Ready, Set, Argue!
Predicting Oxford Debate Winners with Content and Context Features

Girish Ganesan
Rutgers University
gg655

Abstract

In the Oxford-style debate setting where winning is judged by the opinions of a neutral audience before and after the event, can we algorithmically determine which side was likely to have changed more human minds? In this paper, I implement feature extraction for the knowledge base, tonal, and proposition type features compiled by Jo et al. which was excluded from their original work in a publically available Jupyter notebook. I use this feature extraction to train a logistic regression classifier that achieves $> 75\%$ accuracy in predicting the outcome of a debate in the IQ2 Debate Corpus, which meets and exceeds previous results.

1 Introduction

The competitive debate circuit hosts a number of different formats. Intelligence Squared is an organization that organizes Oxford-style debates between leading intellectuals and public figures on matters of social importance. The style is characterized by its uniquely democratic judging mechanism; in contrast to other styles that designate professional judges to determine which side presented better arguments and refuted attacks on their own position more effectively, the winner of an Oxford debate is whichever team swayed more audience minds, measured by a poll taken before and after the event. Specifically, the challenge faced in this paper is to predict the winning side (pro or con) of a corpus of 108 Oxford debates prepared by Zhang et al. (Zhang et al., 2016).

This is a particularly important style of debate since it closely mirrors presidential debates in the United States and debates between candidates for heads of state across the world. News media shows like Crossfire that present one speaker for the pro-side of some contentious issue and one for the con-side also rely on this format implicitly, where the ultimate judge is the television audience. Therefore, a technique which can assess Oxford debate performance could be immensely influential in the political and media spheres.

The first approach to this problem that may spring to mind is one motivated by natural language understanding (NLU). The themes of determining logical fallacy, contradiction, or entailment is one present in many NLU benchmark tasks. However, an approach based only on NLU would be extremely limited in its capacity to evaluate the effect of an argument on the human mind. Rhetorical features, what would be termed "style", would have no bearing on the underlying validity of any statement made by a debater, while they certainly can be expected to bear on how well the human audience responds to the speaker.

Previous solutions focus solely on content features, information inherent to the text, to the exclusion of context, information derived from a general awareness of the surrounding world. This work explores the richness of a larger feature space that attempts to model the general knowledge of the audience members as an important part of their response to the debaters' speeches.

To evaluate overall debate performance, I extract 26 features for each side (pro and con) at the sentence level, which are then aggregated to characterize the entire hour-long event. These features are those identified in "Detecting Attackable Sentences in Arguments" by Jo et al. as being germane to the related sentence-level tasks of assessing which claims made by an interlocutor are rhetorically attackable (Jo et al., 2020). Content-based features describe properties inherent to a statement, like proposition type (including questions and personal stories) and tonal descriptors (like sentiment score). Context-based features are knowledge-base references to the Kialo corpus, a list of posts scraped from Kialo.com (a website that

hosts online anonymous debate), to determine the frequency of an argument for example. A series of classifiers were trained on these features and after hyperparameter tuning, I achieved a testing accuracy of 76.19% in predicting the winner of a given debate.[1] Limitations of the work include the relatively small dataset provided by the IQ2 corpus of only 108 debates. Further, I expect that results would be stronger if transcripts were attached with timestamps for each sentence, allowing a model to infer more tonal information through the speaker's talking pace, or if NLU was implemented alongside the content and context features to enhance the model's logical understanding.

2 Related Work

There is one related work I found on predicting the winner of Oxford-style debates using transcript-based text features, by Brilman and Scherer (Brilman and Scherer, 2015). The main thrust of their work is to assess debate performance via a multimodal approach, employing transcript, video, and audio in their architecture. One of the main questions they set out to answer was how much each source of information adds to the model's understanding of the dynamics of Oxford debates. To answer this, they run ablation tests for each mode of data.

Their video and text classifier achieves a 66.7% accuracy and their audio and text classifier achieves a 76.6% accuracy. The first performs markedly worse than the results presented here, while the second has access to much richer information and still only exceeds the results of this paper by less than half of a percentage point. The major difference in methodology between the work presented here and the work done by Brilman and Scherer is the inclusion of contextual features. When assessing the response of a human audience to rhetoric, their familiarity with ongoing news items and relevant controversies plays into how they relate to the speakers.

3 Method

Features were first extracted at the sentence level and then aggregated to represent the debate as a whole.

3.1 Content Features

Content features are those inherent to the sentence, including proposition type and tone. These features were extracted in three different ways: by regex expression, by an off-the-shelf classifier, or by averaging scores over tokens in a sentence using values in a predefined lexicon.

Regex expressions were used to extract proposition types using keywords and key symbols.[2] For example, personal stories are characterized by words like "believed" and "think" and questions are of course characterized by the appearance of the symbol "?". Sentiment scores were calculated using the VADER sentiment analyzer available from nltk. Three external lexicons were used to determine the concreteness, subjectivity, and arousal/dominance of a sentence based on the average value of the lexicon scores on its tokens (Brysbaert et al., 2014; Wilson et al., 2005; Warriner et al., 2013). These lexicons were crowdsourced through Amazon Mechanical Turk and quality-controlled, giving each word in their respective lists a mean score for the property they were designed to evaluate.

3.2 Context Features

The introduction of external knowledge by cross-referencing statements with the scraped Kialo corpus provides a major improvement over previous results. For each given sentence S, the Kialo database is searched for posts M_i with at least 5 words shared with the sentence. Each entry in the Kialo database comes with the raw text of the post, the number of positive responses, the number of negative responses, and the number of neutral responses. Frequency is given by

$$\log_2(|M|)$$

where $|M|$ is the number of matching statements found for S. Attractiveness is given by

$$\log_2\left(\sum_{i=1}^{|M|} R(M_i)\right)$$

where $R(M_i)$ is the total number of responses to Kialo statement M_i, or the sum of the number of positive, negative, and neutral responses. Extreme-

[1] My code for feature extraction and model training is available here.

