this analysis requires a topic modeling method that discovers topics automatically, does so at scale, and describes a document with something more nuanced than a single topic assignment. Fortunately, such a method exists, and its virtues have made it a stalwart in topic modeling techniques for some time.

2.5.1 Latent Dirichlet Allocation

Since its introduction in 2003, Latent Dirichlet Allocation (LDA) has become a standard method for automatically generating topics from a set of related documents, largely superceding the earlier Latent Semantic Indexing (Deerwester et al., 1990). LDA is an unsupervised, generative graphical model for discovering topics in text (Blei et al., 2003). It takes as input a set of documents and a parameter k, which specifies the number of topics to be discovered. Formally, LDA assumes that each document in the corpus is generated from a mixture of the k topics. This mixture is described as a discrete probability distribution over the topics for each document. In turn, each topic is represented as a distribution over the V words in the vocabulary. A given word in a given document is assumed to be generated by first drawing a topic from the document's topic distribution, then drawing a word from the topic's word distribution. Given this generative framework, a topic model is inferred from the corpus using standard Bayesian inference techniques (such as collapsed Gibbs sampling (Griffiths and Steyvers, 2004)).

LDA takes two hyperparameters, α and η, non-negative real numbers which govern the shape of the inferred topic model. Technically, each topic z has its own associated α_z and each vocabulary item v has its own η_v, so that the hyperparameters are actually a k-dimensional vector and a V-dimensional vector, respectively. Almost always, however, α is set to be the same for all topics and η is set to be same for all vocabulary items. The first hyperparameter, α, influences the sparsity of the topic distributions for each document. If α is low (e.g. 0.05), LDA will tend to assign fewer topics to any given document. The hyperparameter η governs the sparsity of the word distributions for each topic. If η is low (e.g. 0.005), the model will tend to describe a given topic using fewer words.

In the inference process, LDA attempts to find the most likely set of topics that would have generated the given corpus of documents. The inferred model provides two primary outputs:

a) For each document d, a k-dimensional vector θ_d, which is the distribution over topics in d. Intuitively, this represents the relative prominence of each topic in d.
b) For each topic z, a V-dimensional vector β_z, which is the distribution over words in z. Intuitively, words with high probability in β_z may be considered "signatures" of z.[3]

Despite its widespread use, LDA is not without its limitations. The most readily apparent is that it requires the number of topics k to be specified in advance, even though it may not be known ahead of time how many topics ought to be discovered in a given set of documents. Additionally, LDA assumes no underlying relationships between the topics in the corpus, but instead generates a "flat" set of unrelated topics. Finally, as with any unsupervised method, LDA may produce unexpected results in some situations. In particular, when a desired set of topics is known in advance, it is not guaranteed that the topic distinctions that LDA learns will align with the topic distinctions assumed by the experimenter (Mcauliffe and Blei, 2008). Consider a researcher who wishes to assign each document in a corpus of product reviews to a set of product categories (e.g. *electronics*, *health and beauty*, etc.). LDA might perform admirably on this task, but it might just as easily pick up on words like *great* and *horrible* and discriminate between positive and negative reviews instead. The particular outcome depends on which structure LDA deems as more likely to have generated the given corpus, regardless of the experimenter's intentions.

Since LDA's introduction and subsequent popularity, a number of variants have been created to address these limitations. Some of these variants (e.g. HDP (Teh et al., 2005) and hLDA (Blei et al., 2010)) learn a hierarchical tree structure of topics, with successively lower levels in the tree corresponding to lower levels of conceptual abstraction. Other variants, like

[3]There are some notational inconsistencies in the naming of β. In some later descriptions of LDA, β is used instead of η to refer to the document-topic hyperparameter. In those cases, the matrix of topic-word distributions is generally called ϕ or φ instead of β. In this document, I use the original notation.

sLDA, introduce supervision in the form of response variables, which impose an additional constraint on the model's optimization function (Mcauliffe and Blei, 2008). Many other flavors of LDA exist, and more are being introduced every year. For this analysis, though, I will stick with classic, "vanilla" LDA, both because of the wealth of relevant literature and the widespread availability of performant implementations.

2.5.2 Evaluation of Topic Models

For any study that uses machine-learned models, evaluation of those models is a critical step in the analysis. However, robust evaluation of topic models, as with any unsupervised method, presents a challenge. Since there is no information regarding what the "true" topic model should look like, researchers tend to turn to model-intrinsic properties for evaluation. While these measures are not necessarily informative in and of themselves, they can be useful in comparing multiple models trained on the same set of data. In the context of algorithmic design, they are used to evaluate the strength of a given topic modeling algorithm with respect to others. In the case of topic model application, as in the present analysis, they are often used for model selection. Some of the more common evaluation measures are presented here.

Log Likelihood

One standard way to measure the quality of a topic model is to calculate its log likelihood (Blei et al., 2003; Wallach et al., 2009). The likelihood of a model with respect to a set of documents is defined as the probability that the model generated that set of documents. Since, for any large corpus of documents, the likelihood of any model generating that *particular* corpus is very small, these values are generally presented in logged form. To get a stronger notion of how well the model describes the data, this evaluation is often performed on a held-out test set of documents.

If the log likelihood of one model is higher than that of a second, it might be considered a more accurate description of the data, since it was more likely to have generated the data

that were actually observed. As a result, one would prefer the first model to the second. However, it is not always the case that choosing the model with the highest likelihood is best. Consider the example of product reviews presented earlier. Given two topic models, one which captures product categories as topics (as intended) and one which captures review quality as topics, it may well be that the latter has a higher log likelihood than the former. Nevertheless, it would not be preferred for that particular application.

More generally, previous research has shown that log likelihood, even on a held-out test set, does not necessarily correlate with human judgments of topic interpretability (Chang et al., 2009). Another drawback of log likelihood is that its precise computation is not tractable (Wallach et al., 2009). In practice, sampling methods are used to arrive at a reasonable estimation of the true value.

Perplexity

Another common evaluation measure is perplexity. It is closely related to log likelihood, and is included here only for completeness. Perplexity has been used extensively in the context of language modeling, where it is used to compare multiple language models over the same corpus (Bengio et al., 2003). Perplexity can be defined as a function of log likelihood:

$$\text{perplexity}(D) = \exp\left(-\frac{\text{log-likelihood}(D)}{\sum_{d \in D} N_d}\right)$$

Here, exp is the exponential function, D is a set of documents, and N_d is the total number of tokens in document d.

As expected from the equation above, low perplexity is considered preferable to high perplexity in the evaluation of topic models. However, since perplexity is a function of log-likelihood, it suffers from the same drawbacks – namely that its true value is computationally intractable and that it is not always correlated with human judgments.

Given the limitations of measures like log likelihood and perplexity, recent research in topic model evaluation has focused on alternative methods that evaluate topics based on semantic

coherence rather than predictive likelihood.

Topic Coherence

Work on automatically measuring topic coherence largely stems from the discovery, noted above, that traditional measures of predictive likelihood do not always align with human judgments of topic interpretability. Indeed, the first study to use human judgments to evaluate topic coherence found that traditional probabilistic measures were sometimes *negatively* correlated with annotators' judgments of semantic interpretability (Chang et al., 2009).

In that study, human judges were asked to perform two tasks to measure the coherence of individual topics and the appropriateness of a model's document-topic assignments. In the first task – called "word intrusion" – judges were shown the five most probable words from a given topic along with an "intruder word," which was randomly chosen from a set of words with low probability in that topic. They were then asked to identify which word in the set did not belong. The quality of a topic was then defined as the proportion of individual judgments in which the intruder word was correctly identified. The procedure for the second task, which the researchers called "topic intrusion," followed in a similar vein, but using documents and topics rather than topics and words. The researchers found that an increase in predictive likelihood did not correlate with better scores on either the word intrusion or topic intrusion tasks.

Since its publication, Chang et al. (2009) has spurred a number of efforts to find computational proxies for human-judged coherence. One approach that has seen considerable success involves evaluating topics against some external reference corpus that is assumed to be topically organized. Newman et al. (2010), for example, developed a variety of word pair similarity metrics that leveraged information from Google, Wikipedia, and WordNet.[4] To evaluate the coherence of a topic, they aggregated the scores for each pair of words in the

[4]WordNet is a lexical ontology in which edges represent different semantic relationships between words (e.g. hypernym/hyponym, synonymy, and metonymy) (Fellbaum, 1998). It is available online at `http://wordnet.princeton.edu/`

ten most probable words for that topic.

In a follow-up to this study, researchers attempted to automate the word intrusion task presented in Chang et al. (2009) by training a Support Vector Machine model to automatically learn intruder words, using the previously developed similarity measures as features in the model (Lau et al., 2014). They found that the model's accuracy in identifying intruder words was highly correlated with the human-labeled measures of topic coherence used in Chang et al. (2009).

In contrast to approaches that rely on external data, some have attempted to evaluate models using only the data upon which the models were trained. Like external methods, though, they have generally involved calculating word-level similarity scores and aggregating them in various ways. Mimno et al. (2011), for example, calculated word similarity using a corpus-internal document co-occurrence measure, which was found to perform competitively against the external coherence measures outlined above.

The success of both external and internal coherence measures lends credence to the idea that aggregating word-level metrics can be effective for evaluating the semantic coherence of topics. Internal methods in particular demonstrate that human-like accuracy can be achieved without relying on an external reference corpus. These findings will be useful in the selection of a final topic model later in this analysis.

In this chapter, I have reviewed previous efforts from a number of disciplines relevant to this analysis, including framing, computational political science, sentiment analysis, and topic modeling. In the next chapter, I will describe the methods used to carry out my analysis, many of which build on work discussed here.

Chapter 3

METHODS

In this chapter, I will develop the analytic framework necessary to test my hypothesis that applause is associated with polarity. I will begin by introducing and describing the dataset, as well as the steps taken to prepare the data for use in a computational setting. Then, I will proceed to the analysis itself, which falls into three major phases. In the first phase, I will use Latent Dirichlet Allocation to automatically generate a topic model over the corpus. Drawing on previous work in automatic topic coherence evaluation, I will develop a novel method for penalizing models containing party-specific topics. After arriving at a final model, I will present a technique to automatically discover the party polarity of topic-specific lexical frames, inspired by the computational political science literature as well as previous analyses of sentiment and lexical framing. Finally, I will describe two methods for measuring applause in the corpus and, using standard statistical techniques, assess the association between the polarity of debate speech and audience applause.

3.1 Dataset

The dataset for this analysis consists of nearly all U.S. presidential primary debates from both major parties from the 2000, 2004, 2008, 2012, and 2016 election cycles. In this section, I outline the dataset acquisition and preparation process, and give some high-level characterizations of the data itself.

3.1.1 Data Acquisition and Preprocessing

The raw text of the debates comes from the American Presidency Project (APP), housed at the University of California, Santa Barbara.[1] The APP generously makes available a variety of documents related to presidential elections, including transcripts of nearly every presidential primary debate since the 2000 election cycle.[2] However, these transcripts are not designed for use in a computational setting. They are not cleaned or tagged, nor are they presented in a machine-readable format. Rather, they are HTML-formatted news transcripts, and they display a high degree of ambiguity and variability in their formatting. The following excerpts show some examples of this variability with respect to speaker identification, applause annotation, and handling of simultaneous speech.

Usually, the speaker of an utterance is identified by last name at the beginning of a paragraph. Sometimes, however, full names and even titles or associated organizations are included (Peters and Woolley, 2004):

> **GRIFFITH:** Thank you.
> (APPLAUSE)
> **HUME:** John, you're next.
> **JOHN DISTASO, UNION LEADER:** My questions are for Senator Edwards and Reverend Sharpton. Senator Edwards, after voting to authorize [$\cdots$] they inconsistent? How are they consistent?
> **SEN. JOHN EDWARDS, (D-NC) PRESIDENTIAL CANDIDATE:** Because I said from the very beginning, before the first resolution was ever voted on in the Congress, [$\cdots$]

Note that this excerpt also includes an applause event, which is annotated in parentheses.

A common occurrence in presidential primary debates is simultaneous speech. Sometimes, the transcripts indicate this using interleaving utterances, as in the following excerpt (Peters and Woolley, 2000):

> **Gore:** [$\cdots$] I helped to put in place a program called the Partnership for a New Generation of Vehicles, which commits the big three automakers in our country to

[1] `http://www.presidency.ucsb.edu/`.

[2] Excluded are one Democratic debate and two Republican debates from 2000, for which no transcripts were available.

> getting new vehicles into the marketplace that have three times the efficiency...
> **Shaw:** Time.
> **Gore:** ...of today's vehicles. That's part of the answer. [· · ·]

Other times, when too many people are talking, the "crosstalk" annotation is used, as in the following excerpt (Peters and Woolley, 2016c). Note that this excerpt uses square brackets for annotations, as opposed to the parentheses used above. Across all of the transcripts, annotations are variably indicated using parentheses, square brackets, dashes, ellipses, or a combination of these.

> **CRUZ:** Well, listen, I've spent my entire life defending the Constitution before the U.S. Supreme Court. And I'll tell you, I'm not going to be taking legal advice from Donald Trump.
> **TRUMP:** You don't have to. Take it from Lawrence Tribe. [*applause*] [*crosstalk*] Take it from your professors...[*crosstalk*]
> **CRUZ:** The chances of any litigation proceeding and succeeding on this are zero. And Mr. Trump [· · ·]

In light of these and other complications, I carried out the following preprocessing steps to prepare the data for use in an NLP setting:[3]

1. A list of URLs for the transcript of each debate was obtained from the APP's "Presidential Debates" homepage,[4] along with the associated party for each debate.
2. The URLs were automatically crawled to extract the raw HTML of each debate transcript.
3. Most transcripts included a header section from which the list of candidates and moderators could be extracted automatically. For those that did not, the full set of unique speakers was extracted and then manually separated into candidates and moderators.
4. Using a custom parser, each transcript's HTML was parsed into an ordered list of *events*. Events fell into one of two categories:

---

[3]For more details, source code is available at `https://github.com/ethanroday/three-cheers`.

[4]`http://www.presidency.ucsb.edu/debates.php`, accessed 2016/08/23.

(a) *Utterance events*: An utterance event occurred whenever a candidate, moderator, or other debate participant was speaking. The parser automatically identified the speaker for a given utterance using a combination of predictable HTML structure and regular expression heuristics.

(b) *Non-utterance events*: These events captured anything other than utterances, including audience-related events such as laughter, booing, and – most importantly – applause. Non-utterance events were often inconsistently presented in the raw transcripts, but they could be detected with reasonable accuracy using regular expression heuristics.

5. Multi-sentence utterance events were split into individual sentences, with each sentence constituting its own event.
6. The resulting list of utterance and non-utterance events was output in a portable JSON format for use in the analysis.

Limitations of Textual Applause Annotations

As shown in example excerpts above, applause is annotated in a variety of ways in the transcripts. But, aside from the differences in formatting, there is only one *kind* of applause. That is, a textual transcript does not indicate whether a given applause occurrence corresponds to polite clapping or a standing ovation; it only indicates whether applause has occurred or has not. Since my goal is to use applause as a measure of the audience's reaction, the enthusiasm of the applause at any given time would be very useful information. Without it, the granularity of my analysis will certainly be restricted.

There is also substantial variation in exactly *where* applause appears in the transcripts. For some debates, applause is only transcribed between utterances. Other times, it appears in the middle of utterances, with the text on either side of it interrupted by dashes or ellipses. A further complication arises from the fact that non-utterance events – like applause – are generally transcribed after the last utterance that started *immediately* before the non-

Table 3.1: Number of debates by political party and election year.

		Election Year					
		2000	2004	2008	2012	2016	*Total*
Party	Democrats	8	2	19	0	9	*38*
	Republicans	11	0	16	20	19	*66*
	Total	*19*	*2*	*35*	*20*	*28*	*104*

utterance event began. This means that, when a particular candidate utterance generates applause, but a moderator attempts to move on before that applause begins, the applause is transcribed after the *moderator's* utterance, not after the utterance that caused it. I will return to this problem and propose a simple solution in Section 3.4.1.

Finally, some debate transcripts are missing applause annotations entirely. I do not interpret this as an indication that applause did not occur, but rather that it was not transcribed when it did occur. This is based on a manual review of the applause-less debates. In nearly every one, the moderator gives an explicit prompt for applause (usually at the beginning and/or end of the debate), but no applause event is transcribed afterwards. These debates will be excluded in all applause-related portions of the analysis.

3.1.2 Composition of the Dataset

There are 104 debates in the full dataset. Table 3.1 shows the distribution of debates by party and election cycle. Several things are worth noting in the overall distribution of the data.

First, there are some party-year combinations for which there is no data. In these years, one party was running an incumbent president, and so no primary nomination process was held. This makes it difficult to consider this dataset for a party-level analysis over time.

A considerable amount of skew is immediately evident in the distribution of debates.

Table 3.2: Number of debates by political party and election year after removing debates with no applause transcribed.

