The Role of Party and Incumbency in Identification of Argumentation Strategies in Political Debate

Anonymous ACL submission

Abstract

Many facets of language can only be understood in terms of who is speaking, who is listening, and the context of their interaction. Here we explore this in the domain of political debate where prior work suggests strong links between rhetoric and partisan affiliation. In particular, we explore how party and electoral status can be used to improve performance on classification of value and identity arguments. We first consider splitting the data based on these features, training separate models for each group, and then introduce a model which directly combines these features with LSTM-generated latent text representations. We demonstrate that not only does the inclusion of these features in our final model improve performance, it does so with negligible additional computational cost.

1 Introduction

The analysis of political rhetoric was central to some of the earliest considerations of language (Aristotle, 335 BCE) and likely extends back to the first time an angry hominid blamed the next tribe over for the lack of rain. However, computational analysis of this rhetoric is complicated by a variety of issues. For example, some statements are deliberately obscured so that only groups who already agree with a stance will even realize that a position is being taken (López, 2015). One key aspect is the way this domain highlights the interaction between message, source, and target. For example, strong partisans are more likely to accept or reject a claim based on whether the source shares their affiliation (Cohen, 2003; Jost et al., 2009). Identity and ideology can not be separated from linguistic understanding.

In this paper, we focus on how incorporating information about party and electoral status (incumbent versus challenger) can help to automatically identify value and group identity statements. We evaluate this on a corpus of US presidential general election debates annotated based on value and group identity claims (Anonymous: Under review, 2017). This work is driven by the concept that these are features which, while not always directly available in the text, are available to listeners and drive understanding and response.

2 Data

This work makes use of an annotated set of US general election presidential debates (Anonymous: Under review, 2017). The full set will be made publicly available. The current set consists of 18 debates with a total of 2568 debate turns consisting of 16984 sentences and 272k words. Sentences were annotated based on the presence or absence of five discourse strategies commonly used in debates: attacking one’s opponent, description of policy, statement of values, group identity claim, and a discussion of personal characteristics. A single sentence could be tagged with multiple labels and examples of all label combinations were observed. A sentence could be left unlabeled if none of the categories were present. These annotations capture two key axes: the gap between personal and group identity and the gap between discussions of policy and presentations of values. The first of these is important as recent work suggests that group identity is a critical factor in political decision making (Dawes et al., 1990; Carpini et al., 2004). The second is important given that voters can often more easily determine a preferred value orientation as opposed to...
evaluating technical policy proposals (Richardson, 2002; Christiano, 1995).

3 Experiment 1: Split data by party

Given the surface divisions we knew to be present in this data (such as between parties and between incumbents and challengers), our first question was whether separating the training and testing data along these axes would improve overall performance.

Differences in the issue profiles and rhetorical strategies of the two primary US political parties has been extensively studied (Heritage and Greatbatch, 1986). One natural question is whether classification would be better served by training separately on each of the two parties or by combining the two. Our goal in this experiment was to compare classification performance when training and testing using the complete dataset, training and testing on the same party, or training on one party and testing on the other.

3.1 Model structure

Prior work on this dataset found best classification results with a multitask LSTM (Anonymous: Under review, 2017) which we make use of as our baseline model. An embedding layer of 300 dimensional word representations fed into a standard LSTM (Hochreiter and Schmidhuber, 1997) yielding a 300 dimensional latent representation for each turn. This was passed to five separate 100-node dense layers with each feeding into a binary output layer for the five labels. All models were optimized with Adam (Kingma and Ba, 2014) and trained over 30 epochs. We made use of pre-trained word embeddings trained with the Word2Vec model (Mikolov et al., 2013) on the full text of the English Wikipedia.

Models were evaluated on leave-one-out cross validation on the annotated debates. For each of the 18 annotation pairs we separately train a model based on the other 17 (with a 90/10 train/validation split for the learning phase) and evaluate on the held-out debate. We report averaged results.

3.2 Results and discussion

While we expected the intra-party models to do well, we anticipated that they would be hurt by the fact that they had access to less than half the training data of the complete model. However, we found they generally did as well as the complete model and in some cases better. Most dramatic was the attack label where the model trained and tested on democrats had an absolute improvement of 12 points, from 59% to 71%. Equally interesting was how much worse these models did when trained on one party and tested on the other.

As noted, the differences in training sets makes an absolute comparison impossible, however, we can draw some interesting inferences from the relative results. The two cases where the combined model outperformed all of the split data models were for value arguments and personal statements. Of the two, personal statements seems natural—it’s unsurprising for politicians talking about their own charms to sound somewhat similar regardless of party. However, value statements were the category where we had expected the most dramatic differences between parties. The precise reasons for this will require follow-up with more targeted experiments.

4 Experiment 2: Split by incumbent/challenger

Here, we follow the same method described above but apply it to incumbents and challengers. We observed that challengers tended to attack more than incumbents and wanted to test whether the attack category in particular would be better represented in this context. For each year, the candidate from the party in power is treated as the incumbent. In the included years, this was either the current president or vice president.