[2] A full list of these regex expressions can be found in the Jupyter notebook available here.

ness is given by

$$\sum_{i=1}^{|M|} |P(M_i) - N(M_i)|$$

where $P(M_i)$ and $N(M_i)$ are the number of positive and negative responses to Kialo statement M_i, respectively. The Kialo features embed information about the audience's previous exposure to the topic of debate to assess whether the arguments made are well-established in the broader discussion between the two sides or whether it is an invention of the speaker only and not shared by anyone else in the 130k statements collected in the Kialo dump.

3.3 Aggregation

For each side of each debate, the sum of all sentence-level features calculated is incorporated into the 52-dimensional representation of each debate (26 features for the pro side and 26 for the con side). Note that that I chose to sum these features rather than average them. Knowing the unnormalized scores for each feature as a metric of overall performance is preferable to the normalized, per-sentence scores because the latter represents some form of "efficiency" which is not relevant for the purposes of determining which side won the debate. It may be better suited to discussions of who provided the more solid performance on a more technical level, or as a metric for who made their words "count" in the conversation, but this is somewhat removed from the democratic judging scheme of the Oxford debate.

4 Experiments

After features were prepared for all 108 debates, I partitioned the corpus on a 80%-20% split for training and testing (validation is automatically performed within SGDClassifier from sklearn). I experimented with four hyperparameters during tuning: loss, regularization type, regularization strength, and whether to apply early stopping.

4.1 Early Stopping and Regularization Strength

Final testing accuracy values are shown below logistic regression models trained with L2 regularization for different values for the regularization strength parameter α and the options of early stopping or no early stopping.

	Early Stop	No Early Stop
$\alpha = 0.1$	52.38%	52.38%
$\alpha = 0.01$	76.19%	61.90%
$\alpha = 0.001$	57.14%	61.90%
$\alpha = 0.0001$	76.19%	71.43%

Early stopping is a technique used to improve the generalization of the logistic regression classifier to data it is not exposed to during training, so it is expected that the testing accuracy should increase when it turned on. This holds generally true for the results above. For the fixed setting mentioned above, decreasing the regularization strength actually improves performance on this range of values for α.

4.2 Loss and Regularization Type

Testing accuracies are shown for models with $\alpha = 0.0001$ and early stopping enabled.

	L1	L2
Hinge loss	66.67%	71.43%
Log loss	57.14%	76.19%

The choice of hinge loss or log loss amounts to choosing between training a linear SVM or a logistic regression model. For both hinge loss and log loss, L1 regularization was inferior in testing accuracy than L2. L1 more often converges to sparse coefficient matrices, so it is likely that training with L1 regularization was pushing the model to disregard certain features (by setting their coefficient to 0) which eventually hurt its downstream performance.

4.3 Regularization Type and Regularization Strength

Testing accuracies are provided for models trained on log loss with early stopping.

	L1	L2
$\alpha = 0.1$	47.62%	52.38%
$\alpha = 0.01$	57.14%	76.19%
$\alpha = 0.001$	57.14%	57.14%
$\alpha = 0.0001$	57.14%	76.19%

Once the regularization strength is small enough (less than 1e-2), L1 regularization ceases to have much of an impact on the trained model's final testing accuracy. Moreover, L1 regularization with higher α actually performs poorer than with lower α on the downstream task, further suggesting that the L1 penalty is causing the coefficient matrix to become unnecessarily sparse. This may be why L2 regularization outperforms L1 regularization for every setting presented.

Training a logistic regression classifier by stochastic gradient descent (SGD) with L2 regularization and regularization strength set to 0.0001 with early stopping generalized the best to the test set, achieving an accuracy of 76.19%. This surpasses the video and transcript benchmark established by Brilman and Scherer and narrowly falls short of their audio and transcript benchmark.

5 Conclusions

The work presented here is a simple, computationally-efficient technique for accomplishing a challenging tasks: approximating not the structural integrity of an argument but the real-world effect a line of reasoning could have on a human mind. By spending time developing the right content and context features, a mere logistic regression is sufficient to predict the winner of an Oxford-style debate with $> 75\%$ accuracy, which surpasses previous results on the same task while also using unimodal data.

Future Work

One major direction I see for this work is in the computational study of rhetoric. If the work done in this paper is to be used to develop an understanding on how to sway an audience effectively, providing insights for community leaders and political aspirants, more work will have to be done on isolating the effect of each class of feature on the overall debate performance, perhaps through an ablation study. Moreover, the nature of the audience itself must be taken into account if this model is to be applied to evaluate other Oxford-style debates. The IQ2 debate audience is broadly middle-class and college-educated, while no such homogeneous assumption can be made of the audience to the presidential debates, for example. In essence, this amounts to extending context features to understand the background and inclinations of the audience in addition to the general knowledge that they possess. Finally, the Kialo feature collection process could be further refined to extract sentences that match a speaker's statement using a more involved condition than merely sharing some number of words. This would be an opportunity to blend natural language understanding (NLU) techniques with the manual feature collection process already explored.

References

Maarten Brilman and Stefan Scherer. 2015. A multimodal predictive model of successful debaters or how i learned to sway votes. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, page 149–158, New York, NY, USA. Association for Computing Machinery.

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.

Yohan Jo, Seojin Bang, Emaad Manzoor, Eduard H. Hovy, and Chris Reed. 2020. Detecting attackable sentences in arguments. *CoRR*, abs/2010.02660.

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 human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in oxford-style debates. *CoRR*, abs/1604.03114.