		Election Year					
		2000	2004	2008	2012	2016	*Total*
Party	Democrats	4	2	17	0	9	*32*
	Republicans	7	0	13	20	19	*59*
	Total	*11*	*2*	*30*	*20*	*28*	*91*

With respect to time, the data are heavily skewed towards more recent election cycles. Furthermore, along the party dimension, there are almost twice as many Republican debates as Democratic ones. These skews may stem from the fact that the Republican presidential field has been consistently more crowded than the Democratic field in recent history, and especially so in 2012 and 2016. Both dimensions of skew become more pronounced after removing the seven debates for which applause was not transcribed. All seven come from the 2000 election cycle, and together they account for half of Democrats' and slightly more than a third of Republicans' debates from that year. Table 3.2 shows the adjusted distribution by party and year with non-applause debates removed.

Data imbalance is most often discussed in the context of supervised classification, since skewed datasets can present challenges to standard supervised learning methods (Japkowicz and Stephen, 2002). However, simply because I am not taking a supervised approach in this analysis does not mean that this imbalance can be ignored. I will return to it, and present a strategy to address it, in Section 3.2.2.

3.1.3 Distributional Statistics Over the Parsed Dataset

Before continuing, I will present some distributional statistics over the dataset to help orient the discussion. Statistics over utterances and tokens are summarized in Table 3.3, along with

Table 3.3: Distributional statistics over the dataset for utterances and tokens, split in each case by party.

	Utterance-Level Statistics					Token-Level Statistics				
		Utt/Deb		Utt_{cand}/Deb			Tok/Deb		Tok/Utt	
	n	*avg*	*std*	*avg*	*std*	n	*avg*	*std*	*avg*	*std*
Dem	38,425	1,011.2	274.7	72.4%	5.15%	610,407	16,063.3	3,764.2	15.886	12.849
Rep	78,052	1,200.8	359.9	74.1%	5.76%	1,150,338	17,697.5	4,499.7	14.738	11.627
Total	*116,477*	*1,130.8*	*342.2*	*73.5%*	*5.57%*	*1,760,745*	*17,094.6*	*4,297.9*	*15.117*	*12.055*

means and standard deviations where appropriate.

- Utt/Deb is the number of single-sentence utterances per debate.
- Utt_{cand}/Deb is the proportion of utterances spoken by candidates per debate. Non-candidate utterances include all speech from moderators and panelists, as well as from audience members, pre-recorded messages, and other sources.
- Tok/Deb is the number of space-separated, non-punctuation tokens per debate.
- Tok/Utt is the number of tokens per utterance.

On average, Republican debates appear to be longer than Democratic debates by about 200 utterances and 1,600 tokens. They are also more variable in their length. This is likely a reflection of the large number of Republican candidates in the 2012 and 2016 cycles. The proportion of candidate utterances per debate is very consistent between parties, both in average proportion and in per-debate variability. The lengths of utterances are also fairly consistent between parties, with Democrats on average having slightly longer and slightly more variable utterances.

The statistics in Table 3.4 describe the distribution of applause in the dataset. Again, means and standard deviations are provided where appropriate. The applause statistics have the following definitions:

Table 3.4: Distributional statistics over the dataset for applause, split in each case by party.

	Applause-Level Statistics						
		Appl/Deb		*Utt/Appl*		*Tok/Appl*	
	n	*avg*	*std*	*avg*	*std*	*avg*	*std*
Dem	1,999	62.47	47.30	15.10	31.51	239.2	530.8
Rep	3,412	58.83	38.78	18.65	27.95	271.1	404.87
Total	*5,411*	*60.12*	*41.78*	*17.34*	*29.36*	*259.3*	*445.7*

- *Appl/Deb*, where applause is transcribed, is the number of applauses per debate.
- *Utt/Appl*, where applause is transcribed, is the distance in utterances between two applause occurrences.
- *Tok/Appl*, where applause is transcribed, is the distance in tokens between two applause occurrences.

Taken together, the applause behaviors of Democratic and Republican audiences are fairly consistent. On average, there are about 60 applauses per debate, with about 17 utterances in between any two given applause events. Overall, Democrats appear to applaud slightly more often than Republicans, and with somewhat higher variability.

Having presented some details about the makeup and distribution of the dataset, I will now move on to the first phase of the analysis: topic modeling.

3.2 Phase I: Topic Modeling

In this section, I will review the steps taken to prepare the corpus for topic modeling, describe the model selection process along with some application-specific considerations, and present the final topic model used in the remainder of the analysis.

3.2.1 Defining Documents

LDA takes its input in the form of a set of documents. More formally, it takes an $M \times V$ document-term frequency matrix, where M is the number of documents and V is the size of the vocabulary. Often in topic modeling tasks, there is an intuitive unit of text that should be considered a document. In modeling a set of blog posts, for example, each post is its own document. Likewise, in an analysis of Senate press releases, each press release is a document. In the context of semi-structured debate, however, the definition of a document is less clear.

Certainly, an entire debate cannot be considered one document. LDA is a bag-of-words model, so it fundamentally depends on document-level collocation information in inferring a topic model. Using an entire debate as a document would severely hinder LDA's ability to extract such information. By the same token, single sentences as documents would be too small, since one sentence is too sparse to provide helpful collocation information. Somewhere in between, then, lies the optimal document size. One option would be to create a fixed window of l sentences and then, using either a sliding or non-overlapping window, treat each window as a document. Unfortunately, both options are fraught with difficulties. The overlapping window would artificially repeat most of the dataset, which could have unexpected effects. Additionally, the size of the dataset would increase quadratically with the size of the window.[5] The non-overlapping window would be no less troublesome. In this case, the particular set of documents created would be essentially arbitrary. While it's possible that, on the whole, documents might be somewhat coherent, it's just as possible that document cutoffs would occur in extremely unnatural places, perhaps linking together two completely unrelated segments of text or splitting a very topically coherent sequence of sentences into multiple documents. Both window-based approaches also suffer from the fact that the correct choice of l is not immediately evident.

Fortunately, the inherent structure of debates suggests a strategy for heuristically creating a reasonable set of documents from the text of each debate. In general, a debate

[5]Consider a corpus of n sentences, each with m words. A sliding window of length l would result in $m(l^2 - ln + l)$ words in total.

consists of questions from a moderator (or set of moderators), responses from candidates, and perhaps additional discussion among candidates. Conceptually, a debate can be divided into "question segments," each of which begins with a question posed by a moderator and ends just before the next question, after one or more candidates have given their responses to the question. The moderator utterances that open a question segment can be considered one document, and each candidate's response to the question can be its own document as well. This suggests a very simple algorithm for debate segmentation: every time the speaker changes, start a new document.

In reality, of course, debates are not this well-behaved. They are full of interruptions, acknowledgements, requests for clarification, and other short utterances that do not merit the start of a new document. So I impose an utterance cutoff threshold c, which designates the minimum number of contiguous words needed from a single speaker to start a new document. Based on qualitative assessments, I choose $c = 15$. Another practical consideration is that questions are often asked by people besides the official moderators, particularly in town hall settings. In these cases, I use the tag `UNKNOWN` to identify the speaker. Finally, most debates begin with an introduction of the candidates, which usually includes a number of applause events. For each debate ∂, I manually identified an utterance index h^{∂}, representing the index of the first utterance after this introductory section. The pseudocode for the complete document generation algorithm is given in Algorithm 1.

3.2.2 Document Preprocessing and LDA Implementation

After segmenting the debates into documents, but before actually training a topic model, I preprocessed the corpus using a number of standard techniques.

1. All words were lowercased.
2. Common stop words were removed using a stop list. For details, see Appendix A.
3. In addition to common stop words, I removed the names of any debate participants (moderators included), along with the names of offices (e.g. *secretary*) that might

Algorithm 1: Generating documents from debates

Data:

A set of $\mathbb{M}$ debates, with each debate $\partial = 1 \dots \mathbb{M}$ represented using the following:

A list of L^∂ single-sentence utterance events, where each utterance $i = 1 \dots L^\partial$ consists of:

U_i^∂, a V-dimensional frequency vector for the i^{th} utterance in ∂

S_i^∂, the speaker of the i^{th} utterance in ∂ (or `UNKNOWN`)

h^∂, an index indicating the first utterance in ∂ after the introduction

c, an utterance cutoff threshold

Result: an $M \times V$ document-frequency matrix

$C \leftarrow$ an empty matrix with V columns

for $\partial \leftarrow 1$ **to** $\mathbb{M}$ **do**

  $endOfLastTurn \leftarrow h^\partial$

  $curDocStart \leftarrow h^\partial$

  $curDocLength \leftarrow 0$

  for $i \leftarrow h^\partial$ **to** L^∂ **do**

    if $S_i^\partial \neq S_{i-1}^\partial$ **then**

      $curDocStart \leftarrow i$

      $curDocLength \leftarrow 0$

    $curDocLength \leftarrow curDocLength + \sum_{w=1}^{V} U_{i,w}^\partial$

    if *(curDocStart > endOfLastTurn) and (curDocLength > c)* **then**

      $doc \leftarrow$ a V-dimensional vector of zeros

      for $j \leftarrow endOfLastTurn$ **to** $curDocStart$ **do**

        $doc \leftarrow doc + U_j^\partial$

      add doc as a new row in C

      $endOfLastTurn \leftarrow curDocStart$

return C

identify them. Including the debate participants introduces a significant amount of noise in both topic modeling and polarity measurement, since names – especially the names of minor candidates such as *huckabee* and *o'malley* – are highly discriminative with respect to party. The full list of names and offices can be found in Appendix A.

4. The vocabulary was expanded to include bigrams as well as unigrams. This was based on the intuitive assumption that many notable lexical framing strategies are multi-word (e.g. *partial_birth*, *death_tax*, and *pork_barrel*).
5. Terms[6] that appeared fewer than five times in the entire corpus were removed. This shrank the search space considerably, mostly by removing rare bigrams.
6. Terms that appeared in more than 95% of documents were removed, in order to catch other domain-specific stop words.
7. As a coarse method for addressing data skew, documents from Democratic debates were randomly resampled according to the ratio of Democratic utterances to Republican utterances in the parsed dataset.

The document-frequency matrix resulting from these preprocessing steps was the one actually used as input to the topic modeling algorithm. I used vanilla LDA as implemented in the `lda` package for Python.[7] This implementation uses collapsed Gibbs sampling for inference. All models were trained for 1000 iterations unless otherwise noted.

3.2.3 Distributional Statistics Over the Processed Dataset

I will briefly provide a few distributional statistics to characterize the processed dataset C. Table 3.5 shows the values of these statistics, along with means and standard deviations.

- $Terms/Doc$ is the number of terms per document.
- Utt/Doc is the number of utterances per document.

[6]Throughout the rest of this section, I will use *terms* interchangeably with *unigrams and bigrams* for brevity.

[7]Available at `http://pythonhosted.org/lda/api.html`.

Table 3.5: Distributional statistics over the processed dataset, split in each case by party.

	Document-Level Statistics					
		Terms/Doc		*Utt/Doc*		*Appl/Doc*
	n	*avg*	*std*	*avg*	*std*	*avg*
Dem	7,916	57.45	50.00	6.84	5.06	25.48%
Rep	10,228	54.09	45.39	7.30	5.56	24.26%
Total	*18,144*	*55.56*	*47.48*	*7.10*	*5.35*	*24.80%*

- *Appl/Doc* is the proportion of documents that contain at least one applause event.

On average, a document consists of about 56 terms. Note that this does not necessarily indicate the *length* of the document, but the number of terms in the document after the removal of stop words, the expansion of unigrams to bigrams, and all other preprocessing steps. A more intuitive notion of the length of a document is *Utt/Doc*. This statistic shows that the average document is made up of about seven utterances from the original dataset. Finally, about one quarter of all documents contain applause. All three statistics display a high degree of consistency between parties.

3.2.4 Notation for the Processed Dataset

Throughout the rest of this chapter, it will be useful to have a formal representation of the dataset, if only for brevity and consistency in upcoming formulas and equations. I will use the following notation for the various parts of the processed dataset:

- C is the corpus of M documents.
- Each document $C_{d=1...M}$ is a term-frequency vector of length V.
- Each entry $C_{d=1...M,w=1...V}$ is the number of times the term w occurred in document d.
- θ is the $M \times k$ matrix of document-topic distributions from LDA, as described in the previous chapter.

- β is the $k \times V$ matrix of topic-term distributions from LDA, as described in the previous chapter.
- P is an M-dimensional vector where $P_{d=1...M}$ is the party associated with document d. Each P_d is one of $\{dem, rep\}$.

For consistency, I will also generally use the same symbols to denote random variables of the same type in various probabilistic contexts:

- W will always be a random variable for a term, taking on a particular value w.
- D will always be a random variable for a document, taking on a particular value d.
- Z will always be a random variable for a topic, taking on a particular value z.
- $\mathcal{P}$ will always be a random variable for a party, taking on a particular value ρ. Since there are only two parties, it will also be convenient to have $\mathcal{D}$ and $\mathcal{R}$, denoting the events that $\mathcal{P} = dem$ and $\mathcal{P} = rep$, respectively.

Formally, the precise definitions of these variables will differ slightly in various places, but the meaning will always be clear from context. Finally, I will use $\mathbb{P}$ to denote the probability of an event (and $\hat{\mathbb{P}}$ for estimated probabilities), so as to distinguish it from party-related symbols.

3.2.5 Model Selection: The Healthcare Problem

The rest of this section will be devoted to model selection and, at its conclusion, I will have chosen a final topic model for the rest of the analysis. Given that this model will be used to discuss framing with respect to semantically meaningful topics, rather than to predict the topic distribution of unseen documents, it makes sense to turn to measures of topic coherence for evaluation. However, there is a further complication specific to this dataset that standard coherence measures do not capture. It concerns the party-specificity of the topics learned by LDA. To illustrate the issue, I trained a 40-topic model with $\alpha = 0.1$ and $\eta = 0.1$. Table 3.6 shows the most probable terms in each topic for a subset of the topics in the resulting model.

Table 3.6: Two healthcare topics in a 40-topic LDA model with $\alpha = 0.1$ and $\eta = 0.1$.

Index	Most Probable Terms	$prom_{\mathcal{D}}$	$prom_{\mathcal{R}}$
28	*government*, *federal*, *state*, *federal_government*, *states*, *way*...	0.1630	0.8370
10	*spending*, *cut*, *budget*, *billion*, *money*, *debt*, *dollars*, *taxes*...	0.1684	0.8316
⋮	⋮	⋮	⋮
14	*party*, *republican*, *reagan*, *win*, *ronald*, *ronald_reagan*, *election*...	0.3068	0.6932
15	*court*, *rights*, *supreme*, *right*, *supreme_court*, *marriage*...	0.3640	0.6360
35	***health***, ***care***, ***insurance***, ***people***, ***plan***, ***health_care***, ***obamacare***...	**0.3705**	**0.6295**
29	*let*, *just*, *want*, *say*, *going*, *talk*, *respond*, *right*, *let_just*, *ll*...	0.4031	0.5969
⋮	⋮	⋮	⋮
25	*class*, *middle*, *people*, *middle_class*, *country*, *college*, *families*...	0.7961	0.2039
17	***care***, ***health***, ***health_care***, ***insurance***, ***plan***, ***people***, ***universal***...	**0.8216**	**0.1784**

In addition, it shows the percentage of documents in which each topic was the most probable topic, split by party. That is, $prom_{\mathcal{D}}$ for topic 28 indicates that 16.3% of documents in which topic 28 was the most probable topic came from Democratic debates. For the purposes of this illustration, consider $prom_{\mathcal{D}}$ and $prom_{\mathcal{R}}$ to be coarse measures of the overall prominence of each party in each topic. The topics are sorted in the table by $prom_{\mathcal{D}}$ ascending.

An examination of topics 35 and 17 reveals where the problem lies. From the perspective of topic-level coherence, both topics are reasonably interpretable; they both seem to be related to healthcare. In general, ending up with two related topics is not a problem in and of itself, but it is not exactly useful in a study of topic-specific lexical framing and party polarization. Consider the values of $prom_{\mathcal{D}}$ and $prom_{\mathcal{R}}$ in each topic. For topic 35, $prom_{\mathcal{R}}$ is quite high, giving the impression that Republicans "own" this topic. Conversely, $prom_{\mathcal{D}}$ is at its highest in topic 17. LDA has learned a separate healthcare topic for each party. Why is this? A closer look at topic 35 shows that *obamacare* – a term used disparagingly by Republicans to describe President Obama's Affordable Care Act – is the seventh most

probable term. It seems, then, that LDA has picked up on some of healthcare's topic-specific lexical frames – the very phenomena that I am hoping to analyze – and split them into two topics![8]

How can this problem be addressed? One approach might be to train a topic model only over the moderators' questions, under the assumption that the language of the moderators tends to be less partisan than the language of the candidates. One could then fit the candidates' speech to this model. However, there are several issues with this approach. First, each debate is still specific to a particular party, and a moderator's framing of an issue may still assume a tone that reflects the party's position. Second, and perhaps more importantly, merely changing the dataset will not provide an escape from LDA's blindness – it will just be blind in a different way. Table 3.7 shows a subset of the topics from a 20-topic model trained only on moderators' utterances. More precisely, the training set included only those documents in which at least 80% of tokens were spoken by non-candidates. Looking at the results, it is immediately clear that LDA has picked up on several moderator-specific topics, including topics about debate rules, follow-up questions, and broadcast-related logistics. In fact, almost half of the topics learned – the eight topics displayed in the table – are related to moderating and debate procedure.

