Winning on the Merits: The Joint Effects of Content and Style on Debate Outcomes

Lu Wang[1] **Nick Beauchamp**[2,3] **Sarah Shugars**[3] **Kechen Qin**[1]
[1]College of Computer and Information Science, Northeastern University
[2]Department of Political Science, Northeastern University
[3]Network Science Institute, Northeastern University
luwang@ccs.neu.edu, n.beauchamp@northeastern.edu
{qin.ke,shugars.s}@husky.neu.edu

Abstract

Debate and deliberation play essential roles in politics and government, but most models presume that debates are won mainly via superior style or agenda control. Ideally, however, debates would be won on the merits, as a function of which side has the stronger arguments. We propose a predictive model of debate that estimates the effects of linguistic features and the latent persuasive strengths of different topics, as well as the interactions between the two. Using a dataset of 118 Oxford-style debates, our model's combination of content (as latent topics) and style (as linguistic features) allows us to predict audience-adjudicated winners with 74% accuracy, significantly outperforming linguistic features alone (66%). Our model finds that winning sides employ stronger arguments, and allows us to identify the linguistic features associated with strong or weak arguments.

1 Introduction

What determines the outcome of a debate? In an ideal setting, a debate is a mechanism for determining which side has the better arguments and also for an audience to reevaluate their views in light of what they have learned. This ideal vision of debate and deliberation has taken an increasingly central role in modern theories of democracy (Habermas, 1984; Cohen, 1989; Rawls, 1997; Mansbridge, 2003). However, empirical evidence has also led to an increasing awareness of the dangers of style and rhetoric in biasing participants towards the most skillful, charismatic, or numerous speakers (Noelle-Neumann, 1974; Sunstein, 1999).

Motion: *Abolish the Death Penalty*
• Argument 1 (PRO): What is the error rate of convicting people that are innocent? ...when you look at capital convictions, you can demonstrate on innocence grounds a 4.1 percent error rate, 4.1 percent error rate. I mean, would you accept that in flying airplanes? I mean, really. ...
• Argument 2 (CON): ... The risk of an innocent person dying in prison and never getting out is greater if he's sentenced to life in prison than it is if he's sentenced to death. So the death penalty is an important part of our system.
• Argument 3 (PRO): ...I think if there were no death penalty, there would be many more resources and much more opportunity to look for and address the question of innocence of people who are serving other sentences.

Figure 1: In this segment from a debate over abolishing the death penalty, argument 1 is identified as having the linguistic features 'questions' and 'numerical evidence', while arguments 2 and 3 use 'logical reasoning.' Our model infers that both pro arguments are intrinsically "strong," while the con argument is "weak", although arguments 2 and 3 both use 'reasoning' language.

In light of these concerns, most efforts to predict the persuasive effects of debate have focused on the linguistic features of debate speech (Katzav and Reed, 2008; Mochales and Moens, 2011; Feng and Hirst, 2011) or on simple measures of topic control (Dryzek and List, 2003; Mansbridge, 2015; Zhang et al., 2016). In the ideal setting, however, we would wish for the winning side to win based on the strength and merits of their arguments, not based on their skillful deployment of linguistic style. Our model therefore predicts debate outcomes by modeling not just the persuasive effects of directly observable linguistic features, but also by incorporating the inherent, latent strengths of topics and issues specific to each side of a debate.

To illustrate this idea, Figure 1 shows a brief ex-

Transactions of the Association for Computational Linguistics, vol. 5, pp. 219–232, 2017. Action Editor: Noah Smith.
Submission batch: 9/2016; Revision batch: 12/2016; Published 7/2017.

change from a debate about the death penalty. Although the arguments from both sides are on the same subtopic (the execution of innocents), they make their points with a variety of stylistic maneuvers, including rhetorical questions, factual numbers, and logical phrasing. Underlying these features is a shared content, the idea of the execution of innocents. Consistent with the work of Baumgartner et al. (2008), this subtopic appears to inherently support one side – the side opposed to the death penalty – more strongly than the other, *independent of stylistic presentation*. We hypothesize that within the overall umbrella of a debate, some topics will tend to be inherently more persuasive for one side than the other, such as the execution of innocents for those opposed to the death penalty, or the gory details of a murder for those in favor of it. Strategically, debaters seek to introduce topics that are strong for them and weaker for their opponent, while also working to craft the most persuasive stylistic delivery they can. Because these stylistic features themselves vary with the inherent strength of topics, we are able to predict the latent strength of topics even in new debates on entirely different issues, allowing us to predict debate winners with greater accuracy than before.

In this paper, then, we examine *the latent persuasive strength of debate topics, how they interact with linguistic styles, and how both predict debate outcomes*. Although the task here is fundamentally predictive, it is motivated by the following substantive questions: How do debates persuade listeners? By merit of their content and not just their linguistic structure, can we capture the sense in which debates are an exchange of arguments that are strong or weak? How do these more superficial linguistic or stylistic features interact with the latent persuasive effects of topical content? Answering these questions is crucial for modern theories of democratic representation, although we seek only to understand how these features predict audience reactions in the context of a formal debates where substance perhaps has the best chance of overcoming pure style. We discuss in detail the relevance of our work to existing research on framing, agenda setting, debate, persuasion, and argument mining in § 6.

Here, we develop a joint model that simultaneously 1) infers the latent persuasive strength inherent in debate topics and how it differs between opposing sides, and 2) captures the interactive dynamics between topics of different strength and the linguistic structures with which those topics are presented. Experimental results on a dataset of 118 Oxford-style debates show that our topic-aware debate prediction model achieves an accuracy of 73.7% in predicting the winning side. This result is significantly better than a classifier trained with only linguistic features (66.1%), or using audience feedback (applause and laughter; 58.5%), and significantly outperforms previous predictive work using the same data (Zhang et al., 2016) (73% vs. 65%). This shows that the inherent persuasive effect of argument content plays a crucial role in affecting the outcome of debates.

Moreover, we find that winning sides are more likely to have used inherently strong topics (as inferred by the model) than losing sides (59.5% vs. 54.5%), a result echoed by human ratings of topics without knowing debate outcomes (44.4% vs. 30.1%). Winning sides also tend to shift discussion to topics that are strong for themselves and weak for their opponents. Finally, our model is able to identify linguistic features that are specifically associated with strong or weak topics. For instance, when speakers are using inherently stronger topics, they tend to use more first person plurals and negative emotion, whereas when using weaker arguments, they tend to use more second person pronouns and positive language. These associations are what allow us to predict topic strength, and hence debate outcomes, out of sample even for debates on entirely new issues.

2 Data Description

This study uses transcripts from Intelligence Squared U.S. (IQ2) debates.[1] Each debate brings together panels of renowned experts to argue for or against a given issue before a live audience. The debates cover a range of current issues in politics and policy, and are intended to "restore civility, reasoned analysis, and constructive public discourse." Following the Oxford style of debate, each side is given a 7-minute opening statement. The moderator then begins the discussion phase, allowing questions from the audience and panelists, followed by

[1] http://intelligencesquaredus.org

2-minute closing statements.