<table>
<thead>
<tr>
<th>Train</th>
<th>All</th>
<th>Republican</th>
<th>Democrat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attack</td>
<td>0.59</td>
<td>0.58</td>
<td>0.43</td>
</tr>
<tr>
<td>Policy</td>
<td>0.71</td>
<td>0.72</td>
<td>0.57</td>
</tr>
<tr>
<td>Value</td>
<td>0.54</td>
<td>0.46</td>
<td>0.25</td>
</tr>
<tr>
<td>Ingroup</td>
<td>0.14</td>
<td>0.19</td>
<td>0.00</td>
</tr>
<tr>
<td>Personal</td>
<td>0.47</td>
<td>0.46</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Table 1: F1 score for the multitask model trained and tested on Republicans (R) or Democrats (D) and the model (All) trained and tested on the complete training set.

---

1 Although the period covered included two independent candidates, due to the small amount of data we focus on the two primary US political parties.

2 http://wikipedia.org/
Table 2: F1 score for the multitask model when trained on either incumbents (I) or challengers (C) and tested on both.

<table>
<thead>
<tr>
<th>Train</th>
<th>All</th>
<th>Incumbent</th>
<th>Challenger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>All</td>
<td>I</td>
<td>C</td>
</tr>
<tr>
<td>Attack</td>
<td>0.59</td>
<td>0.55</td>
<td>0.42</td>
</tr>
<tr>
<td>Policy</td>
<td><strong>0.71</strong></td>
<td>0.61</td>
<td>0.60</td>
</tr>
<tr>
<td>Value</td>
<td>0.54</td>
<td><strong>0.56</strong></td>
<td>0.40</td>
</tr>
<tr>
<td>Ingroup</td>
<td>0.14</td>
<td>0.08</td>
<td>0.02</td>
</tr>
<tr>
<td>Personal</td>
<td>0.47</td>
<td>0.36</td>
<td>0.25</td>
</tr>
</tbody>
</table>

4.1 Results and discussion

As seen in table 2, the results show that splitting data along these lines can lead to improved performance. Once again, given the differences in the amount of training data, these should not be compared directly, but we can evaluate relative performance. While the split-data models outperformed the baseline on four out of five categories, the results were generally within a few points. Once again, the when trained on either incumbents or challengers and tested on the other group, the results were far worse.

As in the party case, training and testing within these groups captured additional information. For this split, we were particularly interested in the attack category as it seemed the most likely to show a difference on this axis. However, while the performance was improved when training/testing on challengers, some of this may be that democrats were overrepresented in the challenger set and, as seen in experiment 1, their attacks proved far easier to classify. It is also unclear why the model trained and tested on incumbents would do better on value arguments.

However, we do see the same pattern observed in the previous experiment where performance declines dramatically on the cross-group evaluation. The inability of a model trained on one group to predict the behavior of the other is a strong indication that these splits are providing useful information. This motivated our next experiment.

5 Experiment 3: Side information model

The previous two experiments strongly suggested that useful information is contained in these categories. But the approach of splitting training/testing data by category is neither scalable nor extensible. As such, we wanted to explore directly incorporating these features into a single model.

Table 3: F1 scores for the baseline multitask model, the model with party information added (+P), incumbency information added (+I), and both added (+PI).

<table>
<thead>
<tr>
<th></th>
<th>Base</th>
<th>+P</th>
<th>+I</th>
<th>+PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attack</td>
<td>0.59</td>
<td>0.62</td>
<td>0.63</td>
<td><strong>0.66</strong></td>
</tr>
<tr>
<td>Policy</td>
<td>0.71</td>
<td>0.74</td>
<td>0.71</td>
<td><strong>0.75</strong></td>
</tr>
<tr>
<td>Value</td>
<td>0.54</td>
<td>0.54</td>
<td>0.52</td>
<td><strong>0.55</strong></td>
</tr>
<tr>
<td>Ingroup</td>
<td>0.14</td>
<td>0.15</td>
<td><strong>0.16</strong></td>
<td>0.15</td>
</tr>
<tr>
<td>Personal</td>
<td>0.47</td>
<td>0.49</td>
<td><strong>0.53</strong></td>
<td>0.49</td>
</tr>
</tbody>
</table>

Figure 1: The multi-task model with party and incumbency information.

We do this by extending the network described in experiment 1. Rather than simply passing the latent representation generated by the LSTM layers to the dense layers for label prediction, we first generate one-hot vector representations of party and incumbency status and concatenate these with the latent representation before passing this combined representation on to the prediction layers (see figure 1). We then repeated the evaluations on the full dataset using leave-one-out cross validation on the annotation pairs as described previously.