If I were to carry this particular model forward through the rest of the analysis, I'd have to throw away eight of the 20 topics when measuring the relationship between polarity and applause behavior. Losing a considerable amount of analytic opportunity, though, is not the only drawback of this approach. As a consequence of LDA learning so many moderator-specific topics, the remaining non-moderator-specific topics are much coarser than they otherwise would be, reducing the strength of any topic-specific conclusions about framing. In the next section, I will turn to a more systematic method to address this problem that does not involve losing any data.

---

[8]This observation suggests that topic modeling algorithms like LDA may be of use in the actual discovery of lexical framing strategies, though I am hardly the first to come to this conclusion. See Nguyen et al. (2013) for a description of a supervised, hierarchical variant of LDA designed for a similar purpose.

Table 3.7: Moderator-specific topics in a 20-topic LDA model trained using only moderators' utterances, with $\alpha = 0.01$ and $\eta = 0.01$.

Index	Most Probable Terms
0	*want*, *let*, *question*, *ask*, *just*, *right*, *thank*, *going*, *respond*, *ahead*...
⋮	⋮
4	*going*, *ve*, *got*, *right*, *let*, *ll*, *question*, *come*, *talk*, *ve_got*...
⋮	⋮
7	*mr*, *thank*, *know*, *border*, *said*, *does*, *say*, *let*, *chief*, *thank_mr*...
8	*ve*, *want*, *like*, *people*, *say*, *lot*, *going*, *think*, *said*, *know*...
⋮	⋮
12	*said*, *quote*, *just*, *ve*, *think*, *believe*, *good*, *campaign*, *did*, *saying*...
13	*republican*, *party*, *democrats*, *republicans*, *nominee*, *election*...
⋮	⋮
15	*debate*, *candidates*, *thank*, *presidential*, *tonight*, *right*, *questions*...
⋮	⋮
17	*question*, *seconds*, *30*, *30_seconds*, *ll*, *answer*, *thank*, *time*, *going*...
⋮	⋮

3.2.6 Model Selection: Detecting Bad Topic Pairs

Instead of trying to find some subsample of the data for which LDA avoids learning party-specific topics, a more rigorous approach is warranted. Ideally, this approach would be able to automatically detect bad topic pairs like the two healthcare topics above and penalize models which contain them. This penalty could then be used as a measure for model selection. In this subsection, I develop such a method based on pairwise scores between topics. I begin by presenting two pairwise word similarity measures from the topic coherence literature. I then combine these techniques to define a new measure of word similarity that is symmetric, normalized, and does not rely on external data. I use this measure to define the *inter-topic coherence* for a given topic pair. I also define *cross-party distance* to capture the party one-sidedness of a topic pair. Finally, I combine these two methods, resulting in a per-model quantification of the "healthcare problem" presented earlier.

NPMI and Document-Based Topic Coherence

In Section 2.5.2, I provided an overview of the general methodologies used in automatic evaluation of topic coherence. Here, I will present the details of two particular techniques, which I will then extend for use in detecting bad topic pairs.

As discussed in Section 2.5.2, most topic coherence measures are based on pairwise word similarity scores. Given such a score, the coherence measure for a topic is defined as some aggregation (usually the sum or mean) of the scores between all pairs of words in the top n most probable words for that topic. One such similarity score that has been shown to be particularly effective for topic coherence, and which has a number of desirable properties, is Normalized Pointwise Mutual Information (Newman et al., 2010). NPMI was first introduced in 2009, and is a scaled version of Pointwise Mutual Information, a classic measure of similarity in a number of different fields (Bouma, 2009). Given two outcomes x and y from corresponding random variables, standard PMI can be defined as follows:

$$pmi(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \tag{3.1}$$

If x and y are independent, then $pmi(x, y)$ is 0. A negative value of $pmi(x, y)$ indicates that x and y are negatively associated, and vice versa for a positive value. However, these values are highly asymmetric. If x and y never appear together, then $pmi(x, y) = -\infty$. On the other hand, if x and y are perfectly associated, then $pmi(x, y)$ is the smaller of $-\log\ p(x)$ and $-\log\ p(y)$, which is the maximum value that $pmi(x, y)$ can attain. This asymmetric behavior means that PMI can sometimes be a noisy measure, especially for low-frequency outcomes.

NPMI was created to address this problem. It is a simple normalization of PMI, and can be defined as such:

$$npmi(x, y) = \log \frac{p(x, y)}{p(x)p(y)} \Big/ -\log\ p(x, y) \tag{3.2}$$

NPMI's lower bound is -1, which occurs when x and y are mutually exclusive. Its highest value, when x and y always co-occur, is 1. Like PMI, its value is 0 if x and y are independent.

Practically, this normalization has the effect of controlling for the noise generated by low-frequency words in PMI. It also means that the average value of NPMI over many pairs of outcomes has a more intuitive interpretation.

Within the context of topic coherence, the outcomes x and y are terms in the vocabulary, and the joint distribution $p(x, y)$ is the probability that x and y co-occur in some context. The particular definition of co-occurrence varies significantly. In Newman et al. (2010), for example, term co-occurrence was calculated using a ten-token sliding window across the entirety of English Wikipedia.

Resorting to external corpora, however, is not always necessary. Mimno et al. (2011), for instance, used the following as their definition of similarity:[9]

$$docsim(x, y) = \log \frac{D(x, y) + 1}{D(y)} \tag{3.3}$$

Here, $D(x, y)$ is simply the number of documents in which x and y both appear at least once. Likewise, $D(y)$ is the number of documents in which y appears at least once. In many ways, *docsim* is reminiscent of PMI, as with many approaches to word similarity for topic coherence. Research has shown *docsim* to be extremely effective in evaluating topic coherence (Stevens et al., 2012), and it is this outcome that is most relevant to the present analysis. Namely, that corpus-internal, count-based definitions of co-occurrence can be useful in defining word similarity scores. In the next section, I will take advantage of this finding in developing a measure for inter-topic coherence.

Inter-Topic Coherence

Based on the success of *docsim*, I will propose a method for estimating $npmi(x, y)$ based on simple, corpus-internal measures, and will then use that estimation in defining inter-topic coherence. First, let $D(x, y)$, $D(x)$, and $D(y)$ be defined as above, and let M the total

[9]This was called C in the original paper; I have renamed it here to avoid confusion with the corpus C.

number of documents in the corpus. Then:

$$\hat{npmi}(x,y) = \log\frac{D(x,y)/M}{(D(x)/M)(D(y)/M)}\Big/-\log\frac{D(x,y)}{M} = \log\frac{D(x,y)M}{D(x)D(y)}\Big/-\log\frac{D(x,y)}{M} \tag{3.4}$$

In other words, I am estimating the joint and marginal probabilities for x and y using document-level co-occurrence information internal to the corpus. While this estimation could be used to calculate normal topic coherence, I instead use $\hat{npmi}(x,y)$ as the basis for a measure of *inter-topic coherence*. Given two topics, z_1 and z_2, let $T^{z_1}_{i=1\ldots n}$ be a list of the n most probable words in z_1, and likewise for $T^{z_2}_{i=1\ldots n}$. Then, I define the inter-topic coherence to be:

$$ITC(z_1,z_2) = \frac{1}{n}\sum_{i=1}^{n}\hat{npmi}(T^{z_1}_i, T^{z_2}_i) \tag{3.5}$$

In theory, ITC has the same range of potential values as $npmi$, but over pairs of topics instead of words. If the top n words in z_1 and z_2 are perfectly associative, then $ITC(z_1,z_2) = 1$. And if all n pairs are mutually exclusive, it is -1.

Table 3.8 shows the results of calculating ITC over all $\binom{40}{2} = 780$ pairs from the 40-topic model presented above. The mean of these values is 0.0527 with a standard deviation of 0.0819. In the table, I show the two highest and lowest values, along with the median value. Encouragingly, the pair of problematic healthcare topics turns out to have the highest ITC value among all pairs. It is followed by a pair of closely related topics on the economy. The two topics in the median topic pair relate to budgeting and energy. Note that the value of ITC for this pair is 0.0661, indicating that most pairs of topics in this model are at least somewhat positively associated. Finally, the last two rows show topic pairs that are negatively associated: energy vs. gay marriage and healthcare vs. Middle East policy, respectively. This aligns with intuitions, since one would certainly expect these topics to be discussed in mutually exclusive contexts.

At first glance, it may seem naive to make only n comparisons in computing ITC rather than, say, comparing all n^2 pairs in the Cartesian product of T^{z_1} and T^{z_2}. In fact, the overall distributions of values turn out to be similar between these methodologies. The primary

Table 3.8: Values of inter-topic coherence over all pairs of topics in a 40-topic LDA model with $\alpha = 0.1$ and $\eta = 0.1$. The two pairs with the highest values are shown first, followed by the median pair, and, finally, the two pairs with the lowest values.

Index	Most Probable Terms	*ITC*
17	*care*, *health*, *health_care*, *insurance*, *plan*, *people*, *universal*, *health_insurance*...	0.4185
35	*health*, *care*, *insurance*, *people*, *plan*, *health_care*, *obamacare*, *medicare*...	
3	*jobs*, *economy*, *people*, *country*, *america*, *create*, *job*, *work*, *years*, *need*...	0.2840
25	*class*, *middle*, *people*, *middle_class*, *country*, *college*, *families*, *working*...	
10	*spending*, *cut*, *budget*, *billion*, *money*, *debt*, *dollars*, *taxes*, *government*...	0.0661
13	*energy*, *oil*, *change*, *going*, *need*, *climate*, *gas*, *climate_change*, *clean*, *power*...	
13	*energy*, *oil*, *change*, *going*, *need*, *climate*, *gas*, *climate_change*, *clean*, *power*...	-0.3010
15	*court*, *rights*, *supreme*, *right*, *supreme_court*, *marriage*, *constitution*, *law*...	
17	*care*, *health*, *health_care*, *insurance*, *plan*, *people*, *universal*, *health_insurance*...	-0.3991
19	*isis*, *need*, *syria*, *world*, *israel*, *middle*, *east*, *middle_east*, *ground*, *radical*...	

difference is that the variance of ITC's values is generally lower in the latter case. Given my goal of penalizing models that generate certain pairs of extremely similar topics, I prefer the higher variance so as to better discriminate between different levels of similarity. In addition, this formulation significantly reduces computational complexity.

ITC provides a measure of the semantic coherence between two topics. Semantic similarity, though, is not the only necessary condition for a pair of topics to be considered problematic. The two topics must also be split across parties. I will now define a measure, the *cross-party distance*, to quantify this split.

Cross-Party Distance

In Table 3.6, I presented a measure for the "one-sidedness" of a single topic based on the relative prominence of topics in each document's topic distribution. While it allowed for some intuitive conclusions at the time, it has two significant drawbacks that prevent it from being used here. First, it is a coarse measure that unnecessarily discards information by only considering a subset of the data in each calculation. Second, and more importantly, it

is only defined for a single topic at a time, whereas the "healthcare problem" by definition exists between a *pair* of topics. In this section, I will present a measure for the *cross-party distance* between two topics. I will begin by approximating the party distribution of a given topic z. Then, I will use a standard distributional divergence measure to quantify cross-party distance.

Recall that, for each token d_i in each document d, LDA performs the following generative procedure:

1. Draw a topic z from d's topic distribution θ_d.
2. Draw a word w from z's word distribution β_z.
3. Assign the identity of d_i to be w.

The result of the procedure is an $M \times k$ matrix ζ, where $\zeta_{d,z}$ is the number of tokens in d assigned to topic z.[10] This procedure occurs in every iteration, but I will use the state of ζ in LDA's final iteration as a means to estimate the party distribution of topic z. Formally:

$$\hat{\mathbb{P}}(\mathcal{P} = \rho | Z = z) = \frac{\sum_{d=1}^{M} \zeta_{d,z} I_{\mathcal{P}=\rho}(P_d)}{\sum_{d=1}^{M} \zeta_{d,z}} \tag{3.6}$$

Here, I is the indicator function, such that $I_{\mathcal{P}=\rho}(P_d) = 1$ if $P_d = \rho$, and 0 otherwise. In other words, I estimate $\mathbb{P}(\mathcal{P} = \rho | Z = z)$ as the proportion of terms assigned to topic z in the final iteration of LDA that came from documents with party ρ.

Having defined a topic's party one-sidedness as a probability distribution, I can now define the cross-party distance between two topics using any one of a number of standard measures of distributional distance. Kullback-Leibler divergence is one such measure, which has been shown to be useful in a variety of computational linguistic contexts (Lee, 2001). The KL divergence between two distributions p and q over the same random variable X can be defined as follows (Kullback and Leibler, 1951):

$$KLD(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \tag{3.7}$$

[10] This matrix is not explicitly named in the original paper for LDA; the use of ζ is specific to this document.

One drawback of KL divergence in the current context is that it is asymmetric. This property is useful in many cases, but it doesn't make much sense to define the cross-party distance between topics z_1 and z_2 as being different than the distance between z_2 and z_1. Another shortcoming is that KL divergence is not guaranteed to be bounded. If X has two outcomes, as in the present case, then the possible values of KL divergence range between 0 and ∞. This is rather undesirable, since my aim is to combine this measure with ITC, which ranges between -1 and 1.

Fortunately, there is another measure of distance, based on KL divergence, that addresses these limitations. It is called the Jensen-Shannon divergence, and is defined as

$$JSD(p||q) = (KLD(p||m) + KLD(q||m))/2 \tag{3.8}$$

where m is the average distribution of p and q. That is, $m(x) = (p(x)+q(x))/2$.[11] The Jensen-Shannon divergence of two distributions is symmetric, and its value always falls between 0 and 1 inclusive, assuming that $\log_2$ is used in computing the KL divergences. Now, using Jensen-Shannon divergence, the cross-party distance between two topics z_1 and z_2 can be defined as

$$CPD(z_1, z_2) = JSD(\mathbb{P}(\rho|Z = z_1) \;||\; \mathbb{P}(\rho|Z = z_2)) \tag{3.9}$$

where $\mathbb{P}(\mathcal{P} = \rho|Z = z)$ is estimated as in Equation 3.6. The maximum, median, and minimum values of CPD, taken from the same 40-topic model as before, are given in Table 3.9. For orientation, $\hat{\mathbb{P}}(\mathcal{D}|Z = z)$ is also provided for each topic.

Combining ITC and CPD

In the previous two subsections, I presented a means of calculating the inter-topic coherence and cross-party distance between two topics z_1 and z_2. In order to use these measures for model selection, they need to be first combined and then aggregated, so that a single score can be calculated per model.

[11] The original definition of Jensen-Shannon divergence in Lin (1991) allowed the distributions p and q to be parameterized by weights $\pi_p + \pi_q = 1$. Here, though, as is the general practice in NLP, I assume that $\pi_p = \pi_q = 0.5$ in order to weight the distributions equally.

Table 3.9: Values of cross-party distance over all pairs of topics in a 40-topic LDA model with $\alpha = 0.1$ and $\eta = 0.1$, along with $\hat{\mathbb{P}}(\mathcal{D}|Z = z)$ for each topic. The two pairs with the highest values are shown first, followed by the median pair, and, finally, the two pairs with the lowest values.

Index	Most Probable Terms	$\hat{\mathbb{P}}(\mathcal{D}\|Z = z)$	CPD
8	*people*, *border*, *immigration*, *country*...	0.1816	0.2899
17	*care*, *health*, *health_care*, *insurance*...	0.7929	
17	*care*, *health*, *health_care*, *insurance*...	0.7929	0.2748
28	*government*, *federal*, *state*, *federal_government*...	0.1963	
12	*social*, *security*, *social_security*...	0.4472	0.01880
32	*people*, *south*, *thank*, *want*, *carolina*...	0.6081	
22	*street*, *wall*, *wall_street*, *banks*, *big*...	0.4716	3.232×10^{-8}
36	*think*, *important*, *know*, *people*, *good*...	0.4714	
0	*iran*, *nuclear*, *weapons*, *nuclear_weapons*...	0.4768	6.091×10^{-10}
1	*ve*, *got*, *ve_got*, *going*, *think*, *know*, *lot*...	0.4767	

For a given topic pair (z_1, z_2), I first define the score for that pair as simply the product of $ITC(z_1, z_2)$ and $CPD(z_1, z_2)$. A pair should only be considered bad if it is split across parties and, since CPD ranges from 0 to 1, it intuitively acts as the "weight" of the pair in this definition. Thus, if two topics are similar, but their cross-party distance is small, their similarity will not have a significant effect on the overall model's score.