The live audience members record their pre- and post-debate opinions as PRO, CON, or UNDECIDED relative to the resolution under debate. The results are shared only after the debate has concluded. According to IQ2, the winning side is the one that gains the most votes after the debate. 118 debate transcripts were collected, with PRO winning 60.[2] 84 debates had two debaters on each side, and the rest had three per side. Each debate contains about 255 speaking turns and 17,513 words on average.[3]

These debates are considerably more structured and moderated, and have more educated speakers and audience members, than one generally finds in public debates. As such, prediction results of our model may vary on other types of debates. Meanwhile, since we do not focus on formal logic and reasoning structure, but rather on the intrinsic persuasiveness of different topics, it may be that the results here are more general to all types of argument. Answering this question depends on subsequent work comparing debates of varying degrees of formality.

3 The Debate Prediction Model

We consider debate as a process of argument exchange. Arguments have content with inherent (or exogenously determined) persuasive effects as well as a variety of linguistic features shaping that content. We present here a debate outcome prediction model that combines directly observed linguistic features with latent persuasive effects specific to different topical content.

3.1 Problem Statement

Assume that the corpus D contains N debates, where $D = \{d_i\}_{i=1}^{N}$. Each debate d_i comprises a sequence of arguments, denoted as $\mathbf{x}_i = \{x_{i,j}\}_{j=1}^{n_i}$, where n_i is the number of arguments. For the present purposes, an argument is a continuous unit of text on the same topic, and may contain multiple sentences within a given turn (see Figure 1). We use $\mathbf{x}_i^p$ and $\mathbf{x}_i^c$ to denote arguments for PRO and CON. The outcome for debate d_i is $y_i \in \{1, -1\}$, where 1 indicates PRO wins and -1 indicates CON wins.

We assume that each debate d_i has a topic system, where debaters issue arguments from K topics relevant to the motion (K varies for different debates). Each topic has an intrinsic persuasion strength which may vary between sides (e.g. a discussion of innocent convicts may intrinsically help the anti-death-penalty side more than the pro). Thus we have a *topic strength system* $\mathbf{h}_i = \{\mathbf{h}_i^p, \mathbf{h}_i^c\}$, where the strengths for K topics are $\mathbf{h}_i^p = \{h_{i,k}^p\}_{k=1}^{K}$ for PRO, and $\mathbf{h}_i^c = \{h_{i,k}^c\}_{k=1}^{K}$ for CON. Topic strength $h_{*,*}^*$ is a categorical variable in $\{\text{STRONG}, \text{WEAK}\}$.[4] Neither the topics nor their strength are known *a priori*, and thus must be inferred.

For debate d_i, we define $\Phi(\mathbf{x}_i^p, \mathbf{h}_i)$ and $\Phi(\mathbf{x}_i^c, \mathbf{h}_i)$ to be feature vectors for arguments from PRO and CON. We first model and characterize features for each argument and then aggregate them by side to predict the *relative success* of each side. Therefore, the feature vectors for a side can be formulated as the summation of feature vectors of its arguments, i.e. $\Phi(\mathbf{x}_i^p, \mathbf{h}_i) = \sum_{x_{i,j} \in \mathbf{x}_i^p} \phi(x_{i,j}, \mathbf{h}_i)$, and $\Phi(\mathbf{x}_i^c, \mathbf{h}_i) = \sum_{x_{i,j} \in \mathbf{x}_i^c} \phi(x_{i,j}, \mathbf{h}_i)$, where $\phi(x_{i,j}, \mathbf{h}_i)$ is the feature vector of argument $x_{i,j}$.[5]

Each argument feature in $\phi(x_{i,j}, \mathbf{h}_i)$ combines a stylistic feature directly observed from the text with a latent strength dependent on the topic of the argument. For instance, consider an argument $x_{i,j}$ of a topic with an inferred strength of STRONG and which contains 3 usages of the word "you". Then $x_{i,j}$ has two coupled topic-aware features of the form $\phi_{M(\texttt{feature},\texttt{strength})}(x_{i,j}, \mathbf{h}_i)$: $\phi_{M(\text{“\#you”},\text{“strong”})}(x_{i,j}, \mathbf{h}_i)$ takes a value of 3, and $\phi_{M(\text{“\#you”},\text{“weak”})}(x_{i,j}, \mathbf{h}_i)$ is 0. $x_{i,j}$ also has a feature without strength, i.e. $\phi_{M(\text{“\#you”})}(x_{i,j}, \mathbf{h}_i) = 3$. Function $M(\cdot)$ maps each feature to a unique index.

For predicting the outcome of debate $\mathbf{x}_i$, we compute the difference of feature vectors from PRO and CON in two ways: $\tilde{\Phi}^p(\mathbf{x}_i, \mathbf{h}_i) = \Phi(\mathbf{x}_i^p, \mathbf{h}_i) -$

[2] Zhang et al. (2016) also use IQ2 to study talking points and their predictive power on debate outcome. Our dataset includes theirs plus 11 additional debates (excluding one with result as tie).

[3] The dataset can be downloaded from `http://www.ccs.neu.edu/home/luwang/`.

[4] Binary topic strength is better-suited for our proposed discriminative learning framework. In exploratory work we found that continuous-value strength under the same framework tended to be pushed towards extreme values during learning.

[5] This assumes the strength of arguments is additive, though it is possible that a single extremely strong or weak argument could decide a debate, or that debates are won via "rounds" rather than in aggregate.

$\Phi(\mathbf{x}_i^c, \mathbf{h}_i)$ and $\tilde{\Phi}^c(\mathbf{x}_i, \mathbf{h}_i) = \Phi(\mathbf{x}_i^c, \mathbf{h}_i) - \Phi(\mathbf{x}_i^p, \mathbf{h}_i)$. Two decision scores are computed as $f^p(\mathbf{x}_i) = \max_{\mathbf{h}_i}[\mathbf{w} \cdot \tilde{\Phi}^p(\mathbf{x}_i, \mathbf{h}_i)]$ and $f^c(\mathbf{x}_i) = \max_{\mathbf{h}_i}[\mathbf{w} \cdot \tilde{\Phi}^c(\mathbf{x}_i, \mathbf{h}_i)]$. The output is 1 if $f^p(\mathbf{x}_i) > f^c(\mathbf{x}_i)$ (PRO wins); otherwise, the prediction is -1 (CON wins).

Weights $\mathbf{w}$ are learned during training, while topic strengths $\mathbf{h}_i$ are latent variables, and we use integer linear programming to search for $\mathbf{h}_i$ (see § 3.4).

3.2 Learning with Latent Variables

To learn the weight vector $\mathbf{w}$, we use the large margin training objective:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \sum_i l(-y_i \cdot \max_{\mathbf{h}_i}[\mathbf{w} \cdot \tilde{\Phi}(\mathbf{x_i}, \mathbf{h}_i)]) \quad (1)$$

We consider samples based on difference feature vectors $\tilde{\Phi}^p(\mathbf{x}_i, \mathbf{h}_i)$ during training, which is represented as $\tilde{\Phi}(\mathbf{x}_i, \mathbf{h}_i)$ in Eq. 1 and the rest of this section. $l(\cdot)$ is squared-hinge loss function. C controls the trade-off between the two items.