5.1 Results and discussion

As seen in table 3, including this information consistently improved overall model performance, particularly when both party and incumbency information was included. However, although the model with both types of information outperformed the baseline on all cases, the best perfor-
The extensive social scientific literature on political rhetoric has established the importance of group identity and candidate status to the understanding of what is being said at any moment. Beyond political speech, language understanding generally depends on the interplay between who is speaking, who is listening, and what the context is. Language is more than a informational channel, it’s a form of social action (Austin, 1975) and can only be fully understood in context. Unlike tasks such as image recognition, sometimes there is no single right answer.

Computational work is beginning to address these challenges, but there is much work to be done. In particular, this area defines one of the key regions where NLP and the social sciences must intersect. Social scientific experiments provide an essential guide to modeling the dynamics of which types of information are best treated as separate features and which should be learned directly from the data. In turn, those social scientific theories can be informed by the computational studies which incorporate those insights.

In this study, we have approached this question in terms of the particular question of political debate rhetoric, showing that considering “types” of speakers (in terms of party and status) can improve overall model performance and understanding of highly subjective statements. A critical next step is a better analysis of which types of features are best learned from the data and which should be incorporated as separate channels.

While here we focused on including information about the speaker, it is equally critical to model differences in listeners and surrounding contexts. One piece of that is changing how we think of annotation and agreement. In these studies, we have followed the traditional approach of working from annotator agreement. But, the primary assumption here is that different listeners will respond differently depending on their relation to the speaker and context.

While we primarily leave this for future work, we carried out preliminary experiments where, rather than making use of annotator agreement, we considered the union of annotator evaluations. It is plausible that, rather than signaling annotation errors, each annotator would be picking up on different aspects of the target domain leaving the union of annotations as a reasonable way to aggregate the data.

When we retrained both the baseline model and the best performing overall model (the model from experiment 3 with the addition of party and incumbency status) on this dataset, the results were extremely promising. For attack, policy, and value statements positive class F1 score rose to between 0.80 and 0.88. More surprising were the personal and ingroup categories where positive class performance rose from 0.14 and 0.47 to an average of 0.80 and 0.88. This experiment suggests that combining these categories was achieved by the model where only incumbency information was included.

This experiment suggests that combining these types of features with learned latent representations may be an effective strategy for incorporating side information. These models provided the best overall performance observed in any of our experiments. Further, this information can be included with negligible increases in training time and data preparation.

It is unclear why the incumbency information alone outperformed the combined party and incumbent information on the ingroup and personal categories. Given the low overall performance in both these categories, this could simply be noise in the training process.

Overall, the results of this experiment are extremely encouraging. It remains an open question what types of features are best introduced in this manner, but the ease of doing so opens a range of opportunities for modeling.

6 Conclusion and Future Work

The extensive social scientific literature on political rhetoric has established the importance of group identity and candidate status to the understanding of what is being said at any moment. Beyond political speech, language understanding generally depends on the interplay between who is speaking, who is listening, and what the context is. Language is more than a informational channel, it’s a form of social action (Austin, 1975) and can only be fully understood in context. Unlike tasks such as image recognition, sometimes there is no single right answer.

Computational work is beginning to address these challenges, but there is much work to be done. In particular, this area defines one of the key regions where NLP and the social sciences must intersect. Social scientific experiments provide an essential guide to modeling the dynamics of which types of information are best treated as separate features and which should be learned directly from the data. In turn, those social scientific theories can be informed by the computational studies which incorporate those insights.

In this study, we have approached this question in terms of the particular question of political debate rhetoric, showing that considering “types” of speakers (in terms of party and status) can improve overall model performance and understanding of highly subjective statements. A critical next step is a better analysis of which types of features are best learned from the data and which should be incorporated as separate channels.

While here we focused on including information about the speaker, it is equally critical to model differences in listeners and surrounding contexts. One piece of that is changing how we think of annotation and agreement. In these studies, we have followed the traditional approach of working from annotator agreement. But, the primary assumption here is that different listeners will respond differently depending on their relation to the speaker and context.

While we primarily leave this for future work, we carried out preliminary experiments where, rather than making use of annotator agreement, we considered the union of annotator evaluations. It is plausible that, rather than signaling annotation errors, each annotator would be picking up on different aspects of the target domain leaving the union of annotations as a reasonable way to aggregate the data.

When we retrained both the baseline model and the best performing overall model (the model from experiment 3 with the addition of party and incumbency status) on this dataset, the results were extremely promising. For attack, policy, and value statements positive class F1 score rose to between 0.80 and 0.88. More surprising were the personal and ingroup categories where positive class performance rose from 0.14 and 0.47 to an average of 0.80 and 0.88. This experiment suggests that combining these categories was achieved by the model where only incumbency information was included.


Robyn M Dawes, Alphans JC Van de Kragt, and John M Orbell. 1990. Cooperation for the benefit of usnot me, or my conscience.


Ian Haney López. 2015. *Dog whistle politics: How coded racial appeals have reinvented racism and wrecked the middle class*. Oxford University Press.