Given this means of scoring a particular pair of topics, individual pairs' scores now need to be summarized into a single score for each model. Intuitively, this summarization should capture the overall "badness" of a model based on the presence or absence of problematic topic pairs. One option would be to simply take the mean of the $ITC \cdot CPD$ scores for each of the $\binom{k}{2}$ topic pairs in a given k-topic model. However, including all pairs would introduce a significant amount of unnecessary information. Consider, for example, the bad pair of healthcare topics presented earlier. From the perspective of one of these topics, call it $health_1$, it is sufficient to know that $health_2$ is extremely similar and cross-party distant. The relationship between $health_1$ and, say, a topic about the economy, is irrelevant.

Based on this observation, I restrict the set of topic pairs that contribute to the overall model score. Specifically, for each topic z, I only consider the topic z' that is *most similar*; that is, for which the value of $ITC(z, z')$ is highest. Then, given this set of k pairs, I define the overall model score to be not the mean, but the standard deviation, of the $ITC \cdot CPD$ values for each of the pairs. My motivation for doing so can be seen in Figure 3.1, which shows the distribution of $ITC \cdot CPD$ scores for all most-similar topic pairs across a set of 225 LDA models with varying parameters (the makeup of this set of models will be described in the next section). As is evident from the figure, most pairs of topics have very small $ITC \cdot CPD$ values, so using a mean would make it very difficult to differentiate models with bad topic pairs. Standard deviation, on the other hand, naturally summarizes the influence of bad topic pairs. To state this definition more formally, let $sim(z)$ be the following function, for any topic z in a model with k topics:

$$sim(z) = \underset{1 \leq z' \leq k, z' \neq z}{\arg\max} \left(ITC(z, z')\right) \tag{3.10}$$

Then, the score for a given model with k topics is

$$\sqrt{\frac{1}{k} \sum_{z=1}^{k} \left(ITC(z, sim(z)) CPD(z, sim(z)) - \mu\right)^2} \tag{3.11}$$

where $\mu = \frac{1}{k} \sum_{z=1}^{k} \left(ITC(z, sim(z)) CPD(z, sim(z))\right)$.

Recall that ITC ranges from -1 to 1, with a value of 1 indicating very similar topics. So, given this formulation, a *lower* score for a model is better, since it indicates that there are fewer pairs of topics that are both similar to each other and oppositely distributed between parties. Thus, any selection from among a set of candidate models should focus on those models with the lowest scores.

3.2.7 Model Selection: The Final Model

In order to select a final model, I perform a grid search over a set of k's, α's, and η's, training a total of 225 models. The full set of parameter values is summarized in Table 3.10.

Figure 3.1: Histogram of $ITC \cdot CPD$ val-

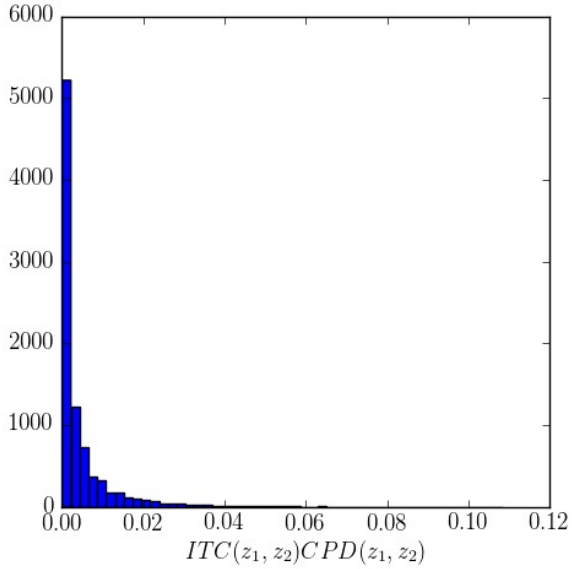


Figure 3.2: Distribution of model scores

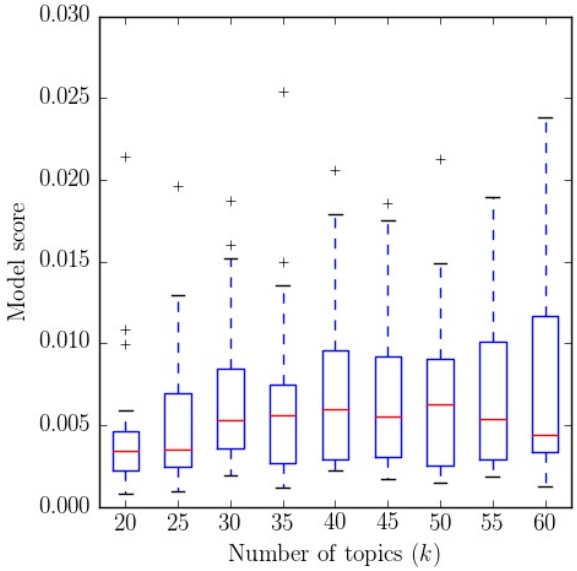


Figure 3.2 shows the distribution of model scores for all 225 models, split by k. It seems that, overall, models with more topics have a slightly higher chance of containing bad topic pairs.

The objective of this model selection effort is to select a model that will be useful in a topic-specific analysis of framing. And, while the model scoring method developed above usefully quantifies the influence of bad topic pairs, it certainly does not capture all factors relevant to this objective. As such, I do not blindly take the model with the lowest score as the final topic model. Rather, I opt to use the models' scores as a sort function, starting from the model with the lowest score and performing a manual inspection of each model's topics until arriving at one suitable for use in the remaining phases of the analysis. The final model as chosen by this procedure is a 25-topic model with $\alpha = 0.01$ and $\eta = 0.01$. It has the second lowest model score – the lowest, a 20-topic model, was rejected on the grounds that it conflated most social issues into one topic. The top terms for each topic in this final model are shown in Table 3.11. Additionally, Figure 3.3 plots the log-likelihood of the model

Table 3.10: Parameter values for LDA grid search over k, α, and η.

Param	Values	Count
k	20, 25, 30, 35, 40, 45, 50, 55, 60	9
α	1, 0.5, 0.1, 0.05, 0.01	5
η	0.1, 0.05, 0.01, 0.005, 0.001	5
	Number of Models	*225*

in each iteration of training to establish convergence.

Before continuing, I will point out that several of the topics in this model are not suitable for use in the rest of the analysis. First, topic 11 does not appear to be immediately interpretable. Topics 0, 9, and 15 are perhaps too general to warrant inclusion in a topic-specific analysis. Finally, topics 12 and 22 are moderator-specific. One focuses on debate procedure, while the other seems to be about the opening and closing of debates. Aside from these six topics, the remaining 19 are semantically interpretable concepts that correspond to political issues. I have assigned each of these a short-hand label in order to be able to refer to them throughout the rest of the analysis without using their topic index.[12,13,14]

With a topic model in hand, I will now move on to measuring polarity.

An Aside: Reranking Terms for Topic Readability

Traditionally, topics generated by LDA and related models are represented using the n most probable terms in that topic, as they have been up to this point. Often, however, top terms are common across many topics in the corpus, which reduces the descriptiveness of this repre-

[12]The MidEast label is not meant to encompass the IraqWar label. Rather, it refers to any Middle East discussion that is *not* specifically related to the Iraq War.

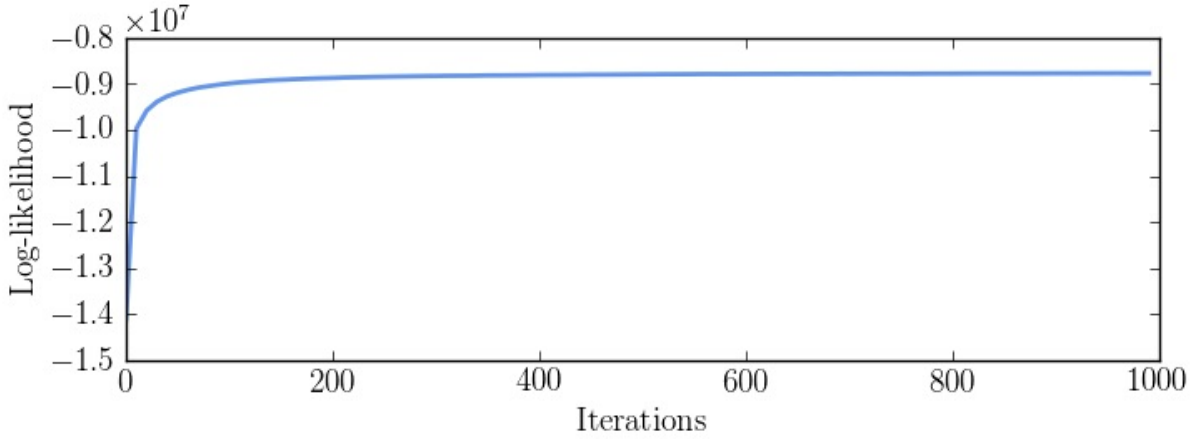
[13]DADT stands for *Don't Ask, Don't Tell*, a controversial policy regarding gays in the military.

[14]CampFin stands for campaign finance reform.

Table 3.11: Top terms for each topic in the final model – a 25-topic model with $\alpha = 0.01$ and $\eta = 0.01$ – along with a semantic label for each topic. Uninterpretable topics and moderator-related topics have no label.

Index	Label	Top Terms
0	–	*people*, *think*, *going*, *ve*, *country*, *make*, *want*, *know*, *america*, *american* ...
1	SocialSec	*security*, *social*, *social_security*, *going*, *spending*, *budget*, *cut*, *billion*, *money* ...
2	Courts	*court*, *law*, *supreme*, *supreme_court*, *constitution*, *states*, *right*, *amendment*...
3	Race	*think*, *americans*, *african*, *american*, *know*, *america*, *country*, *community*...
4	PrimaryEd	*education*, *school*, *children*, *schools*, *kids*, *need*, *child*, *teachers*, *parents*, *think* ...
5	MidEast	*war*, *iran*, *need*, *world*, *nuclear*, *isis*, *military*, *think*, *going*, *iraq* ...
6	DADT	*government*, *state*, *think*, *federal*, *states*, *federal_government*, *make*, *people*, *right* ...
7	College	*women*, *country*, *college*, *men*, *young*, *veterans*, *need*, *people*, *years*, *men_women* ...
8	Abortion	*life*, *reagan*, *ronald*, *ronald_reagan*, *pro*, *believe*, *abortion*, *right*, *pro_life*, *rights* ...
9	–	*ve*, *years*, *state*, *people*, *got*, *did*, *jobs*, *record*, *time*, *new* ...
10	Immigration	*immigration*, *people*, *border*, *country*, *illegal*, *come*, *immigrants*, *going*, *need*, ...
11	–	*said*, *think*, *know*, *say*, *just*, *did*, *ve*, *people*, *didn*, *right* ...
12	–	*let*, *question*, *want*, *just*, *going*, *seconds*, *ask*, *30*, *ll*, *respond* ...
13	WallSt	*people*, *government*, *money*, *going*, *know*, *street*, *wall*, *economy*, *wall_street*, *big* ...
14	Economy	*tax*, *percent*, *taxes*, *plan*, *income*, *economy*, *people*, *pay*, *jobs*, *middle* ...
15	–	*party*, *going*, *think*, *republican*, *know*, *republicans*, *people*, *win*, *democrats* ...
16	Climate	*energy*, *oil*, *going*, *change*, *need*, *ve*, *climate*, *new*, *gas*, *climate_change* ...
17	ForeignPol	*united*, *united_states*, *states*, *think*, *world*, *policy*, *foreign*, *israel*, *america*...
18	Trade	*trade*, *jobs*, *china*, *america*, *ve*, *going*, *country*, *workers*, *american*, *world* ...
19	Guns	*gun*, *south*, *people*, *carolina*, *south_carolina*, *think*, *guns*, *said*, *just*, *state* ...
20	CampFin	*money*, *people*, *campaign*, *interests*, *think*, *special*, *washington*, *know*, *reform*...
21	Health	*health*, *care*, *health_care*, *insurance*, *people*, *plan*, *health_insurance*, *going* ...
22	–	*question*, *thank*, *debate*, *candidates*, *right*, *going*, *tonight*, *want*, *let*, *ll* ...
23	CrimJust	*people*, *new*, *need*, *city*, *crime*, *law*, *justice*, *police*, *think*, *ve* ...
24	IraqWar	*iraq*, *war*, *troops*, *going*, *ve*, *time*, *military*, *end*, *american*, *home* ...

Figure 3.3: Log-likelihood per iteration of collapsed Gibbs sampling in the final topic model.



sentation. Instead, one can rank a term w within a topic z based on the mutual information between w and z, which captures how much the presence of w reduces uncertainty about whether or not the current document is predominantly about z. Unless otherwise noted, all LDA results going forward, including Table 3.11, are presented using this information-based ranking method, rather than using the topics' most probable terms. For more details on the calculation of this measure, refer to Grimmer (2010, Section 7.2).

3.3 Phase II: Measuring Polarity

In this section, I will derive a method for quantifying the party polarity of different lexical frames, then use those frame-specific measures to calculate the topic-specific polarity of entire documents. For the purposes of this analysis, a lexical frame is nothing more than a unigram or bigram (i.e. a term) used in the context of a particular topic. I will begin by describing a standard measure used for ranking words, and show how it can be adapted for topic-specific use in this analysis. I will then progressively improve the measure over three stages, pointing

out weaknesses and presenting refinements in each stage.[15] Once I have arrived at a final measure for topic-specific lexical frame polarity, I will present a method for aggregating those term polarity scores to the level of documents.

3.3.1 Log Odds Ratio

The *odds ratio* is a standard statistical measure of association. Between two events, A and B, it can be defined as such (Sinclair and Bracken, 1994):

$$OR(A, B) = \frac{\mathbb{P}(A)/\mathbb{P}(A^C)}{\mathbb{P}(B)/\mathbb{P}(B^C)} \tag{3.12}$$

Odds ratios are commonly found in biomedical research as a means of assessing risk. Recently, however, they have found success in computational political science as a way to associate words with specific parties (Grimmer and Stewart, 2013). More specifically, due to the asymmetric nature of odds ratios, *log* odds ratios have been used instead:

$$LOR(A, B) = \log\left(\frac{\mathbb{P}(A)}{\mathbb{P}(A^C)}\right) - \log\left(\frac{\mathbb{P}(B)}{\mathbb{P}(B^C)}\right) \tag{3.13}$$

Now, rather than events A and B, consider an event X, conditioned in one case on event Y_1 and in another case on event Y_2. One could then define a modified LOR_X as:

$$LOR_X(Y_1, Y_2) = \log\left(\frac{\mathbb{P}(X|Y_1)}{\mathbb{P}(X^C|Y_1)}\right) - \log\left(\frac{\mathbb{P}(X|Y_2)}{\mathbb{P}(X^C|Y_2)}\right) \tag{3.14}$$

The usefulness of LOR for lexical analysis is clearer when restated in this fashion. In this definition, $LOR_X(Y_1, Y_2)$ will be very negative if X is highly associated with Y_1 and very positive if it's associated with Y_2.

Based on its previous successes in computational political science, I will use LOR as a starting point in developing a definition of polarity for lexical frames. For each term w in the corpus, I'd like to calculate the log-odds ratio between Democrats' and Republicans' usage of the term. So, given $\mathbb{P}(W = w|\mathcal{P} = \rho)$, I need to calculate the following:

$$LOR_{W=w}(\mathcal{D}, \mathcal{R}) = \log\left(\frac{\mathbb{P}(W = w|\mathcal{D})}{1 - \mathbb{P}(W = w|\mathcal{D})}\right) - \log\left(\frac{\mathbb{P}(W = w|\mathcal{R})}{1 - \mathbb{P}(W = w|\mathcal{R})}\right) \tag{3.15}$$

[15] This course of refinements is heavily inspired by Monroe et al. (2009).

Table 3.12: Top 20 terms for each party, weighted by $LOR_w^{\mathcal{D}-\mathcal{R}}$.