This objective function is non-convex due to the maximum operation (Yu and Joachims, 2009). We utilize Alg. 1, which is an iterative optimization algorithm, to search for the solution for $\mathbf{w}$ and $\mathbf{h}$. We first initialize latent topic strength variables as $\mathbf{h}_0$ (see next paragraph) and learn the weight vector as $\mathbf{w}^*$. Adopted from Chang et al. (2010), our iterative algorithm consists of two major steps. For each iteration, the algorithm first decides the latent variables for positive examples. In the second step, the solver iteratively searches for latent variable assignments for negative samples and updates the weight vector $\mathbf{w}$ with a cutting plane algorithm until convergence. Global variable H_i is maintained for each negative sample to store all the topic strength assignments that give the highest scores during training.[6] This strategy facilitates efficient training while a local optimum is guaranteed.

Topic strength initialization. We investigate three approaches for initializing topic strength variables. The first is based on the *usage frequency per topic*. If one side uses more arguments of a given topic, then its strength is likely to be strong for them and weak for their opponent. Another option is to initialize *all topics as strong for both sides*, then $\mathbf{w}_0$ learns the association between strong topics and features that lead to winning. The third option is to initialize *all topics as strong for winners* and weak for losers.

Input : $\{\mathbf{x}_i, y_i\}_i$: training samples of arguments $\mathbf{x}_i$ and outcome y_i, $\tilde{\Phi}(\cdot,\cdot)$: feature vectors, C: trade-off coefficient, τ: iteration number threshold
Output: feature weights $\mathbf{w}^*$

foreach $\mathbf{h}_i$ **do**
  Initialize $\mathbf{h}_i$ as $\mathbf{h}_i^0$ (see § 3.2)
end
$\mathbf{w}^* \leftarrow \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \sum_i l(-y_i \cdot [\mathbf{w} \cdot \tilde{\Phi}(\mathbf{x}_i, \mathbf{h}_i^0)])$
// H_i: storing h_i^* for negative samples
foreach negative sample $\mathbf{x}_i$ $(y_i = -1)$ **do**
  $H_i \leftarrow \emptyset$
end
$t \leftarrow 0$
while w^* not converge and $t < \tau$ **do**
  // Assign strength for positive samples
  foreach positive sample $\mathbf{x}_i$ $(y_i = 1)$ **do**
    $\mathbf{h}_i^* \leftarrow \arg\max_{\mathbf{h}_i}[\mathbf{w} \cdot \tilde{\Phi}(\mathbf{x}_i, \mathbf{h}_i)]$ (*)
  end
  // Iteration over negative samples
  $t' \leftarrow 0$
  while w^* not converge and $t' < \tau$ **do**
    foreach negative sample $\mathbf{x_i}, y_i = -1$ **do**
      $\mathbf{h}_i^* \leftarrow \arg\max_{\mathbf{h}_i}[\mathbf{w}^* \cdot \tilde{\Phi}(\mathbf{x}_i, \mathbf{h}_i)]$ (*)
      $H_i \leftarrow H_i \cup \{\mathbf{h}_i^*\}$
    end
    $\mathbf{w}^* \leftarrow \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C \cdot \sum_{i,y_i=1} l(-y_i \cdot [\mathbf{w} \cdot \tilde{\Phi}(\mathbf{x}_i, \mathbf{h}_i^*)]) + C \cdot \sum_{i,y_i=-1} l(-y_i \cdot \max_{\mathbf{h} \in H_i}[\mathbf{w} \cdot \tilde{\Phi}(\mathbf{x}_i, \mathbf{h}_i)])$
    $t' \leftarrow t' + 1$
  end
  $t \leftarrow t + 1$
end

Algorithm 1: Iterative algorithm for learning weights $\mathbf{w}$ and latent topic strength variables $\mathbf{h}$. Iteration threshold τ is set as 1000. Steps with (*) are solved as in § 3.4.

3.3 Features

We group our directly observed linguistic features, roughly ordered by increasing complexity, into categories that characterize various aspects of arguments. For each linguistic feature, we compute two versions: one for the full debate and one for the discussion phase.

Basic features. We consider the frequencies of words, numbers, named entities per type, and each personal pronoun. For instance, usage of personal pronouns may imply communicative goals of the speaker (Brown and Gilman, 1960; Wilson, 1990). We also count the frequency of each POS tag output by Stanford parser (Klein and Manning, 2003). Sentiment and emotional language usage is prevalent

[6] A similar latent variable model is presented in Goldwasser and Daumé III (2014) to predict the objection behavior in courtroom dialogues. In their work, a binary latent variable is designed to indicate the latent relevance of each utterance to an objection, and only relevant utterances contribute to the final objection decision. In our case our latent variables model argument strength, and all arguments matter for the debate outcome.

in discussions on controversial topics (Wang and Cardie, 2014b). We thus count words of positive and negative sentiment based on MPQA lexicon (Wilson et al., 2005), and words per emotion type according to a lexicon from Mohammad and Turney (2013). Moreover, based on the intuition that agreement carries indications on topical alignment (Bender et al., 2011; Wang and Cardie, 2014a), occurrence of agreement phrases ("I/we agree", "you're right") is calculated. Finally, audience feedback, including applause and laughter, is also considered.

Style features. Existing work suggests that formality can reveal speakers' opinions or intentions (Irvine, 1979). Here we utilize a formality lexicon collected by Brooke et al. (2010), which counts the frequencies of formal or informal words in each argument. According to Durik et al. (2008), hedges are indicators of weak arguments, so we compile a list of hedge words from Hyland (2005), and hedging of verbs and non-verbs are counted separately. Lastly, we measure word attributes for their concreteness (perceptible vs. conceptual), valence (or pleasantness), arousal (or intensity of emotion), and dominance (or degree of control) based on the lexicons collected by Brysbaert et al. (2014) and Warriner et al. (2013), following Tan et al. (2016), who observe correlations between word attributes and their persuasive effect on online arguments. The average score for each of these features is then computed for each argument.

Semantic features. We encode semantic information via semantic frames (Fillmore, 1976), which represent the context of word meanings. Cano-Basave and He (2016) show that arguments of different types tend to employ different semantic frames, e.g., frames of "reason" and "evaluative comparison" are frequently used in making claims. We count the frequency of each frame, as labeled by SEMAFOR (Das et al., 2014).

Discourse features. The usage of discourse connectors has been shown to be effective for detecting argumentative structure in essays (Stab and Gurevych, 2014). We collect discourse connectives from the Penn Discourse Treebank (Prasad et al., 2007), and count the frequency of phrases for each discourse class. Four classes on level one (temporal, comparison, contingency, and expansion) and sixteen classes on level two are considered. Finally, pleading behavior is encoded as counting phrases of "urge", "please", "ask you", and "encourage you", which may be used by debaters to appeal to the audience.