Democrats		Republicans	
public_financing	-18.624646	*pro_life*	18.233145
college_tuition	-18.624630	*flat_tax*	18.233013
fossil_fuel	-18.624621	*radical_islamic*	18.232894
public_colleges	-18.624613	*fox*	18.232881
steagall	-18.624610	*free_enterprise*	18.232874
glass_steagall	-18.624610	*job_creation*	18.232868
150_month	-18.624606	*repeal_obamacare*	18.232867
political_revolution	-18.624606	*religious_liberty*	18.232867
family_medical	-18.624602	*wanna*	18.232863
fuel_industry	-18.624602	*creator*	18.232863
medical_leave	-18.624602	*caliphate*	18.232859
hiv	-18.624599	*islamic_terrorism*	18.232852
gun_safety	-18.624595	*debt_ceiling*	18.232850
free_tuition	-18.624593	*unborn*	18.232847
profit_health	-18.624593	*bubble*	18.232845
childhood_education	-18.624593	*come_legally*	18.232843
corrupt_campaign	-18.624588	*socialized*	18.232843
universal_pre	-18.624588	*verify*	18.232841
black_caucus	-18.624586	*shrink*	18.232841
residual	-18.624586	*bain*	18.232841

For notational compactness, I will call this measure $LOR_w^{\mathcal{D}-\mathcal{R}}$. Calculating this number globally is not difficult. I estimate $\mathbb{P}(W = w|\mathcal{D})$ directly (and likewise for $\mathbb{P}(W = w|\mathcal{R})$) using maximum likelihood estimation. That is, I simply aggregate over all of the documents from Democratic and Republican debates and calculate the observed proportion of w in that set of documents:

$$\hat{\mathbb{P}}(W = w|\mathcal{P} = \rho) = \frac{\sum_{d=1}^{M} C_{d,w} I_{\mathcal{P}=\rho}(P_d)}{\sum_{d=1}^{M}\sum_{v=1}^{V} C_{d,v} I_{\mathcal{P}=\rho}(P_d)} \tag{3.16}$$

Once again, as in Equation 3.6, I is the indicator function.

$LOR_w^{\mathcal{D}-\mathcal{R}}$ provides a measure of the party specificity of each term. Very Democratic

terms will have highly negative values, and very Republican terms will be highly positive. Table 3.12 shows the 20 most highly weighted terms for each party.

These results seem to align with what one might expect, at least in a broad sense. Two problems, however, are readily apparent. First, a significant number of the top words are extremely conceptually specific. The term *glass_steagall*, for example, refers to several provisions of the U.S. Banking Act of 1933 regarding the regulation of banks' activities. It was brought up several times in Democratic primary debates in the wake of the 2008 financial crisis. In this sense, it is a highly Democratic word, but it is far too conceptually specific – and too rare – to be considered a major contributor to party polarity. Other terms which are only incidentally party-specific, like *residual* for Democrats and *bain* for Republicans, should not be so highly weighted either. The second, and more pressing, problem is that these global results, while interesting, are not very useful in an analysis of framing, which is inherently tied to a particular topic or issue. I will address issue this now, and will return to the specificity problem afterwards.

In order to make LOR applicable to the present investigation, it needs to be defined in the context of a given topic. That is, I need to calculate not $LOR_w^{\mathcal{D}-\mathcal{R}}$, but $LOR_{w,z}^{\mathcal{D}-\mathcal{R}}$ for all terms w and all topics z. Introducing this new parameter into the previous equation, we have:

$$LOR_{w,z}^{\mathcal{D}-\mathcal{R}} = \log\left(\frac{\mathbb{P}(W=w|\mathcal{D},Z=z)}{1-\mathbb{P}(W=w|\mathcal{D},Z=z)}\right) - \log\left(\frac{\mathbb{P}(W=w|\mathcal{R},Z=z)}{1-\mathbb{P}(W=w|\mathcal{R},Z=z)}\right) \tag{3.17}$$

Estimating $\mathbb{P}(W=w|\mathcal{P}=\rho,Z=z)$, however, is far less straightforward than its global counterpart, $\mathbb{P}(W=w|\mathcal{P}=\rho)$. Unlike with parties, there isn't one single topic assignment for any given document in the corpus. Indeed, the fundamental assumption of LDA is that every document is a mixture of *multiple* topics. Likewise, any given term in any given document cannot be assigned to a single topic either. So, what is the path forward?

Recall that LDA provides a distribution over topics for each document in the form of the matrix θ. That is, $\theta_{d,z} = \mathbb{P}(Z=z|D=d)$ for a given document d and topic z. One potential approach, then, would be to simply look at each document's topic distribution and select the

most probable topic as the "representative" topic for that document, similar to the approach in Table 3.6. Indeed, this method has been recommended as a heuristic for identifying important or dominant topics in a corpus (Boyd-Graber et al., 2014). Then, after partitioning the documents by most probable topic, $LOR_{w,z}^{\mathcal{D}-\mathcal{R}}$ could be calculated over each topic's "sub-corpus." Formally, let $R(d,z)$ (for "representative") equal $I_{D=d,Z=z}(z = \arg\max_{1\leq z'\leq k}(\theta_{d,z'}))$. In other words, $R(d,z) = 1$ when $\theta_{d,z}$ is the largest value in θ_d and is 0 otherwise. Then:

$$\hat{\mathbb{P}}(W = w|\mathcal{P} = \rho, Z = z) = \frac{\sum_{d=1}^{M} C_{d,w} R(d,z) I_{\mathcal{P}=\rho}(P_d)}{\sum_{d=1}^{M}\sum_{v=1}^{V} C_{d,v} R(d,z) I_{\mathcal{P}=\rho}(P_d)} \tag{3.18}$$

Having been calculated as such, $\hat{\mathbb{P}}(W = w|\mathcal{D}, Z = z)$ and $\hat{\mathbb{P}}(W = w|\mathcal{R}, Z = z)$ can be substituted back in to Equation 3.17.

The leftmost columns in Tables 3.13, 3.14, and 3.15 show the results of this technique for a few selected topics. The results look promising from the perspective of lexical framing, but there are some clear problems. In Immigration, for example (Table 3.14), some differing perspectives have emerged, with Democrats talking about paths to citizenship and Republicans stressing the legal aspects of the issue. In Health as well (Table 3.15), indications of each party's framing strategies have begun to appear. Overall, though, the measure seems very noisy, and gives high weight to extremely specific terms that don't capture topic-wide patterns. This mirrors the specificity issue observed earlier in the global results.

The erratic results stem from the fact that $LOR_{w,z}^{\mathcal{D}-\mathcal{R}}$ gives high weight to infrequent terms that happen to be party-specific in this particular dataset. Introducing some probabilistic assumptions into the calculation can control for this problem.

3.3.2 Variance-Weighted Log Odds Ratio

So far I have directly estimated $\mathbb{P}(W = w|\mathcal{P} = \rho, Z = z)$ using the observed counts in the corpus. While I have acknowledged that these are only estimates, I have ignored the level of uncertainty that these estimates have. Now, to remedy this, I instead take a Bayesian

Table 3.13: Most Democratic (top) and most Republican (bottom) terms from progressively improved polarity measurements for the ABORTION topic.

$LOR^{\mathcal{D}-\mathcal{R}}_{w,z}$	$Z(LOR^{\mathcal{D}-\mathcal{R}}_{w,z})$	$Z(LOR^{\mathcal{D}-\mathcal{R}}_{w,z})_{reg}$
catholic_doctrine	*public*	*right_choose*
remedies	*courage*	*woman_right*
decisive_moment	*right*	*decisive_moment*
took_position	*fought*	*decisive*
came_campaign	*answer*	*spiritual*
rallied	*pray*	*moment_life*
support_woman	*public_life*	*public*
make_peace	*woman_right*	*health*
proudly	*values*	*women_rights*
empower_american	*moment*	*courage*
⋮	⋮	⋮
issue_life	*world*	*planned*
unborn_child	*life*	*god*
abortion_wrong	*great*	*liberty*
freedom_born	*conservative*	*abortion*
body_parts	*liberty*	*ronald*
new_birth	*america*	*ronald_reagan*
certain_inalienable	*ronald*	*reagan*
birth_freedom	*ronald_reagan*	*pro_life*
inalienable_rights	*reagan*	*pro*
inalienable	*pro*	*life*

Table 3.14: Most Democratic (top) and most Republican (bottom) terms from progressively improved polarity measurements for the IMMIGRATION topic.

$LOR^{\mathcal{D}-\mathcal{R}}_{w,z}$	$Z(LOR^{\mathcal{D}-\mathcal{R}}_{w,z})$	$Z(LOR^{\mathcal{D}-\mathcal{R}}_{w,z})_{reg}$
relationship_mexico	*reform*	*immigration_reform*
people_path	*comprehensive*	*comprehensive_immigration*
dapa	*immigration_reform*	*comprehensive*
employers_taking	*workers*	*reform*
2007_immigration	*undocumented*	*undocumented*
workers_drive	*comprehensive_immigration*	*undocumented_workers*
workers_punished	*families*	*need_comprehensive*
rights_thrown	*congress*	*relationship*
poverty_law	*path*	*million_undocumented*
southern_poverty	*children*	*2007*
⋮	⋮	⋮
enforce_border	*legal*	*secure_border*
language_government	*illegal_immigration*	*legally*
permanent_resident	*people*	*legal*
come_border	*going*	*illegal_immigration*
american_express	*secure*	*secure*
permanent_residents	*amnesty*	*amnesty*
employment_verification	*illegally*	*come*
aviation_assets	*border*	*illegally*
residency	*illegal*	*illegal*
immigration_control	*come*	*border*

Table 3.15: Most Democratic (top) and most Republican (bottom) terms from progressively improved polarity measurements for the HEALTH topic.

$LOR^{\mathcal{D}-\mathcal{R}}_{w,z}$	$Z(LOR^{\mathcal{D}-\mathcal{R}}_{w,z})$	$Z(LOR^{\mathcal{D}-\mathcal{R}}_{w,z})_{reg}$
universal_single	*affordable*	*universal*
junk_food	*children*	*universal_health*
mandate_parents	*drug*	*health_care*
month_voucher	*companies*	*health*
create_children	*universal*	*affordable*
congressional_plan	*drug_companies*	*companies*
independent_experts	*coverage*	*drug*
believe_universal	*know*	*drug_companies*
food_schools	*middle*	*care*
leave_15	*important*	*coverage*
⋮	⋮	⋮
like_market	*repeal*	*account*
insurance_government	*federal_government*	*socialized*
insurance_policies	*free*	*health_savings*
government_insurance	*massachusetts*	*insured*
state_lines	*buy*	*market*
called_saving	*market*	*massachusetts*
lives_saving	*federal*	*medicine*
free_riders	*obamacare*	*state*
free_care	*state*	*government*
everybody_insured	*government*	*obamacare*

perspective, assuming that some underlying but unknown multinomial distribution exists over the vocabulary for each party and topic. Let $\pi^{\rho,z}_{w=1\ldots V}$ be the vector of these true probabilities $\mathbb{P}(W = w|\mathcal{P} = \rho, Z = z)$. Additionally, let $C^{\rho,z}_{w=1\ldots V}$ be a frequency vector of term counts for each topic and party's "sub-corpus." Finally, let $N^{\rho,z}$ be the total number of terms in that sub-corpus. Then, I am introducing the following assumption:

$$C^{\rho,z} \sim \text{Multinomial}(N^{\rho,z}, \pi^{\rho,z}) \tag{3.19}$$

In other words, the observed corpus $C^{\rho,z}$ is simply the result of repeatedly choosing an outcome from $\pi^{\rho,z}$ over $N^{\rho,z}$ trials. Though the true value of $\pi^{\rho,z}$ is not known, an estimate is readily available. For any word w, I estimate $\pi^{\rho,z}_w$ as $\hat{\pi}^{\rho,z}_w = \hat{\mathbb{P}}(W = w|\mathcal{P} = \rho, Z = z)$.

This representation is more than mere abstraction. One major advantage of the multinomial assumption is that I can now calculate the variance of each w in $C^{\rho,z}$ and, by extension, the variances for $LOR^{\mathcal{D}-\mathcal{R}}_{w,z}$. Following Monroe et al. (2009), this variance can be estimated as:[16]

$$\text{Var}(LOR^{\mathcal{D}-\mathcal{R}}_{w,z}) \approx \frac{1}{C^{\mathcal{D},z}_w} + \frac{1}{C^{\mathcal{R},z}_w} \tag{3.20}$$

That is, the variance of a given estimate of $LOR^{\mathcal{D}-\mathcal{R}}_{w,z}$ will be relatively large if w is rare in either party's sub-corpus for topic z.

This variance can be used to improve a term's polarity score. Intuitively, if little information is available for a given term w (i.e. if w is infrequent and its variance therefore high), its polarity score should be less extreme than that of a more frequent but equally one-sided term. One way to capture this is to standardize each $LOR^{\mathcal{D}-\mathcal{R}}_{w,z}$ from a raw estimate to a z-score.[17]

$$Z(LOR^{\mathcal{D}-\mathcal{R}}_{w,z}) = \frac{LOR^{\mathcal{D}-\mathcal{R}}_{w,z}}{\sqrt{\text{Var}(\text{LOR}^{\mathcal{D}-\mathcal{R}}_{\text{w,z}})}} \tag{3.21}$$

---

[16]This is actually a slight abuse of notation. Technically, $C^{\mathcal{D},z}_w$ should be $C^{dem,z}_w$ and $C^{\mathcal{R},z}_w$ should be $C^{rep,z}_w$. I opted for the former, given that the actual values *dem* and *rep* have not been used since their original definition.

[17]The reason the mean is missing from this equation is that it is assumed to be zero. Another way to frame this calculation is as a z-test for each value of $LOR^{\mathcal{D}-\mathcal{R}}_{w,z}$ under the null hypothesis that $LOR^{\mathcal{D}-\mathcal{R}}_{w,z} = 0$.

This standardized value captures both the magnitude of the estimate $LOR^{\mathcal{D}-\mathcal{R}}_{w,z}$ and certainty about its correctness. The center columns in Tables 3.13, 3.14, and 3.15 show the results of this calculation, and the improvements over the previous estimate are immediately apparent. It now is clear from the data, for example, that Republicans prefer to frame the discussion of IMMIGRATION from the perspective of legality, with terms like *illegal* and *amnesty* dominating. Democrats, on the other hand, are more likely stress the need for *reform*, with a focus on the *workers* who are already in the country. And in HEALTH, Democrats can be seen stressing *universal coverage* and Republicans stressing the role of the *states* and using the *obamacare* moniker. There is room for improvement, however. Consider the ABORTION topic. While there are some indications of positions on abortion (e.g. *pro* for Republicans and *right* for Democrats), some of the most highly-weighted terms, like *fought* and *great*, are not topic-specific.

3.3.3 Variance-Weighted Log Odds Ratio With An Informative Bayesian Prior

This problem can be addressed by introducing a form of regularization that pushes the values of $Z(LOR^{\mathcal{D}-\mathcal{R}}_{w,z})$ towards zero, especially for terms that are common but not topic-specific. To enable this, assume that there is a Bayesian prior over the true probability distribution $\pi^{\rho,z}$ by representing it as a draw from a Dirichlet distribution:

$$\pi^{\rho,z} \sim \text{Dirichlet}(\alpha^{\rho,z}) \tag{3.22}$$

Previously, the estimate $\hat{\pi}^{\rho,z}$ was just $\hat{\mathbb{P}}(W = w|Z = z, \mathcal{P} = \rho)$. With this new prior, the estimate changes. Fortunately, there is still a direct analytic solution (once again, after Monroe et al. 2009):

$$\hat{\pi}^{\rho,z}_w = \frac{C^{\rho,z}_w + \alpha^{\rho,z}_w}{N^{\rho,z} + \sum_{v=1}^{V} \alpha^{\rho,z}_v} \tag{3.23}$$

Given this formula, the values used for $\alpha^{\rho,z}$ have a surprisingly intuitive interpretation. For any term w, the value of $\alpha^{\rho,z}_w$ adjusts the value of $\hat{\pi}^{\rho,z}_w$ as there had been $\alpha^{\rho,z}_w$ more occurrences of w in the corpus $C^{\rho,z}$.

What, then, should the values of $\alpha^{\rho,z}$ be, given that the goal is to reduce the impact of words like *fought* and *great*? One approach would be to define a global V-dimensional vector α of size A, and set each α_w proportional to the frequency of w in the full corpus C. That is:

$$\alpha_w^{\rho,z} = \alpha_w = \frac{\sum_{d=1}^{M} C_{d,w}}{\sum_{d=1}^{M}\sum_{v=1}^{V} C_{d,v}} \cdot A \tag{3.24}$$

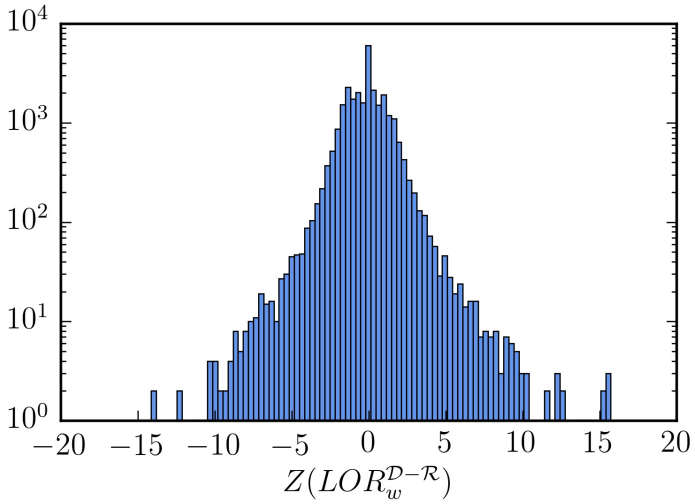
Smoothing each $\hat{\pi}^{\rho,z}$ with this vector, as in Equation 3.23, essentially gives each sub-corpus $C^{\rho,z}$ more information about the "true" relative proportions of each word in the vocabulary. Effectively, it pushes the values of $Z(LOR_{w,z}^{\mathcal{D}-\mathcal{R}})$ close to zero, leaving at the extremes only those terms which are very topic-specific and very polar. How great of an effect this smoothing has depends on the choice of A. Based on qualitative evaluation, I choose $A = 10^4$.