Sentence-level features. We first consider the frequency of questioning since rhetorical questions are commonly used for debates and argumentation. To model the sentiment distribution of arguments, sentence-level sentiment is labeled by the Stanford sentiment classifier (Socher et al., 2013) as positive (sentiment score of 4 or 5), negative (score of 1 or 2), and neutral (score of 3). We then count single sentence sentiment as well as transitions between adjacent sentences (e.g. positive $\rightarrow$ negative) for each type. Since readability level may affect how the audience perceives arguments, we compute readability levels based on Flesch reading ease scores, Flesch-Kincaid grade levels, and the Coleman-Liau index for each sentence. We use the maximum, minimum, and average of scores as the final features. The raw number of sentences is also calculated.

Argument-level features. Speakers generally do not just repeat their best argument ad infinitum, which suggests that arguments may lose power with repetition. For each argument, we add an indicator feature (i.e. each argument takes value of 1) and an additional version with a decay factor of $\exp(-\alpha \cdot t_k)$, where t_k is the number of preceding arguments by a given side which used topic k, and α is fixed at 1.0. Interruption is also measured, when the last argument (of more than 50 words) in a turn is cut off by at most two sentences from opponent or moderator. Word repetition is often used for emphasis in arguments (Cano-Basave and He, 2016), so we measure bigram repetition more than twice in sequential clauses or sentences.

Interaction features. In addition to independent language usage, debate strategies are also shaped by interactions with other debaters. For instance, previous work (Zhang et al., 2016) finds that debate winners frequently pursue talking points brought up by their opponents'. Here we construct different types of features to measure how debaters address opponents' arguments and shift to their favorable subjects. First, for a given argument, we detect if there is an argument of the same topic from the previous turn by the opponent. If yes, we further measure the number of words of the current argument, the number of common words between the two argu-

ments (after lemmatization is applied), the concatenation of the sentiment labels, and the concatenation of the emotion labels of the two arguments as features; these interactions thus capture interactive strategies regarding quantity speech and sentiment. We also consider if the current argument is of a different topic from the previous argument in the same turn to encode topic shifting behavior.

Feature functions $\phi_{M(\texttt{feature},\texttt{strength})}(x_{i,j}, \mathbf{h}_i)$ in § 3.1 only consider the strengths of single arguments. To capture interactions between sides that relate to their relative argument strengths, we add features $\phi_{M(\texttt{feature},\texttt{strength}^{\texttt{self}},\texttt{strength}^{\texttt{oppo}})}(x_{i,j}, \mathbf{h}_i)$, so that strengths of pairwise arguments on the same topic from both sides are included. For instance, for topic "execution of innocents", side PRO with STRONG strength uses an argument of 100 words to address the challenge raised by CON with WEAK strength. We add four grouped features associated with the number of words addressing an opponent: $\phi_{M(\text{"\#words to oppo"},\text{"strong,weak"})}(x_{i,j}, \mathbf{h}_i)$ is 100, while $\phi_{M(\text{"\#words to oppo"},\text{"weak,weak"})}(x_{i,j}, \mathbf{h}_i)$, $\phi_{M(\text{"\#words to oppo"},\text{"strong,strong"})}(x_{i,j}, \mathbf{h}_i)$, and $\phi_{M(\text{"\#words to oppo"},\text{"weak,strong"})}(x_{i,j}, \mathbf{h}_i)$ are all 0.

3.4 Topic Strength Inference

Topic strength inference is used both for training (Alg. 1) and for prediction. Our goal is to find an assignment $\mathbf{h}_i^*$ that maximizes the scorer $\mathbf{w}^* \cdot \tilde{\Phi}(\mathbf{x}_i, \mathbf{h}_i)$ for a given $\mathbf{w}^*$. We formulate this problem as an integer linear programming (ILP) instance.[7] Since topic strength assignment only affects feature functions that consider strengths, we discuss how to transform those functions into the ILP formulation.

For each topic k of a debate d_i, we create binary variables $r^p_{k,strong}$ and $r^p_{k,weak}$ for pro, where $r^p_{k,strong} = 1$ indicates the topic is STRONG for pro and $r^p_{k,weak} = 1$ denotes the topic is WEAK. Similarly, $r^c_{k,strong}$ and $r^c_{k,weak}$ are created for con.

Given weights associated with different strengths $w_{M(\texttt{feature},strong)}$ and $w_{M(\texttt{feature},weak)}$, the contribution of any feature to the objective (i.e. scoring difference between pro and con) transforms from

$w_{M(\texttt{feature},strong)} \cdot [\sum_{x_{i,j} \in \mathbf{x}_i^p} \phi_{M(\texttt{feature},strong)}(x_{i,j}, \mathbf{h}_i)$
$- \sum_{x_{i,j} \in \mathbf{x}_i^c} \phi_{M(\texttt{feature},strong)}(x_{i,j}, \mathbf{h}_i)]$
$+ w_{M(\texttt{feature},weak)} \cdot [\sum_{x_{i,j} \in \mathbf{x}_i^p} \phi_{M(\texttt{feature},weak)}(x_{i,j}, \mathbf{h}_i)$
$- \sum_{x_{i,j} \in \mathbf{x}_i^c} \phi_{M(\texttt{feature},weak)}(x_{i,j}, \mathbf{h}_i)]$

to the following form:

$w_{M(\texttt{feature},strong)} \cdot [\sum_{x_{i,j} \in \mathbf{x}_i^p} \phi_{M(\texttt{feature})}(x_{i,j}, \mathbf{h}_i) \cdot r^p_{k,strong}$
$- \sum_{x_{i,j} \in \mathbf{x}_i^c} \phi_{M(\texttt{feature})}(x_{i,j}, \mathbf{h}_i) \cdot r^c_{k,strong}]$
$+ w_{M(\texttt{feature},weak)} \cdot [\sum_{x_{i,j} \in \mathbf{x}_i^p} \phi_{M(\texttt{feature})}(x_{i,j}, \mathbf{h}_i) \cdot r^p_{k,weak}$
$- \sum_{x_{i,j} \in \mathbf{x}_i^c} \phi_{M(\texttt{feature})}(x_{i,j}, \mathbf{h}_i) \cdot r^c_{k,weak}]$

The above equation can be reorganized into a linear combination of variables $r^*_{*,*}$. We further include constraints as discussed below, and solve the maximization problem as an ILP instance.

For features that consider strength for pairwise arguments, i.e. $\phi_{M(\texttt{feature},\texttt{strength}^{\texttt{self}},\texttt{strength}^{\texttt{oppo}})}$, we have binary variables $r^{p,c}_{k,strong,strong}$ (strength is strong for both sides), $r^{p,c}_{k,strong,weak}$ (strong for pro, weak for con), $r^{p,c}_{k,weak,strong}$ (weak for pro, strong for con), and $r^{p,c}_{k,weak,weak}$ (weak for both).

Constraints. We consider three types of topic strength constraints for our ILP formulation:

• C1 – *Single Topic Consistency*: each topic can either be strong or weak for a given side, but not both. Pro: $r^p_{k,strong} + r^p_{k,weak} = 1$; con: $r^c_{k,strong} + r^c_{k,weak} = 1$

• C2 – *Pairwise Topic Consistency*: for pairwise arguments from pro and con on the same topic, their joint assignment is true only when each of the individual assignments is true. C2 applies only for features of pairwise arguments.

$r^{p,c}_{k,strong,strong} = r^p_{k,strong} \wedge r^c_{k,strong}$;
$r^{p,c}_{k,strong,weak} = r^p_{k,strong} \wedge r^c_{k,weak}$;
$r^{p,c}_{k,weak,strong} = r^p_{k,weak} \wedge r^c_{k,strong}$;
$r^{p,c}_{k,weak,weak} = r^p_{k,weak} \wedge r^c_{k,weak}$

• C3 – *Exclusive Strength*: a topic cannot be strong for both sides. This constraint is optional and will be tested in experiments. $r^p_{k,strong} + r^c_{k,strong} \leq 1$

3.5 Argument Identification

In order to identify the topics associated with a debate and the contiguous chunks of same-topic text that we take to be arguments, for each separate debate we utilize a hidden topic Markov model (HTMM) (Gruber et al., 2007) which jointly models the topics and topic transitions between sentences. For details on HTMM, we refer the readers to Gruber et al. (2007).

The HTMM assigns topics on the sentence level, assuming each sentence is generated by a topic draw

[7] We use LP Solve: `http://lpsolve.sourceforge.net/5.5/`.

followed by word draws from that topic, with a transition probability determining whether each subsequent sentence has the same topic as the preceding sentence, or is a fresh draw from the topic distribution. Unlike the standard HTMM process, however, we presume that while both sides of a debate share the same topics, they may have different topic distributions reflecting the different strengths of those topics for either side. We thus extend the HTMM by allowing different topics distributions for the pro and con speech transcripts, but enforce shared word distributions for those topics. To implement this, we first run HTMM on the entire debate, and then rerun it on the pro and con sides while fixing the topic-word distributions. Consecutive sentences by the same side with the same topic are treated as a single argument.

4 Experimental Results

4.1 Experimental Setup

We test via leave-one-out for all experiments. For logistic regression classifiers, ℓ_2 regularization with a trade-off parameter of 1.0 is used. For support vector machines (SVM) classifiers and our models, we fix the trade-off parameter between training error and margin as 0.01. Real-valued features are normalized to $[0, 1]$ via linear transformation.

Our modified HTMM is run on each debate, with number of topics between 10 and 20. Topic coherence, measured via Röder et al. (2015), is used to select the topic number that yields highest score. On average, there are 13.7 unique topics and about 322.0 arguments per debate.

4.2 Baselines and Comparisons

We consider two baselines trained with logistic regression and SVMs classifiers: (1) NGRAMS, including unigrams and bigrams, are used as features, and (2) AUDIENCE FEEDBACK (applause and laughter) are used as features, following Zhang et al. (2016). We also experiment with SVMs trained with different sets of features, presented in § 3.3.

4.3 Results

The debate outcome prediction results are shown in Table 1. For our models, we only display results with latent strength initialization based on frequency per topic, which achieves the best performance. Results for different initialization methods are exhibited and discussed later in this section. As can be seen, our model that leverages learned latent topic strengths and their interactions with linguistic features significantly outperform the non-trivial baselines[8] (bootstrap resampling test, $p < 0.05$). Our latent variable models also obtain better accuracies than SVMs trained on the same linguistic feature sets. Without the audience feedback features, our model yields an accuracy of 72.0%, while SVM produces 65.3%. This is because our model can predict topic strength out of sample by learning the interaction between observed linguistic features and unobserved latent strengths. During test time, it infers the latent strengths of entirely new topics based on observable linguistic features, and thereby predicts debate outcomes more accurately than using the directly observable features alone. Using the data in Zhang et al. (2016) (a subset of our dataset), our best model obtains an accuracy of 73% compared to 65% based on leave-one-out setup.

As mentioned above, we experimented with a variety of latent topic strength initializations: argument frequency per topic (*Freq*); all topics strong

	SVMs	**Our Model** (*w Latent Strength*)
Baselines		
NGRAMS	61.0	–
AUDIENCE FEEDBACK	56.8	–
Features (as in § 3.3)		
BASIC	57.6	59.3
+ STYLE, SEMANTICS, DISCOURSE	59.3	65.3
+ SENTENCE, ARGUMENT	62.7	69.5
+ INTERACTION (all features)	66.1	**73.7**

Table 1: Debate outcome prediction results for baseline models and SVMs using the various linguistic feature categories, compared to our model that includes latent argument strengths in addition to the linguistic features. The best performing system (in **bold**) is achieved by our system with topic strength as latent variables when all features are used, which significantly outperforms the baselines via bootstrap resampling test ($p < 0.05$). For the lower section, each row shows features included in addition to those in the rows above.

[8] For baselines with logistic regression classifiers, the accuracy is 63.6 with ngram features, and 58.5 with audience feedback features.

for both sides (*AllStrong*); strong just for winners (*AllStrong*$_{win}$); and *Random* initialization. From Table 2, we can see that there is no significant difference among different initialization methods. Furthermore, the strength constraints make little difference, though their effects slightly vary with different initializations. Most importantly, the constraint that topics cannot be strong for both sides (i.e., C3) does not systematically help, suggesting that in many cases topics may indeed be strong for both sides, as discussed below.

	Initialization			
Constraints	*Freq*	*AllStrong*	*AllStrong*$_{win}$	*Random*
C1, C2	73.7	71.2	70.3	67.8
C1, C2, C3	72.9	73.7	69.5	68.6

Table 2: Prediction results (in accuracy) with different initialization and topic strength constraints. *C3* denotes a constraint that a topic cannot be strong for both sides.

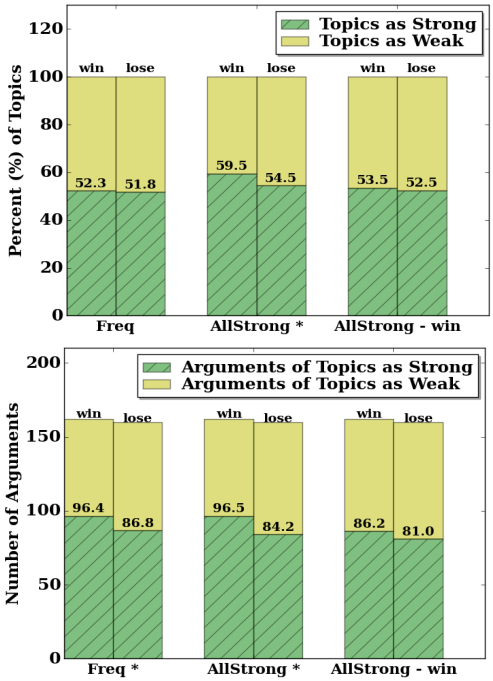


Figure 2: [Upper] Average percentage of topics inferred as STRONG and WEAK for winning ("win") and losing sides ("lose"). [Lower] Raw number of arguments of STRONG and WEAK topics. Numbers are computed for three types of topic strength initialization: initialized by frequency (Freq), all topics are strong for both sides (AllStrong), and all topics are strong for winners (AllStrong - win). Two-sided Mann-Whitney rank test is conducted on values of STRONG topics ($*$: $p < 0.05$).