I denote this regularized measure $Z(LOR_{w,z}^{\mathcal{D}-\mathcal{R}})_{reg}$, and adopt it as my final definition of the polarity of topic-specific lexical frames. The improvements of this method over unregularized $Z(LOR_{w,z}^{\mathcal{D}-\mathcal{R}})$ are subtle, but nonetheless apparent. Looking at the rightmost column in Table 3.13, the new scores are far more reflective of the party-specific framing strategies for Abortion, with Democrats emphasizing a woman's *right* to *choose* and Republicans positioning themselves as *pro-life*.

3.3.4 Document-Level Polarity

So far, I have developed a means of measuring the party polarity of topic-specific lexical frames. However, the unit of analysis in this investigation is not an individual term, but a document. As such, these term scores need to be aggregated to the document level, generating document-topic polarity scores that will be used in the rest of the analysis. The approach I will carry forward is very simple: the topic-specific polarity of a document is the sum of the topic-specific polarities of its constituent terms. I will call this measure Pol_d^z for

Figure 3.4: Distribution of $Z(LOR_w^{\mathcal{D}-\mathcal{R}})$ scores for each term on a log scale.



a given document d and topic z. That is:

$$Pol_d^z = \sum_{w=1}^{V} C_{d,w} Z(LOR_{w,z}^{\mathcal{D}-\mathcal{R}})_{reg} \tag{3.25}$$

I define Pol_d^z using a simple sum – rather than, say, a mean – so as not to penalize polar documents that happen to be long. This decision is supported by Figure 3.4, which shows that the global distribution of term polarity scores is symmetric. In other words, given a document of length n, and the addition of a new word at position $n+1$ (say, v_1), that new word is equally likely to have positive or negative polarity. Furthermore, given another new word at position $n+2$ (say, v_2), $\mathrm{E}[Z(LOR_{v_1}^{\mathcal{D}-\mathcal{R}})_{reg} + Z(LOR_{v_2}^{\mathcal{D}-\mathcal{R}})_{reg}] = 0$. Note that the scale in the figure is logarithmic. This is because the vast majority of term polarities are concentrated around zero, a finding in keeping with the subjectivity detection literature (Fernández-Gavilanes et al., 2016).

3.4 Phase III: Associating Applause and Polarity

One final step remains before testing the hypothesis that applause is associated with polarity. In this section, I will define how to actually attribute and measure applause.

3.4.1 Attributing Applause

In the parsed debate transcripts, applauses are represented as non-utterance events. Their place in the parsed transcript corresponds exactly to their place in the raw transcript, that is, immediately after the last utterance that started before the applause.

Given that the unit of analysis is documents, it would seem natural to attribute a given applause to the document that contains the utterance immediately before the applause occurs. In most situations, this may be perfectly logical, but this naive method fails to account for the problem outlined in Section 3.1.1, in which applauses are not always annotated immediately after the utterance that caused them. Consider the following snippet from the transcript of a 2016 Republican debate (Peters and Woolley, 2016c):

> **RUBIO:** [$\cdots$] And when I'm president, we're not just going to have a president that gives a State of the Union and says America is the greatest country in the world. When I'm president, we're going to have a president that acts like it.
> **BARTIROMO:** Thank you, senator.
> Mr. Trump, South Carolina Governor Nikki Haley in her response to the State of the Union address...[*applause*]...appeared [$\cdots$]

It is clearly Senator Rubio's utterance that has caused the applause, but the naive attribution method would instead attribute it to the moderator's utterance immediately following. The applause is transcribed this way because, strictly speaking, it did not start until *after* the moderator started speaking.

This pattern appears regularly in the data. Fortunately, compensating for it is not difficult. In cases where the speaker of the utterance immediately preceding an applause event is not a candidate, I simply step backwards to the last utterance by a candidate, and

attribute the applause to that utterance. Then, the applause event is attributed to the document in which that utterance falls.[18]

3.4.2 Measuring Applause

In this section, I describe two methods – one direct and one indirect – for measuring the applause in a given document. I will use both of these measures when testing the association between polarity and applause.

Direct Measurement

The most obvious – and direct – way to measure applause is to simply consider applause as a binary response variable. If a given document contains an utterance to which an applause event is attributed, then the value is 1. Otherwise, it is 0. I define this response variable per document, and call it $HasAppl_d$. Given that the transcripts do not provide indications of applause strength – such as loudness or duration – this binary response is the most specific measure of applause directly available from the data. While its coarseness certainly restricts the granularity of the analysis, it also means that the results of the analysis will serve as a lower bound on the strength of association between applause and polarity. If a significant relationship is observed using the binary response, then any further refinements to direct applause measurement would, in theory, only make the relationship stronger.

Indirect Measurement

The second method for measuring applause is less direct. With it, I attempt to measure what might be called the "applause-worthiness" of a given document. That is, given the makeup of a document, how much do we *expect* applause to occur, relative to other documents? To do this, I need some way of measuring which terms in a document are highly associated with

[18] Another strategy would be to take advantage of applause interruptions, in which an utterance abruptly ends – often with a "- -" or "..." – an applause event occurs, and then the utterance continues. However, the simpler method above seems to work well already.

applause in the larger corpus. Fortunately, a means of doing so was already presented in Section 3.3.1: variance-weighted log odds ratios with informative Bayesian priors. I follow the same method outlined there, but using applause vs. no applause rather than Democratic vs. Republican. And, since Republicans and Democrats certainly don't applaud for the same things, I calculate applause scores for each term, topic, *and party*, rather than just term and topic. Define $\mathcal{A}$ to be the event that applause occurs in a given document, and define $\mathcal{N}$ to be the event that it does not. Then, $LOR^{\mathcal{N}-\mathcal{A}}_{w,z,\rho}$ – the log odds ratio between no applause and applause given term, topic, and party – can be defined as such:

$$LOR^{\mathcal{N}-\mathcal{A}}_{w,z,\rho} = \log\left(\frac{\mathbb{P}(W=w|\mathcal{A},\mathcal{P}=\rho,Z=z)}{1-\mathbb{P}(W=w|\mathcal{A},\mathcal{P}=\rho,Z=z)}\right) - \log\left(\frac{\mathbb{P}(W=w|\mathcal{N},\mathcal{P}=\rho,Z=z)}{1-\mathbb{P}(W=w|\mathcal{N},\mathcal{P}=\rho,Z=z)}\right) \tag{3.26}$$

Again, I segment topics by partitioning the corpus into subsets of documents where the given topic is the most probable. The probabilities are estimated, normalized, and regularized analogously to the methods in Section 3.3.1.

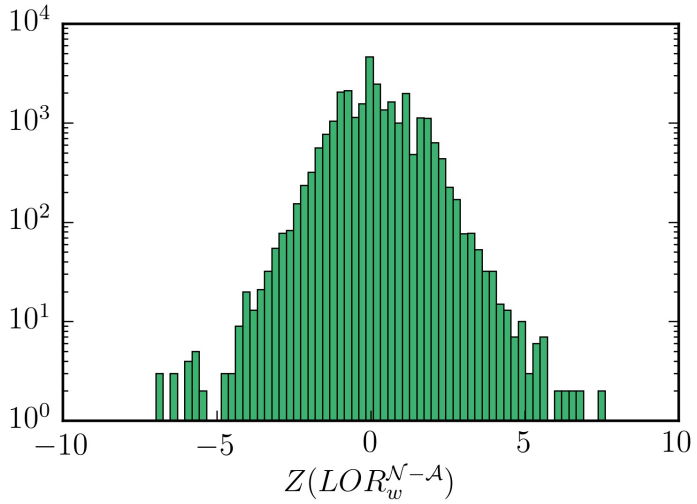
Figure 3.5 shows the distribution of these term applause scores. As before, the scores are shown on a log scale, since most terms are not strongly associated with the presence or absence of applause. Note that, once again, the distribution is symmetric. Thus, as with polarity, I summarize the "applause-worthiness" of a document – which I will denote $Appl^z_d$ – using the sum of its individual term scores:

$$Appl^z_d = \sum_{w=1}^{V} C_{d,w} Z(LOR^{\mathcal{N}-\mathcal{A}}_{w,z,P_d})_{reg} \tag{3.27}$$

3.4.3 Testing the Hypothesis

It is finally time to test the hypothesis which launched this investigation – whether the applause behavior of debate audiences is associated with the polarity of candidates' language. I use two standard statistical tests, logistic and linear regression, for each party and topic. The motivation for using simple tests such as these – rather than, say, training a machine learning model using word polarities as features – is that I am testing a hypothesis of association, not predictiveness. If I had originally posited that topic-specific lexical frame polarity

Figure 3.5: Distribution of $Z(LOR_w^{\mathcal{N}-\mathcal{A}})_{reg}$ scores for each term on a log scale.



would be an accurate *predictor* of applause, then training a model and evaluating it using standard precision, recall, and accuracy measures would be a suitable course of action. However, I would like to understand whether there is a significant *association* between polarity and applause. As such, more traditional statistical significance tests are warranted.

Note that, for all tests, I have excluded the seven debates in which applause is not transcribed as well as all documents in which the proportion of words spoken by a non-candidate is greater than 0.5.

Logistic Regression

Logistic regression is a standard statistical method for testing the association between one or more continuous independent variables and a binary response variable. Here, it is used to test the hypothesis with the direct measure of applause, $HasAppl_d$. For each topic, I consider each document in which that topic is most probable as a data point. I use a single independent variable for each document: Pol_d^z, the sum of the topic-specific lexical frame polarity scores of all of the terms in the document. Since it will be interesting to compare

the results across parties, the set of documents is further split between Democrats and Republicans, and two regressions are performed for every topic. To evaluate significance, I use the p-value of the log likelihood ratio. For goodness of fit, I use McFadden's pseudo-R^2.

Linear Regression

I use simple linear regression to test association using the indirect continuous response variable, $Appl_d^z$. Once again, I segment the documents by most probable topic, then segment them further by party. Again, all regressions use the single independent variable Pol_d^z. I use the p-value of the F-measure for significance and Pearson's R^2 for goodness of fit.

Chapter 4

RESULTS AND DISCUSSION

I will now present and discuss the results of the analysis developed in the previous chapter. I will begin with a presentation of the raw numbers and will then proceed to an in-depth examination of the results. Finally, I will conclude with a discussion of the study's limitations and some possible future directions.

4.1 Results

Recall that the goal of the present analysis is to test my original hypothesis that the topic-specific polarity of debate speech is associated with audience applause. For a given topic, a significant relationship between Pol_d^z and $HasAppl_d^z$ or $Appl_d^z$ (as determined by the results of the logistic and linear regressions, respectively) would offer support for my hypothesis, whereas a non-significant relationship would not. I consider 0.01 as the significance threshold for both tests. That is, I contend that there is a significant relationship between applause and the polarization of lexical frames whenever $p_{llr} < 0.01$ (for logistic regression) or $p_f < 0.01$ (for linear regression). Pol_d^z, $HasAppl_d^z$, and $Appl_d^z$ are calculated according to their definitions in Equation 3.25, Section 3.4.2, and Equation 3.27, respectively.

Tables 4.1 and 4.2 summarize the results of the logistic and linear regressions, respectively, for each party and topic. For each test, three values are reported:

1. n is the number of data points used for each regression. That is, for a given topic z and party ρ, n is the number of documents in which z was the most probable topic that came from ρ's debates.

2. p_{llr} (for logistic regression) and p_f (for linear regression) are the results of the signifi-

cance tests on the model's log likelihood ratio and F-measure, respectively. For aid of visualization, values less than 0.01 are shown as <**0.01**.

3. R^2_{McF} (for logistic regression) and R^2 (for linear regression) measure goodness of fit for each regression model. Again, for aid of visualization, values are cut off at 0.01.

4.2 Interpretation of the Results

In this section, I will begin with some general observations about the results of the regression tests and will then explore in detail some of the more interesting patterns that emerge from the results.

4.2.1 Statistical Significance

Despite the recent (and warranted) criticism of p-values across a wide variety of academic disciplines,[1] I am comfortable concluding that, overall, the values of p_{llr} and p_f (for logistic and linear regression respectively) indicate a significant relationship. Sixteen out of the nineteen topics analyzed displayed a significant relationship between polarity and applause for at least one party and applause measurement. And in both tests, at least half of the topics saw p-values below 0.01 in at least one party. This evidence lends preliminary but strong support to the hypothesis that applause is associated with polarity.

4.2.2 Goodness of Fit

It is also important to evaluate the results in the context of goodness of fit. In statistical regression, goodness of fit tests attempt to quantify how well the response variable is explained by each of the independent variables, or by the entire model. For linear regression, the canonical measure is the coefficient of determination, or R^2, which ranges between 0 and 1. Intuitively, R^2 measures how much of the variance in the response variable can be

[1] See Section 4.3.3 for further discussion on this issue.

Table 4.1: Results of the logistic regression tests for each topic and party. n is the number of documents in the regression, p_{llr} is the p-value of the log likelihood ratio, and R^2_{McF} is McFadden's pseudo-R^2.

		Democrats			Republicans		
Index	Label	n	p_{llr}	R^2_{McF}	n	p_{llr}	R^2_{McF}
1	SocialSec	178	0.365	<0.01	462	0.888	<0.01
2	Courts	150	0.640	<0.01	347	0.893	<0.01
3	Race	370	0.196	<0.01	138	0.097	0.015
4	PrimaryEd	249	0.445	<0.01	158	**<0.01**	0.032
5	MidEast	458	**<0.01**	0.016	884	**<0.01**	0.034
6	DADT	212	0.031	0.020	588	**<0.01**	0.023
7	College	258	0.092	<0.01	164	**<0.01**	0.110
8	Abortion	80	0.026	0.045	408	0.940	<0.01
10	Immigration	175	**<0.01**	0.121	434	**<0.01**	0.019
13	WallSt	226	0.133	<0.01	706	**<0.01**	<0.01
14	Economy	202	0.377	<0.01	528	**<0.01**	0.036
16	Climate	242	0.144	<0.01	177	0.228	<0.01
17	ForeignPol	448	0.681	<0.01	490	**<0.01**	0.030
18	Trade	202	0.729	<0.01	284	**<0.01**	0.029
19	Guns	197	0.084	0.011	127	0.026	0.031
20	CampFin	373	0.062	<0.01	153	0.026	0.025
21	Health	396	**<0.01**	0.017	266	**<0.01**	0.049
23	CrimJust	284	**<0.01**	0.086	157	0.016	0.031
24	IraqWar	339	0.216	<0.01	226	**<0.01**	0.101

Table 4.2: Results of the linear regression tests for each topic and party. n is the number of documents in the regression, p_f is the p-value of the F-measure, and R^2 is Pearson's R^2.