5 Discussion

In this section, we first analyze argument strengths for winning and losing sides, followed by a comparison of these results with human evaluations (§ 5.2). We then examine the interactive topic shifting behavior of debaters (§ 5.3) and analyze the linguistic features predictive of debate outcome, particularly the ones that interact with topic strength (§ 5.4). The results are reported by training our model on the full dataset. Initialization of topic strength is based on usage frequency unless otherwise specified.

5.1 Topic and Argument Usage Analysis

We start with a basic question: *do winning sides more frequently use strong arguments?* For each side, we calculate the proportion of strong and weak topics as well as the total number of strong and weak arguments on each side. Figure 2 shows that under all three topic strength initializations, our model infers a greater number of strong topics for winners than for losers. This result is echoed by human judgment of topic strength, as described in § 5.2. Similarly, winners also use significantly more individually strong arguments.

As can be seen in Table 2, the constraint that a topic be strong for at most one side only increased accuracy for one initialization case. This indicates that, in general, the model was improved by allowing some topics to be strong for both sides. Interestingly, while the majority (53%) of topics are STRONG for one side and WEAK for the other, about a third (31%) of topics are inferred as STRONG for both sides. While it is clear what it means for a topic to be strong for one side and not the other (as in our death penalty example), or weak for both sides (as in a digression off of the general debate topic), the importance of both-strong for prediction is a somewhat surprising result. Figure 3 illustrates an example as judged by our model. What this shows is that even on a given topic within a debate (Syrian refugees: resettlement), there are different subtopics that may be selectively deployed (resettlement success; resettlement cost) that make the general topic strong for both sides in different ways. For subsequent work, a hierarchical model with nested strength relationships (McCombs, 2005; Nguyen et al., 2015) can be designed to better characterize the topics.

Lastly, we display the usage of strong arguments

Motion: *The U.S. Should Let In 100,000 Syrian Refugees*
Topic: Refugee resettlement
• Pro (STRONG): 415 Syrians resettled by the IRC. Our services show that last year, 8 out of 10 Syrians who we resettled were in work within six months of getting to the United States. And there's one other unique resource of this country: Syrian-American communities across the country who are successful. ...
• Con (STRONG): It costs about 13 times as much to resettle one refugee family in the United States as it does to resettle them closer to home. ... They're asking you to look only at the 400 – the examples of the 415 Syrians that David Miliband's group has so well resettled, and to ignore what is likely to happen as the population grows bigger.

Figure 3: A sample exchange where the argument topic is strong for both sides.

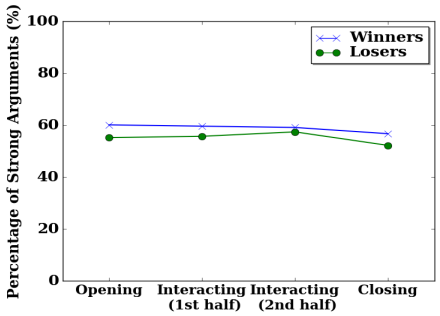


Figure 4: Usage of arguments with strong topics at different stages for winners and losers. Similar results are achieved by counting the number of words in arguments.

during the course of debates in Figure 4. Each debate is divided into opening statements, two interacting phases (equal number of turns), and closing statements. Similar usage of strong arguments are observed as debates progress, though a slight, statistically non-significant drop is noted in the closing statement. One possible interpretation is that debaters have fully delivered their strong arguments during opening and interactions, while only weaker arguments remain when closing the debates.

5.2 Human Validation of Topic Strength

Here we evaluate *whether our inferred topic strength matches human judgment.* We randomly selected 20 debates with a total of 268 topics. For each debate, we first displayed its motion and a brief description constructed by IQ2. Then for each topic, the top 30 topic words from the HTMM model were listed, followed by arguments from PRO and CON. Note that debate results were not shown to the annotators.

We hired three human annotators who are native speakers of English. Each of them was asked to first

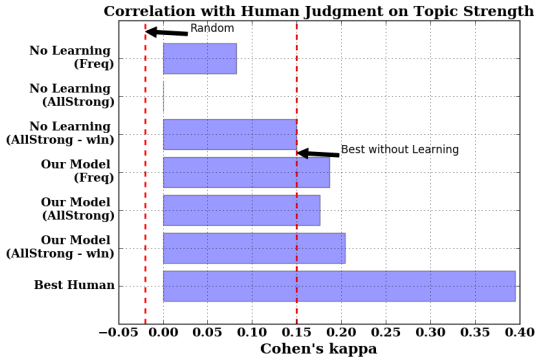


Figure 5: Topic strength correlation with human judgment using Cohen's κ. The left red dotted line indicates the best correlation between a random assignment and human, and the right red dotted line shows the best correlation without learning.

evaluate topic coherence by reading the word list and rate it on a 1-3 scale (1 as incoherent, 3 as very coherent). If the judgment was coherent (i.e., a 2 or 3), they then read the arguments and judged whether (a) both sides are strong on the topic, (b) both sides are weak, (c) pro is strong and con is weak, or (d) con is strong and pro is weak.

54.9% of the topics were labeled as coherent by at least two annotators. However, since topics are estimated separately for each debate, even the less coherent topics generally had readily interpretable meanings in the context of a given debate. Among coherent topics, inter-annotator agreement for topic strength annotation had a Krippendorff's α of 0.32. Judging argument strength is clearly a more difficult and subjective task.

Nevertheless, without knowing debate outcomes, all three human judges identified more strong topics for winning sides than losing sides. Among the coherent topics, a (macro-)average of 44.4% of topics were labeled as strong for winners, compared to 30.1% for losers. This echoes the results from our models as illustrated in Figure 2.

Furthermore, we calculate the correlation between topic strength inferred by our models and the ones labeled by each human judge using Cohen's κ. The results are illustrated in Figure 5, which shows our three different initializations, with and without learning. The highest human κ is also displayed. Our trained models clearly match human judgments better than untrained ones.

	$(topic_{self}, topic_{oppo}) \rightarrow (topic'_{self}, topic'_{oppo})$	*Percent*
Winners	(Strong, Weak) → (Strong, Weak)	12.7%
	(Strong, Strong) → (Strong, Strong)	10.5%
	(Strong, Strong) → (Strong, Weak)	8.9%
Losers	(Weak, Strong) → (Weak, Strong)	11.9%
	(Strong, Strong) → (Strong, Strong)	9.0%
	(Strong, Weak) → (Strong, Weak)	8.8%

Table 3: Top 3 types of shifts. $topic_{self}$ and $topic_{oppo}$ are the strengths of the current topic for one side and their opponent. $topic'_{self}$ and $topic'_{oppo}$ are the strengths of the topic for the following arguments.

Category	**Topic Strength**	
	STRONG	WEAK
BASIC	# "we"$_{full}$	# "you"$_{inter}$ *
	# "they"$_{inter}$	# "I"$_{inter}$
	# "emotion:sadness"$_{full}$	# "emotion:joy"$_{full}$ *
	# "emotion:disgust"$_{full}$	# "emotion:trust"$_{full}$ **
		# unique nouns$_{full}$ *
		# "numbers"$_{full}$
		# named entities$_{inter}$ *
STYLE, SEMANTIC, DISCOURSE	# non-verb hedging $_{full}$	# non-verb hedging $_{full}$
	avg concreteness$_{full}$ *	avg arousal score$_{full}$ *
	# formal words$_{full}$ *	# PDTB:temporal$_{inter}$ *
	# FS:capability$_{full}$	# PDTB:contrast$_{inter}$
	# FS:information$_{full}$	# FS:certainty$_{full}$
SENTENCE, ARGUMENT	Flesch Reading Ease$_{full}$	Flesch Reading Ease$_{full}$
	# sentiment:negative$_{full}$ *	# sentiment:neutral$_{inter}$ *
	# question$_{full}$	# question$_{full}$
	# audience laughter$_{inter}$ *	
	decayed argument count$_{full}$ *	
INTERACTION	# words addressing opponent's argument$_{full}$	if addressing opponent's argument$_{full}$
	# common words with opponent's argument$_{full}$	

Table 4: Top weighted features joint with topic strength. "$full$" and "$inter$" indicates features that are calculated for full debates or the interactive (discussion) phase only. "FS" denotes frame semantic. Two-sided Mann-Whitney rank test is conducted on between features of wining and losing sides (*: $p < 0.05$, **: $p < 0.01$).

5.3 Topic Shifting Behavior Analysis

Within competitive debates, strategy can be quite interactive: one often seeks not just to make the best arguments, but to better the opponent from round to round. Agenda setting, or shifting the topic of debate, is thus a crucial strategy. An important question is therefore: *do debaters strategically change topics to ones that benefit themselves and weaken their opponent?* According to the HTMM results, debaters make 1.5 topical shifts per turn on average. Both winning and losing teams are more likely to change subjects to their strong topics: winners in particular are much more likely to change the topic to something strong for them (61.4% of shifts), although debate losers also attempt this strategy (53.6% of shifts).

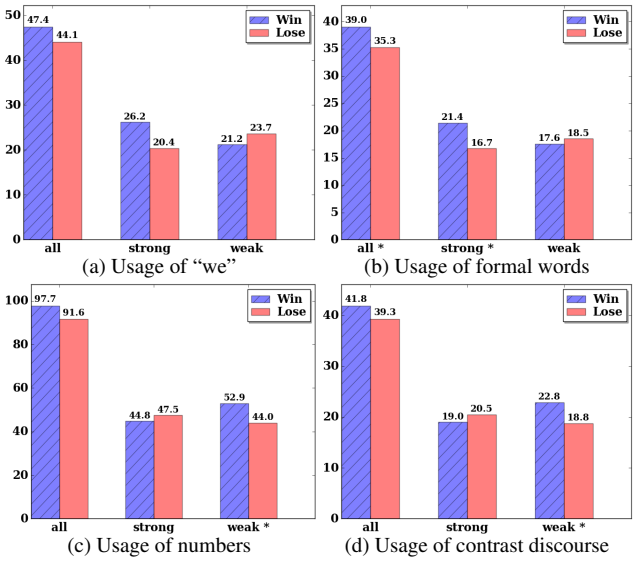
A more sophisticated strategy is if the debaters also attempt to put their opponents at a disadvantage with topic shifts. We consider the topic strengths of a current argument for both the speaker ("*self*") and their "*opponent*", as well as the strength of the following argument. The top 3 types of shifts are listed in Table 3. As can be seen, winners are more likely to be in a strong (for them) and weak (for the opponent) situation and to stay there, while losers are more likely to be in the reverse. Both sides generally stay in the same strength configuration from argument 1 to argument 2, but winners are also likely (row 3) to employ the strategy of shifting from a topic that is strong for both sides, to one that is strong for them and weak for the opponent.

5.4 Feature Analysis

Lastly, we investigate the *linguistic features associated with topics of different strengths that affect the audience*. Table 4 displays some of the 50 highest weighted features that interact with strong and weak topics. Personal pronoun usage has been found to be related to communicative goals in many previous studies (Brown and Gilman, 1960; Wilson, 1990). We find that strong topics are associated with more first person plurals, potentially an indicator of group responsibility (Wilson, 1990). On the other hand, our model finds that weak topics are associated with second person pronouns, which may be arguments either attacking other discussants or addressing the audience (Simons and Jones, 2011). For sentiment, previous work (Tan et al., 2016) has found that persuasive arguments are more negative in online discussions. Our model associates negative sentiment and anger words with strong topics, and neutral and joyful languages with weak topics.

In terms of style and discourse, debaters tend to use more formal and more concrete words for arguments with strong topics. By contrast, arguments with weak topics show more frequent usage of words with intense emotion (higher arousal scores), and contrast discourse connectives. Figure 6 shows how some of these features differ between winners and losers, illustrating the effects on outcome via strong or weak arguments in particular.

Interaction features also play an important role for



(a) Usage of "we" (b) Usage of formal words

(c) Usage of numbers (d) Usage of contrast discourse

Figure 6: Values of sample features with substantial difference between weights associated with "strong" and "weak" topics are plotted next to feature values of "all" arguments. Two-sided Mann-Whitney rank test is conducted between wining and losing sides ($*$: $p < 0.05$).

affecting audiences' opinions. In particular, debaters spend more time (i.e. use more words) addressing their opponents' arguments if it is a strong topic for their opponents. But even for weak topics, it appears helpful to address opponents' arguments.

6 Related Work

Previous work on debate and persuasion has studied the dynamic of audience response to debates and the rhetorical frames the speakers use (Boydstun et al., 2014). However, this work is limited by the scarcity of data and does not focus on the interactions between content and language usage. Topic control, operationalized as the tendency of one side to adopt or avoid the words preferentially used by the other, is investigated in Zhang et al. (2016) to predict debate outcome using the Intelligence Squared data. Our work complements theirs in examining topic interactions, but brings additional focus on the latent persuasive strength of topics, as well as strength interactions. Tan et al. (2016) examines various structural and linguistic features associated with persuasion on Reddit; they find that some topics correlate more with malleable opinions. Here we develop a more general model of latent topic strength and the linguistic features associated with strength.

Additional work has focused on the influence of agenda setting — controlling which topics are discussed (Nguyen et al., 2014), and framing (Card et al., 2015; Tsur et al., 2015) — emphasizing certain aspects or interpretations of an issue. Greene and Resnik (2009) study the syntactic aspects of framing, where syntactic choices are found to correlate with the sentiment perceived by readers. Based on the topic shifting model of Nguyen et al. (2014), Prabhakaran et al. (2014) finds that changing topics in presidential primary debates positively correlates with the candidates' power, which is measured based on their relative standing in recent public polls. This supports our finding that both sides *seek* to shift topics, but that winners are more likely to shift to topics which are strong for them but weak for their opponents.