		Democrats			Republicans		
Index	Label	n	p_f	R^2	n	p_f	R^2
1	SocialSec	178	0.968	<0.01	462	<**0.01**	0.036
2	Courts	150	0.262	<0.01	347	<**0.01**	0.057
3	Race	370	0.257	<0.01	138	0.022	0.038
4	PrimaryEd	249	0.483	<0.01	158	<**0.01**	0.055
5	MidEast	458	0.025	0.011	884	<**0.01**	0.131
6	DADT	212	0.026	0.023	588	<**0.01**	0.171
7	College	258	<**0.01**	0.053	164	<**0.01**	0.263
8	Abortion	80	0.700	<0.01	408	0.014	0.015
10	Immigration	175	<**0.01**	0.051	434	0.782	<0.01
13	WallSt	226	<**0.01**	0.169	706	0.016	<0.01
14	Economy	202	<**0.01**	0.037	528	0.032	<0.01
16	Climate	242	0.138	<0.01	177	0.217	<0.01
17	ForeignPol	448	<**0.01**	0.020	490	<**0.01**	0.049
18	Trade	202	<**0.01**	0.055	284	<**0.01**	0.084
19	Guns	197	<**0.01**	0.038	127	0.429	<0.01
20	CampFin	373	<**0.01**	0.048	153	<**0.01**	0.438
21	Health	396	0.389	<0.01	266	<**0.01**	0.179
23	CrimJust	284	<**0.01**	0.183	157	0.458	<0.01
24	IraqWar	339	0.472	<0.01	226	<**0.01**	0.070

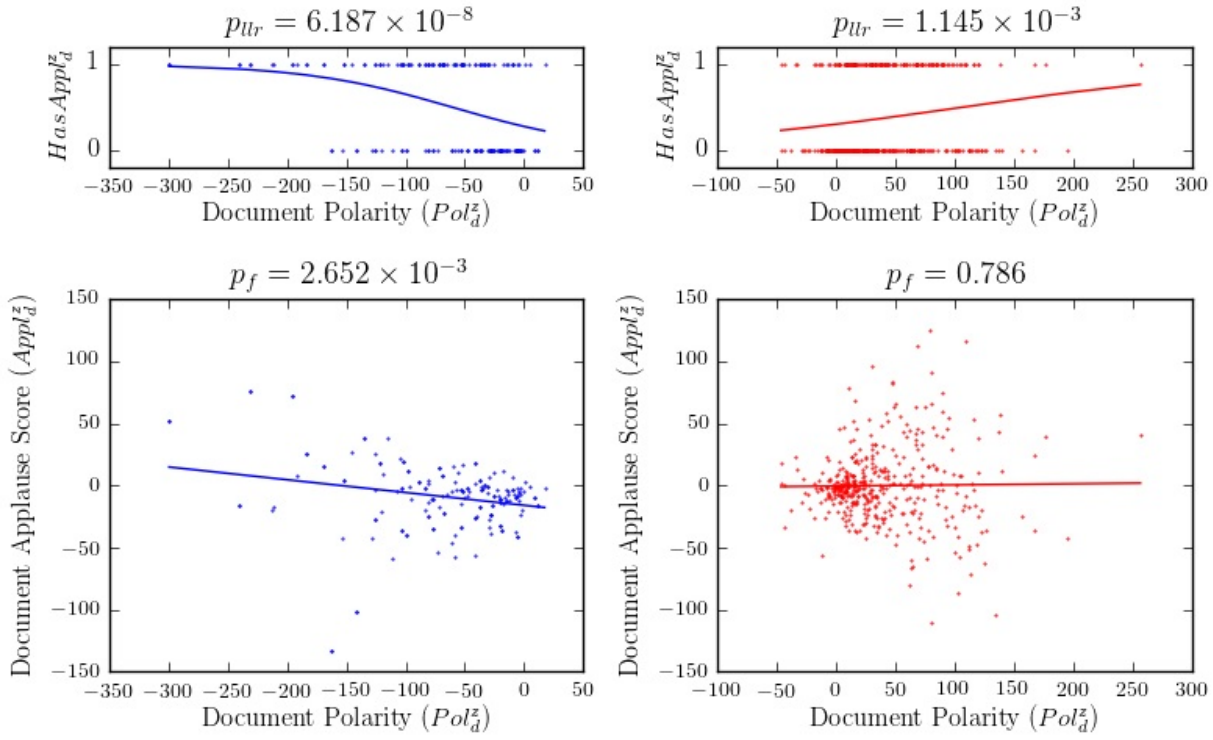
explained by the independent variables. An R^2 value of 0 indicates that the none of the response's variance is explained, and 1 indicates a perfect fit. For logistic regression, there is no analogous measure, but a number of "pseudo-R^2" metrics have been proposed. Here, I use McFadden's pseudo-R^2, R^2_{McF}, which has at least roughly the same intuitive interpretation as canonical R^2 in linear regression (McFadden, 1974, 1978).

Looking at Tables 4.1 and 4.2, it is immediately evident that the goodness of fit values for the regression models are not high. The average value of R^2_{McF} is 0.0248 and the average R^2 is 0.0614. However, this should not be taken to mean that the models are poor, or that the relationship between polarity and applause is not strong. Goodness of fit indicates not how *well* the regressors predict the response, but how *much*. In fact, small R^2 values are expected here, and high values should give us considerable pause. If R^2 were close to 1 for a given topic, that would mean that lexical framing polarity was the *only* factor in determining applause for that topic. This is, of course far from true. Innumerable factors influence whether the audience applauds in a particular context. Aside from linguistic factors that aren't captured here – such as a candidate's intonation and rhetorical structure – applause behavior at a given debate may be influenced by geographical location, the particular point in the nomination process, or even how many candidates are on stage.

4.2.3 Ideological Unity

For some topics, the strength of the relationship between applause and polarity varies substantially between parties. Consider, for example, IMMIGRATION. In linear regression, a significant relationship between applause and polarity is observed for Democrats. However, despite highly polarized, Republican-leaning frames within the topic – like *amnesty* and *border* – the result of the linear regression for Republicans is far from significant. In fact, the p_f value for this regression, 0.782, is the second largest out of all of the 28 linear regressions

Figure 4.1: Results for logistic and linear regression on the IMMIGRATION topic, with Democrats on the left and Republicans on the right.

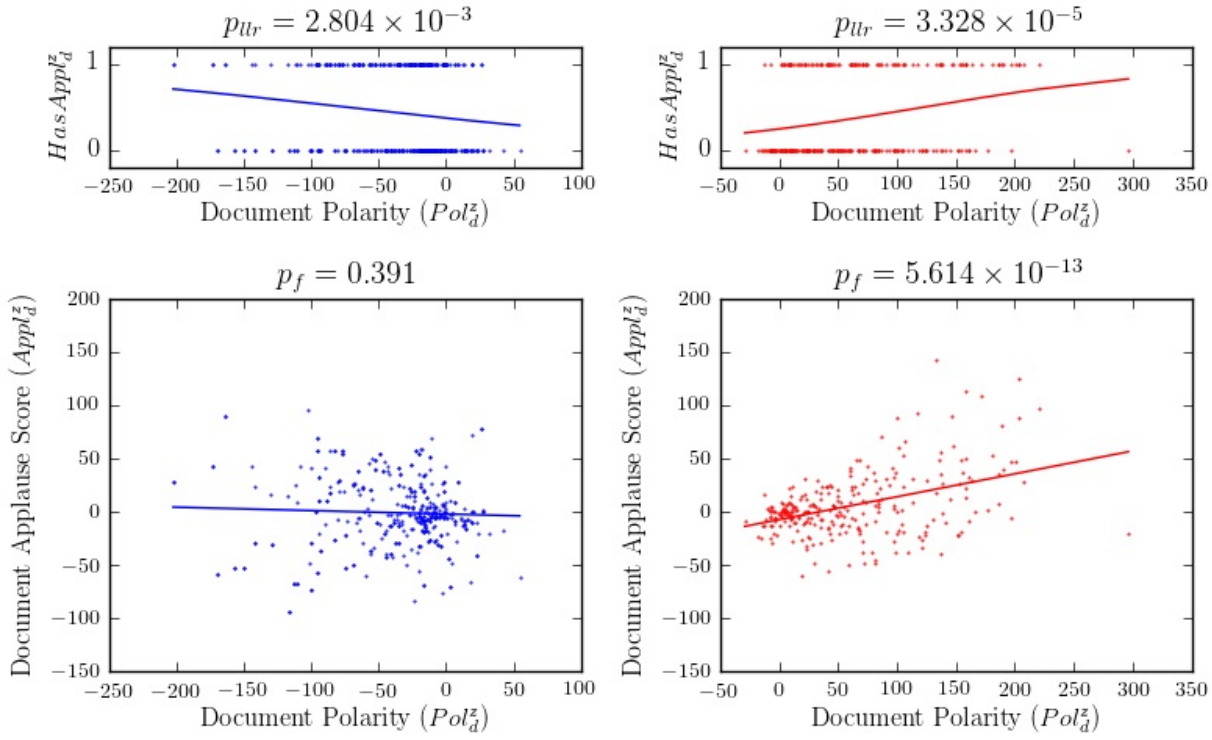


in Table 4.2. Figure 4.1 shows a graphical representation of these results.[2]

One possible explanation of the surprising lack of significance for Republicans may be the diversity of perspectives on immigration within the party. Particularly within the last several election cycles, candidates have expressed a wide array of positions, ranging from calls for mass deportation of illegal immigrants to other, more centrist proposals. Indeed, two prominent Republican candidates in this dataset – Senators Marco Rubio and John McCain – were members of the "Gang of Eight" who were responsible for proposing a bi-

[2]Figures 4.1, 4.2, and 4.3 show the results of the logistic and linear regressions for a selected topic. For logistic regression, the curve in the figure is the fitted logistic function. Likewise for linear regression, the curve is the line of best fit.

Figure 4.2: Results for logistic and linear regression on the HEALTH topic, with Democrats on the left and Republicans on the right.



partisan, comprehensive immigration reform bill in 2013. Among other things, this bill called for a path to citizenship for undocumented workers, a far cry from the more hard-line positions espoused by candidates like Rudy Giuliani, Donald Trump, and others.

This can be contrasted with HEALTH, in which both logistic and linear regression show an extremely significant relationship for Republicans (see Figure 4.2 for a graphical representation). Unlike immigration, healthcare is a topic on which Republicans have presented extreme ideological consistency. Over the past two election cycles in particular, the Republican party has been sharp and unrelenting in its criticism of the Affordable Care Act, and their use of the term *obamacare* in leveling that criticism has provided an excellent exam-

ple of effective lexical framing in action. To illustrate the magnitude of *obamacare*'s effect within the HEALTH topic, I removed it from the dataset completely and reran the polarity analysis. Doing so resulted in two major differences. First, the average party polarity of Republican documents in HEALTH was reduced by nearly 15%. Perhaps more importantly, the relationship between polarity and applause was *no longer significant* for logistical regression once *obamacare* was removed. Thus it seems that the Republican party's ideological unity – and the consistency with which they presented that unity – is a major factor in driving the association between applause and polarity for this topic.

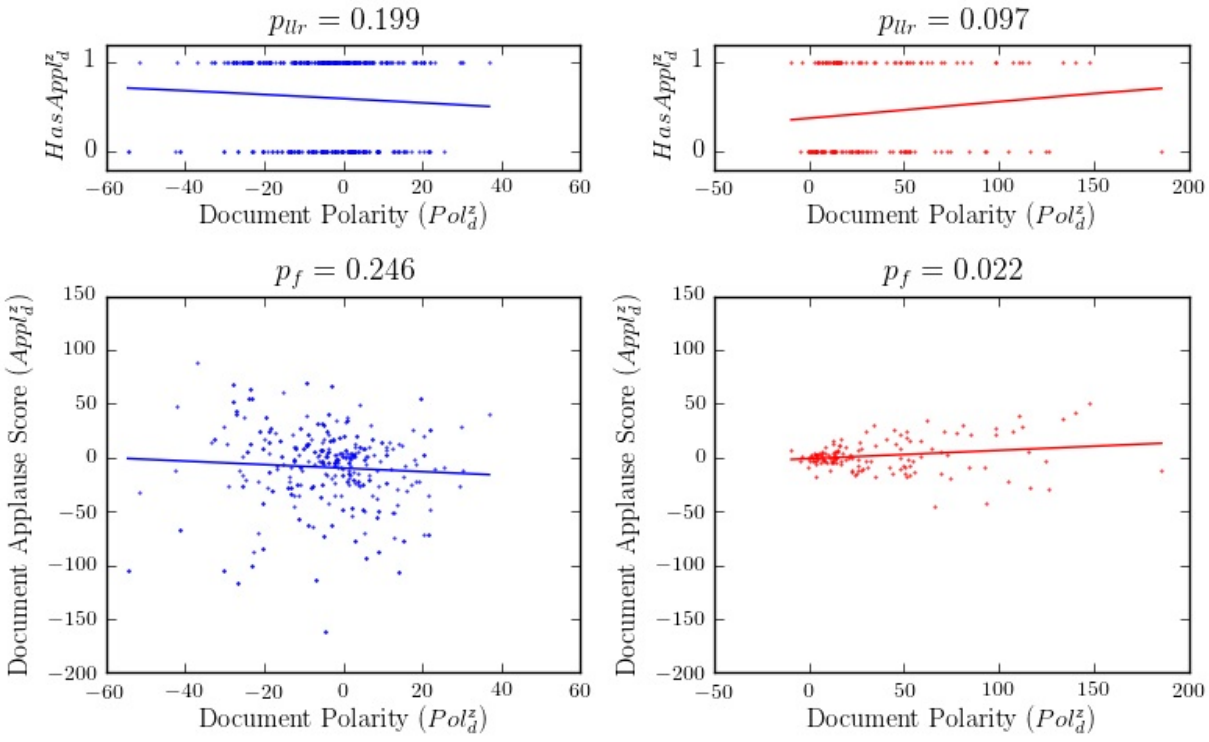
4.2.4 Topic Ownership

In Section 3.2.6, I used an estimation of the "party one-sidedness" for a given topic merely as a means to weed out models with bad topic pairs. However, the one-sidedness of a topic is also valuable information in and of itself. In fact, a major focus of computational political science is the characterization and analysis *topic ownership* (Tsur et al., 2015). One way of understanding topic ownership is by looking at how much each party talks about a given issue. If most of the discussion on a topic comes from a single party, that party is said to "own" the topic. My analysis of polarity and applause did not address topic ownership directly, but some of its effects can be seen in analyzing the results.

The Democratic party is commonly closely associated with issues of identity politics. The data used in this analysis bear this out. Looking at the RACE topic in Tables 4.1 and 4.2, n is nearly three times as large for Democrats than for Republicans. In addition, no significant results are observed in either regression test for either party (Figure 4.3 shows a graphical representation). I posit that Democrats' ownership of the RACE topic might be a possible explanation for this.

For one, the data imbalance results in poor quality polarity measurement for Republicans. The top five most polar lexical frames for Republicans in this topic are *hispanic*, *hispanics*, *hispanic_community*, *bigotry*, and *moms*. These are reminiscent of the incidentally party-specific frames discovered by the naive log-odds ratio measurement in Section 3.3.1.

Figure 4.3: Results for logistic and linear regression on the RACE topic, with Democrats on the left and Republicans on the right.



Poor quality of polarity scores, however, is certainly not the only contributing factor. Consider another imbalanced topic, ABORTION. Here, n is more than five times larger for Republicans than for Democrats. Again, no significant results are observed in any of the regressions. However, as shown in Table 3.13 in the previous chapter, the lexical frames for ABORTION are coherent and logical for both parties. Why then, considering the strong partisan divisions and emotional weight of this topic, is there no significant association between polarity and applause? Perhaps the debate on abortion is so long-standing that the traditional partisan positions on the issue are *expected*, not applauded. Or perhaps it is simply not socially appropriate to applaud in the context of such a sensitive and emotional issue. Whatever the reason, it would no doubt be interesting to explore this question from other perspectives based on the results achieved here.

4.3 Limitations

The conclusions made in this analysis are limited by a number of factors which bear mentioning here. Broadly, they fall into three categories: limitations related to the dataset, the modeling process, and the actual hypothesis testing.

4.3.1 Data Limitations

The dataset for this study was constructed using news transcripts from the debates. Besides the data preprocessing challenges already mentioned, using this particular dataset imposed two important limitations on the scope of the analysis:

1. Since these transcripts were not created with the purpose of measuring applause, I cannot be entirely confident that every applause that occurred was transcribed. Further, as noted in Section 3.4.1, strict chronological order was sometimes preferred in transcribing applause events, as opposed to placing them directly after the applause-causing utterance.

2. At several points, I have addressed the homogeneity of applause annotations. Since

the dataset provided no information by which applauses could be categorized, I simply considered whether an applause occurred or did not. This is, of course, not the full story. Applauses come in many forms, from polite clapping to full-throated cheers. Obtaining and measuring acoustic information for applause events – especially loudness and duration – would have allowed a more nuanced analysis.

An additional limitation is related not to how the data were presented, but to the nature of the data itself. Throughout this analysis, I have considered applause to be a measure of positive response within the context of a presidential primary debate, and I have found there to be a significant association between polarity and applause for a substantial number of topics. It is tempting to extend these findings beyond presidential primary debate audiences and claim that they apply to the supporters of each party in general. However, without evidence to suggest that the behavior of debate audiences is representative of their parties as a whole, such a claim is entirely unwarranted. Thus, the generalizability of my conclusions is limited.

4.3.2 Modeling Limitations

Despite efforts made in Section 3.2 to ameliorate its effect, the fact that LDA and related generative topic modeling methods are considerably unstable cannot be ignored. Relying on such methods to draw conclusions about topic-specific phenomena is far from ideal.

I also made a number of simplifying assumptions in structuring the modeling phase of the analysis. As discussed briefly in Section 2.3, a single left-right axis is certainly not a complete representation of the ideological spectrum of American politics (see Djemili et al. (2014) for an attempt to describe positions on this spectrum more precisely). Relatedly, partitioning the data simply along party lines – the division on which the polarity modeling and association measurements were predicated – may have been overly simplistic. One could easily argue – and with good cause – that certain candidates, or even certain debates, should be considered outliers for the purposes of an analysis of left-right polarization, or that the

ideological positions of each candidate should be considered as a factor in the modeling process.