Our work is in line with argumentation mining. Existing work in this area focuses on argument extraction (Moens et al., 2007; Palau and Moens, 2009; Mochales and Moens, 2011) and argument scheme classification (Biran and Rambow, 2011; Feng and Hirst, 2011; Rooney et al., 2012; Stab and Gurevych, 2014). Though stance prediction has also been studied (Thomas et al., 2006; Hasan and Ng, 2014), we are not aware of any work that extracts arguments according to topics and position. Argument strength prediction is also studied largely in the domain of student essays (Higgins et al., 2004; Stab and Gurevych, 2014; Persing and Ng, 2015). Notably, none of these distinguishes an argument's strength from its linguistic surface features. This is a gap we aim to fill.

7 Conclusion

We present a debate prediction model that learns latent persuasive strengths of topics, linguistic style of arguments, and the interactions between the two. Experiments on debate outcome prediction indicate that our model outperforms comparisons using audience responses or linguistic features alone. Our model also shows that winners use stronger arguments and strategically shift topics to stronger ground. We also find that strong and weak arguments differ in their language usage in ways relevant to various behavioral theories of persuasion.

Acknowledgments

This work was supported in part by National Science Foundation Grant IIS-1566382 and a GPU gift from Nvidia. We thank the TACL reviewers for valuable suggestions on various aspects of this work.

References

Frank R. Baumgartner, Suzanna L. De Boef, and Amber E. Boydstun. 2008. *The decline of the death penalty and the discovery of innocence*. Cambridge University Press.

Emily M. Bender, Jonathan T. Morgan, Meghan Oxley, Mark Zachry, Brian Hutchinson, Alex Marin, Bin Zhang, and Mari Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in Wikipedia talk pages. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 48–57, Portland, OR, June. Association for Computational Linguistics.

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, 5(04):363–381.

Amber E. Boydstun, Rebecca A. Glazier, Matthew T. Pietryka, and Philip Resnik. 2014. Real-time reactions to a 2012 presidential debate a method for understanding which messages matter. *Public Opinion Quarterly*, 78(S1):330–343.

Julian Brooke, Tong Wang, and Graeme Hirst. 2010. Automatic acquisition of lexical formality. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 90–98, Stroudsburg, PA, USA. Association for Computational Linguistics.

Roger Brown and Albert Gilman, 1960. *The pronouns of power and solidarity*, pages 253–276. MIT Press, Cambridge, MA.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.

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

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. 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. Association for Computational Linguistics.

Ming-Wei Chang, Dan Goldwasser, Dan Roth, and Vivek Srikumar. 2010. Discriminative learning over constrained latent representations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 429–437, Los Angeles, CA, June. Association for Computational Linguistics.

Joshua Cohen. 1989. *Deliberation and Democratic Legitimacy*. The Good Polity: Normative Analysis of the State. Basil Blackwell.

Dipanjan Das, Desai Chen, André F.T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.

John S. Dryzek and Christian List. 2003. Social choice theory and deliberative democracy: A reconciliation. *British Journal of Political Science*, 33(01):1 – 28.

Amanda M. Durik, M. Anne Britt, Rebecca Reynolds, and Jennifer Storey. 2008. The effects of hedges in persuasive arguments: A nuanced analysis of language. *Journal of Language and Social Psychology*.

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

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Dan Goldwasser and Hal Daumé III. 2014. "I object!" modeling latent pragmatic effects in courtroom dialogues. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 655–663, Gothenburg, Sweden, April. Association for Computational Linguistics.

Stephan Greene and Philip Resnik. 2009. 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, CO, June. Association for Computational Linguistics.

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic Markov models. In *International conference on artificial intelligence and statistics*, pages 163–170.

Jürgen Habermas. 1984. *The theory of communicative action*. Beacon Press, Boston.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar, October. Association for Computational Linguistics.

Derrick Higgins, Jill Burstein, Daniel Marcu, and Claudia Gentile. 2004. Evaluating multiple aspects of coherence in student essays. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, Boston, Massachusetts, USA, May 2-7, 2004*, pages 185–192.

Ken Hyland. 2005. *Metadiscourse: Exploring interaction in writing*. Continuum, London.

Judith T. Irvine. 1979. Formality and informality in communicative events. *American Anthropologist*, 81(4):773–790.

Joel Katzav and Chris Reed. 2008. Modelling argument recognition and reconstruction. *Journal of Pragmatics*, 40(1):155–172.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July. Association for Computational Linguistics.

Jane Mansbridge. 2003. Rethinking representation. *The American Political Science Review*, 97(04):515–528.

Jane Mansbridge, 2015. *A Minimalist Definition of Deliberation*, book section 2, pages 27–50. Equity and Development series. World Bank Publications.

Maxwell McCombs. 2005. A look at agenda-setting: Past, present and future. *Journalism studies*, 6(4):543–557.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 225–230. ACM.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, Deborah A. Cai, Jennifer E. Midberry, and Yuanxin Wang. 2014. Modeling topic control to detect influence in conversations using nonparametric topic models. *Machine Learning*, 95(3):381–421.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. 2015. 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. Association for Computational Linguistics.

Elisabeth Noelle-Neumann. 1974. The spiral of silence a theory of public opinion. *Journal of communication*, 24(2):43–51.

Raquel Mochales Palau and Marie-Francine Moens. 2009. Argumentation mining: the detection, classification and structure of arguments in text. In *Proceedings of the 12th international conference on artificial intelligence and law*, pages 98–107. ACM.

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

Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. 2014. Staying on topic: An indicator of power in political debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1481–1486, Doha, Qatar, October. Association for Computational Linguistics.

Rashmi Prasad, Eleni Miltsakaki, Nikhil Dinesh, Alan Lee, Aravind Joshi, Livio Robaldo, and Bonnie L. Webber. 2007. The Penn discourse treebank 2.0 annotation manual.

John Rawls. 1997. The idea of public reason revisited. *The University of Chicago Law Review*, 64(3):765–807.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408, New York, NY, USA. ACM.

Niall Rooney, Hui Wang, and Fiona Browne. 2012. Applying kernel methods to argumentation mining. In *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference, Marco Island, Florida. May 23-25, 2012*.

Herbert W. Simons and Jean Jones. 2011. *Persuasion in society*. Routledge, 2nd ed. edition.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and

Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

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

Cass R. Sunstein. 1999. The law of group polarization.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 613–624.

Matt Thomas, Bo Pang, and Lillian Lee. 2006. 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 (EMNLP)*, pages 327–335. Association for Computational Linguistics.

Oren Tsur, Dan Calacci, and David Lazer. 2015. 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. Association for Computational Linguistics.

Lu Wang and Claire Cardie. 2014a. Improving agreement and disagreement identification in online discussions with a socially-tuned sentiment lexicon. In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 97–106, Baltimore, MD, June. Association for Computational Linguistics.

Lu Wang and Claire Cardie. 2014b. A piece of my mind: A sentiment analysis approach for online dispute detection. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–699, Baltimore, MD, June. Association for Computational Linguistics.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4):1191–1207.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA. Association for Computational Linguistics.

John Wilson. 1990. *Politically speaking: The pragmatic analysis of political language*. Basil Blackwell.

Chun-Nam John Yu and Thorsten Joachims. 2009. Learning structural SVMs with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1169–1176, New York, NY, USA. ACM.

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