In Sections 3.3.1 and 3.4.2, I presented a method for calculating log odds ratios for each topic and term that made use of the most probable topic in each document. While existing literature supports this approach, it throws away the rich probabilistic representation provided by LDA's posteriors, particularly topic distributions over documents. Additionally, no such distributional information is available for applause, and so I could only resort to partitioning the data into two groups depending on whether the document contained an applause event or did not. This resulted in a coarse measurement of the "applause affinity" of each term for each topic and party. A more refined analysis would be more precise in this measurement. One approach might be to consider each applause event as being caused by the tokens preceding it, assigning each successively distant token a decreasing "probability of contribution." Given this representation, one could generalize to an applause distribution over the vocabulary, after which any number of probabilistic measures could be used to quantify the association between terms and applause. How these distributions over tokens and vocabulary would be generated, though, remains unresolved.

4.3.3 Analysis Limitations

The use of p-values for significance testing has recently come under heavy criticism from the scientific community at large, based on the arbitrariness of significance thresholds, the inability to interpret p-values intuitively, and widespread evidence of "p-hacking" – intentional or otherwise – in a variety of disciplines (Head et al., 2015). In light of these discussions, it would perhaps have been wiser for me to assess significance in a different way.

4.4 Future Directions

While the analysis presented here seems to lend strong support to an association between applause and polarity, the establishment of this relationship is only the first step towards a more complete understanding of the effect of polarized language in the political arena. A

number of potential avenues of future investigation present themselves.

In Section 3.2.6, I presented a method for evaluating whether a model had learned party-specific versions of certain topics. While this was certainly helpful in the model selection phase, it would have been more useful to integrate this penalty directly into the training of the topic model. Future work could address this by designing a variant of LDA that penalizes models which have low inter-topic distance with respect to some arbitrary response variable as defined by the experimenter.

A number of times in the course of this analysis, I have touched on the idea of topic ownership, which is a well-studied phenomenon in the political science literature (Tsur et al., 2015). Especially based on the analysis in Section 4.2.4, a fuller treatment of the effect of topic ownership on both polarity and applause would no doubt be interesting.

Another potential direction involves more linguistically informed measures of framing. Framing, as noted in the background discussion on that topic, is a complex linguistic phenomenon. It is an active area of research in the linguistics community, and its many dimensions are still only roughly defined (Chong and Druckman, 2007). In this analysis, I have focused on lexical framing, which looks at framing through the narrow lens of individual words and phrases. But framing manifests itself in all levels of linguistic representation, from phonetics to syntax to discourse structure. Preliminary computational studies indicate that people on different sides of the death penalty debate use distinct syntactic frames when discussing the issue (Greene, 2007). That is, particular tree structures can be identified that are predictive of someone's position on the issue. Exploring other types of framing in political debate could be a promising way of refining the present analysis.

Another potential improvement to the framing portion of the analysis would be to integrate classic positive/negative sentiment analysis in the measurement of party polarity. The disparaging sentiment of the *obamacare* frame especially, taken in combination with its substantial influence on the results for the HEALTH topic, suggests that sentiment could play an important role in whether or not a given utterance generates applause.

Finally, further study is needed to accurately place this analysis in the larger context of

American politics. As I noted in the previous section, the generalizability of my conclusions is limited by the fact that the applause behavior of debate audiences is not necessarily representative of the voting public at large. This gap could be addressed in a number of ways, including correlating national opinion polls with audience applause behavior, or determining in what kinds of situations applause is truly an expression of support. One potential avenue for addressing the latter would be to correlate the applause measurements from this study with previous analyses of political debates where audience response has been measured more finely, such as in Boydstun et al. (2014).

Chapter 5

CONCLUSION

Polarization in the American political system is at its highest in recent history and, at least as of this publication, shows few signs of abatement (Pew Research Center, 2014). It is starkly reflected in the political ideologies of citizens, the voting patterns of the politicians who represent them, and even in the media that members of each party choose to consume (Mitchell et al., 2014). In this work, I have explored the linguistic aspects of this phenomenon, examining the effects of polarized language in presidential primary debates through the lens of lexical framing.

I hypothesized that, given a particular topic, debate audiences would be more likely to applaud for rhetoric that was polarized with respect to that topic. Using a corpus of 104 Republican and Democratic presidential primary debates since 2000, I developed an analysis in three phases to test this hypothesis. First, I generated a topic model over the corpus, developing a novel model evaluation metric in the process. I then presented a technique to automatically quantify the polarity of topic-specific lexical frames. Finally, after presenting two methods for measuring applause, I used standard statistical association measures to test the hypothesis.

The results of the analysis showed a significant relationship between polarity and applause for a substantial number of topics in the model, lending preliminary but strong support to my original hypothesis. In reviewing the results, I observed a number of interesting patterns. First, ideological unity appeared to play a role in the establishment and effectiveness of polarized lexical frames. Despite highly polarized language, Republicans' ideological fragmentation on immigration may have been a factor in the lack of a significant relationship between polarity and applause. On the other hand, in healthcare, their unified criticism

of the Affordable Care Act – and their use of the lexical frame *obamacare* in particular – contributed strongly to the significance of the relationship in that topic.

An interesting pattern concerning topic ownership emerged as well. When the discussion of an issue was particularly one-sided, such as racial politics for Democrats or abortion for Republicans, significant associations between polarity and applause were not as likely to be present. In some cases, this may have reflected a limitation of my polarity quantification method. Other times, however, the lexical frames discovered were logical and coherent, but still no significant relationship was found.

These and other observations also suggest a number of potential avenues for future work, including more robust topic modeling, incorporation of classic positive/negative sentiment analysis, and a fuller treatment of the effect of topic ownership and agenda setting.

In this work, I have attempted to shed some light on the linguistic phenomena at play in polarized political rhetoric. And, while I am wary of using the conclusions made here to speculate about larger phenomena, the primary finding – that debate audiences are more likely to applaud for polarized language – suggests a vicious cycle that may be contributing to America's increasing political divisiveness. It may be, when voters react positively to polarized language, that politicians, pundits, and others respond with even more polar rhetoric, which in turn heightens the intensity of voters' reactions. The increasing isolation of the "filter bubbles" of television and social media could be interpreted as a manifestation of this cycle, as could the gradually shrinking center of the American political spectrum.

This work stands as only the latest entry in an ancient and ever-growing body of literature concerning political rhetoric and its effects on the voting public. I hope the conclusions presented here encourage further study in this area, especially as the extreme polarization of America's political climate continues.

BIBLIOGRAPHY

Clio Andris, David Lee, Marcus J. Hamilton, Mauro Martino, Christian E. Gunning, and John Armistead Selden. The rise of partisanship and super-cooperators in the U.S. House of Representatives. *PloS One*, 10(4), 2015.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. URL `http://www.jmlr.org/papers/v3/blei03a.html`.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7, 2010. URL `http://www.cs.princeton.edu/~blei/papers/BleiGriffithsJordan2009.pdf`.

Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Biennial GSCL Conference*, volume 156, 2009.

Jordan Boyd-Graber, David Mimno, and David Newman. *Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements*. CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, Florida, 2014.

Amber E. Boydstun, Rebecca A. Glazier, Mathew T. Pietryka, and Philip Resnik. Real-time reactions to a 2012 presidential debate: A method for understanding which messages matter. *Public Opinion Quarterly*, 78(S1):330–343, June 2014. URL `http://poq.oxfordjournals.org/cgi/content/long/78/S1/330`.

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

Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. A density-based method for adaptive LDA model selection. *Neurocomputing*, 72(7-9):1775–1781, March 2009. URL http://dx.doi.org/10.1016/j.neucom.2008.06.011.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China, July 2015. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P15-2072.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. Reading tea leaves: How humans interpret topic models. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc., 2009. URL http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf.

Dennis Chong and James N. Druckman. Framing theory. *Annual Review of Political Science*, 10(1):103–126, 2007. URL http://faculty.wcas.northwestern.edu/~jnd260/pub/Chong%20Druckman%20Annual%20Review%202007.pdf.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

Sarah Djemili, Julien Longhi, Claudia Marinica, Dimitris Kotzinos, and Georges-Elia Sarfati. What does Twitter have to say about ideology? In *NLP 4 CMC: Natural Language Processing for Computer-Mediated Communication/Social Media – Pre-conference workshop at KONVENS 2014*, volume 1, pages 16–25. Universitätsverlag Hildesheim, 2014.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

Milagros Fernández-Gavilanes, Tamara Álvarez-López, Jonathan Juncal-Martínez, Enrique Costa-Montenegro, and Francisco Javier González-Castaño. Unsupervised method for sentiment analysis in online texts. *Expert Systems with Applications*, 58(C):57–75, October 2016. URL `https://doi.org/10.1016/j.eswa.2016.03.031`.

Sean Gerrish and David M. Blei. Predicting legislative roll calls from text. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 489–496, 2011.

Erving Goffman. *Frame Analysis: An Essay on the Organization of Experience*. Northeastern, 1974.

Stephan Greene. *Spin: Lexical Semantics, Transitivity, and the Identification of Implicit Sentiment*. PhD thesis, University of Maryland, 2007. URL `http://drum.lib.umd.edu/bitstream/handle/1903/7293/umi-umd-4694.pdf?sequence=1&isAllowed=y`.

Stephan Greene and Philip Resnik. More than words: Syntactic packaging and implicit sentiment. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 503–511, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N/N09/N09-1057.pdf`.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

Justin Grimmer. A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in senate press releases. *Political Analysis*, 18(1):1–35, 2010. URL `http://www.jstor.org/stable/25791991`.

Justin Grimmer and Brandon M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, January 2013. URL `http://pan.oxfordjournals.org/content/21/3/267.abstract`.

Megan L. Head, Luke Holman, Rob Lanfear, Andrew T. Kahn, and Michael D. Jennions. The extent and consequences of p-hacking in science. *PLoS Biololgy*, 13(3), 2015.

John Heritage and David Greatbatch. Generating applause: A study of rhetoric and response at party political conferences. *American Journal of Sociology*, 92(1):110–157, 1986.

Mohit Iyyer, Peter Enns, Jordan Boyd-Graber, and Philip Resnik. Political ideology detection using recursive neural networks. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1113–1122, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P14-1105`.

Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Pearson Education, Inc., 2nd edition, 2009.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, March 1951. URL `http://dx.doi.org/10.1214/aoms/1177729694`.

George Lakoff and Mark Johnson. *Metaphors We Live By*. The University of Chicago Press, 1980.

Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden, April 2014. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/E14-1056`.

Lillian Lee. On the effectiveness of the skew divergence for statistical language analysis. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, pages 65–72, 2001.

Jianhua Lin. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

Jon D. Mcauliffe and David M. Blei. Supervised topic models. In *Advances in Neural Information Processing Systems 21*, pages 121–128, 2008. URL `http://papers.nips.cc/paper/3328-supervised-topic-models.pdf`.

Daniel L. McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, 1974.

Daniel L. McFadden. Quantitative methods for analyzing travel behaviour of individuals: Some recent developments. In David Hensher and David Stopher, editors, *Behavioral Travel Modeling*, pages 279–318. Croom Helm London, 1978.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. Political polarization & media habits. *Pew Research Center, Washington, D.C.*, October 2014.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, pages 372–403, 2009.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1857999.1858011`.

Viet-An Nguyen. *Guided Probabilistic Topic Models for Agenda-Setting and Framing*. PhD thesis, University of Maryland, 2015. URL `http://drum.lib.umd.edu/bitstream/handle/1903/16600/Nguyen_umd_0117E_16056.pdf?sequence=1&isAllowed=y`.

Viet-An Nguyen, Jordan L. Boyd-Graber, and Philip Resnik. Lexical and hierarchical topic regression. In Christopher J. C. Burges, Leon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1106–1114. Curran Associates, Inc., 2013. URL `http://papers.nips.cc/paper/5163-lexical-and-hierarchical-topic-regression.pdf`.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. Tea Party in the House: A hierarchical ideal point topic model and its application to Republican legislators in the 112th Congress. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1438–1448, Beijing, China, July 2015. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P15-1139`.

Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the*

Seventh International Conference on Language Resources and Evaluation (LREC '10), Valletta, Malta, May 2010. European Language Resources Association (ELRA).

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Gerhard Peters and John T. Woolley. Presidential candidates debates: Democratic presidential candidates debate in Los Angeles, California. Online, March 2000. URL `http://www.presidency.ucsb.edu/ws/?pid=105446`.

Gerhard Peters and John T. Woolley. Presidential candidates debates: Democratic presidential candidates debate in Manchester, New Hampshire. Online, January 2004. URL `http://www.presidency.ucsb.edu/ws/?pid=74351`.

Gerhard Peters and John T. Woolley. Presidential candidates debates: Republican candidates debate in Simi Valley, California. Online, September 2015. URL `http://www.presidency.ucsb.edu/ws/?pid=110756`.

Gerhard Peters and John T. Woolley. Presidential candidates debates: Democratic candidates debate in Charleston, South Carolina. Online, January 2016a. URL `http://www.presidency.ucsb.edu/ws/?pid=111409`.

Gerhard Peters and John T. Woolley. Presidential candidates debates: Republican candidates debate in Detroit, Michigan. Online, March 2016b. URL `http://www.presidency.ucsb.edu/ws/?pid=111711`.

Gerhard Peters and John T. Woolley. Presidential candidates debates: Republican candidates debate in North Charleston, South Carolina. Online, January 2016c. URL http://www.presidency.ucsb.edu/ws/?pid=111395.

Pew Research Center. Political polarization in the American public. *Pew Research Center, Washington, D.C.*, June 2014.

Campbell Robertson. Millions on election day make a different decision: Not voting. *The New York Times*, November 2016. URL https://www.nytimes.com/2016/11/09/us/politics/voter-turnout.html.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, July 1948.

John C. Sinclair and Michael B. Bracken. Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology*, 47(8):881–889, 1994.

Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961. Association for Computational Linguistics, 2012.

Yee W. Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Sharing clusters among related groups: Hierarchical Dirichlet processes. In Lawrence K. Saul, Yair Weiss, and Leon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 1385–1392. MIT Press, 2005. URL http://papers.nips.cc/paper/2698-sharing-clusters-among-related-groups-hierarchical-dirichlet-processes.pdf.

Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 Conference*

on Empirical Methods in Natural Language Processing, pages 327–335. Association for Computational Linguistics, 2006.

Oren Tsur, Dan Calacci, and David Lazer. A frame of mind: Using statistical models for detection of framing and agenda setting campaigns. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1629–1638, Beijing, China, July 2015. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P15-1157`.

Hanna M. Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112. ACM, 2009.

Appendix A

LIST OF STOP WORDS

For standard English stop words, I used the stop list provided with `scikit-learn` for Python. The list is available at `https://github.com/scikit-learn/scikit-learn/blob/master/sklearn/feature_extraction/stop_words.py`. This appendix contains the complete list of names and offices that were also considered stop words in the preprocessing of the corpus.

al	bret	clinton	edwards	gregory
alan	brian	cokie	errol	griffith
albin	brien	congressman	fahey	gwen
ambassador	brit	congresswoman	ferrechio	harris
anderson	brokaw	cooney	fiorina	harwood
andrea	brown	cooper	forbes	hatch
arraras	brownback	cordes	fred	hemmer
bachman	bush	crowley	garrett	herman
bachmann	byron	cruz	gary	hewitt
baier	cain	cummings	geha	hillary
baker	cameron	dana	george	holt
barack	campbell	david	gerald	hook
bartiromo	candy	dean	gerard	howard
bash	carl	demint	gibson	huckabee
bauer	carly	dennis	gilmore	hugh
becky	carolyn	diane	gingrich	hume
ben	carson	diaz-balart	giuliani	hunter
bernard	cavuto	dickerson	gloria	huntsman
bernie	chafee	dinan	goldman	ifill
biden	charlie	distaso	goler	inskeep
bill	chris	dodd	gore	jake
blitzer	christie	don	governor	janet
bobby	christopher	donald	graham	jeanne
borger	chuck	doyle	gravel	jeb
bradley	clark	duncan	greenfield	jeff

jennings
jim
jindal
joe
john
johns
johnson
jon
jorge
jose
josh
juan
judy
juliana
karen
kasich
kathie
keith
kelly
kerry
kevin
keyes
kimberly
king
koppel
kucinich
larry
lemon
lester
lieberman
lincoln
lindsey
lopez
louis
maccallum
maddow
major
malley
malveaux
marco
maria
martha
martin
matthews
mccain
mcelveen
mcmanus
megyn
michele
mike
mitchell
mitt
morales
muir
nancy
natalie
neil
newt
norris
o'brien
o'malley
obama
obradovich
olbermann
orrin
pataki
paul
pawlenty
pelley
perry
peter
president
quick
quintanilla
rachel
raddatz
ramos
rand
regan
richardson
rick
robert
roberts
romney
ron
rose
rubio
rudy
russert
ryerson
salinas
sam
sanders
sandra
santorum
sawyer
scott
secretary
seib
senator
sharpton
shaw
siegel
smiley
smith
soledad
speaker
spradling
stanton
stephanopoulos
stephen
steve
strassel
susan
suzanne
tancredo
tapper
tash
tavis
ted
thompson
tim
todd
tom
tommy
trish
trump
tumulty
vandehei
vice
walker
wallace
washburn
webb
wendell
wesley
williams
wolf
woodruff
yepsen
york