Political Argumentation and Attitude Change in Online Interactions

Supplementary Information

Isaac D. Mehlhaff*

August 31, 2022

Contents

*The University of North Carolina at Chapel Hill; mehlhaff@live.unc.edu.

S1 Classifier Training and Details

S1.1 Training Data

I rely on two corpora to train the classifiers in this paper. For four of the five tasks, I use the Internet Argument Corpus (IAC), a collection of posts extracted from several online debate and discussion forums very similar to `r/ChangeMyView` (Abbott, Ecker, et al. 2016; Walker et al. 2012). Using a corpus gathered from sources similar to those on which the model is used for inference helps ameliorate concerns about the classifiers' applicability to different contexts. The discussions in the corpus cover a variety of controversial topics relevant to politics and social life in the United States, such as same-sex marriage, gun control, and the existence of God. This diversity of issues is especially useful for training domain-general classifiers, as it prevents the models from over-fitting on words or phrases relevant to specific topics.

Each post is annotated by five to seven human coders on each characteristic. Each coder assigns each document a scalar value in $[-5, 5]$ on each characteristic, and all coders' scores are then averaged to get the final real-valued score reported in the corpus.[1] The authors report that the coders found the assignment of these scores rather difficult and highly subjective, reflecting the often-idiosyncratic nature of debate and argumentation as well as the difficulty of argument mining. Across all topics, however, coders nevertheless achieve an average Cohen's κ of 0.47, a value indicating moderate agreement (Landis and Koch 1977).

I take data for the final task (argument quality) from IBM-Rank-30k, a corpus of approximately 24,000 crowd-sourced arguments across a similarly diverse set of 71 common topics (Gretz et al. 2020). Ten human coders assign each argument a binary value indicating whether they find it a satisfactory argument for a particular viewpoint, regardless of their personal opinion. Two ranking algorithms then translate these binary annotations into a continuous value of argument quality in $[0, 1]$. Across all topics, the authors report an average Cohen's κ of 0.83, a value indicating strong agreement. Additional information on data preparation is included in the following subsection.

A wide array of studies have used the IAC to construct unique tasks (Galitsky, Ilvovsky, and Pisarevskaya 2018; Hartmann et al. 2019; Misra, Ecker, and Walker 2016) and train models (Lukin et al. 2017; Misra and Walker 2013; Oraby, Harrison, et al. 2016). One of the tasks I pursue here has previous state-of-the-art performance benchmarks: On the disagreement classification task, Abbott, Walker, et al. (2011) achieve an accuracy of 0.682 and Wang and Cardie (2014) achieve an F1 score of 0.636.

[1] Snow et al. (2008) show that taking the mean of scalar annotations reduces noise in evaluations given by non-expert human coders.

S1.2 Data Preparation

To prepare the data for a classification task, I first need to convert the real-valued annotations to binary labels. The simplest way to do this would be to assign a 0 to all documents less than the scale midpoint (0 for the IAC tasks and 0.5 for the argument quality task) and a 1 to all documents greater than the scale midpoint. Unfortunately, this strategy would likely create more problems than it would solve. Although the scale midpoint theoretically represents the dividing line between, for example, whether a document expresses agreement or disagreement, it likely does not represent such a clear-cut demarcation in practice. Annotators likely have different implicit understandings of how each value in the scale maps onto the concept they are annotating—a source of bias known in survey research as differential item functioning (Stegmueller 2011). The "true" dividing line between class labels is likely to be somewhere around the scale midpoint, but not the scale midpoint exactly. Any choice of a hard cutoff is therefore arbitrary and would introduce an additional source of bias into the class labels. Additionally, it is difficult to know why documents in the middle of the scale receive the score they do. For example, documents may score close to the scale midpoint because they express *both* agreement and disagreement, because they express *neither* agreement nor disagreement, because it is difficult to accurately gauge their relative degree of disagreement, or because coders simply disagree with each other. Training classifiers with such noisy class labels is not desirable.

I therefore follow the practice of Oraby, Reed, et al. (2015) and remove documents scoring in $[-1, 1]$ on the IAC tasks and in $[0.4, 0.6]$ on the argument quality task. Documents are then dichotomized after this middle range has been removed. Table S1 provides descriptive statistics of the data used for all five tasks, with the total N and class balance representing the final, dichotomized corpora. Eighty percent of the data are used for model training, with ten percent set aside for validation and a further ten percent for the final test set.

Table S1: Descriptive Statistics of Document Annotations

Task	N	Range	Mean	SD	Class Balance
Disagreement	28,171	[-5, 5]	-0.916	1.689	0.802 / 0.198
Object of Address	5,997	[-5, 5]	-1.271	2.09	0.816 / 0.184
Scope of Argument	5,603	[-5, 5]	-0.671	2.245	0.673 / 0.327
Counterargue vs. Rebut	5,831	[-5, 5]	-0.479	2.293	0.625 / 0.375
Question vs. Assert	6,216	[-5, 5]	0.717	2.368	0.319 / 0.681
Quality of Argument	24,055	[0, 1]	0.83	0.182	0.062 / 0.938

S1.3 Feature Extraction

I conduct feature extraction for the deep learning models with bidirectional encoder representations from transformers (BERT), a neural network architecture that relies on self-attention mechanisms to relate different portions of a document to each other in order to represent the document as a whole (Devlin et al. 2019; Vaswani et al. 2017). I use the base BERT model, which contains twelve encoding layers, twelve attention heads, and 110 million parameters and has been pre-trained on English Wikipedia and the BooksCorpus (Zhu et al. 2015), which collectively provide a training corpus of over 3.3 billion words. The precise design and function of BERT's architecture is beyond the scope of this paper, but it is useful to highlight a key benefit it imparts to NLP applications.

BERT is a deeply bidirectional model, meaning that it learns the meaning of a word from the context it appears in, and this context can be imparted by words appearing both before and after the target word. This attention to context closely represents how the human brain understands and deciphers language, and it is critical in building software to understand human speech. Word embedding models such as Word2Vec (Mikolov et al. 2013)—a popular choice in political science for those wishing to go beyond "bag of words" approaches (Rodriguez and Spirling 2022)—are non-contextual; they calculate a single embedding representation for each token regardless of how it contributes to the meaning of a sentence or phrase. Unidirectional models like OpenAI's GPT (Radford et al. 2018) "read" text from left to right and draw context from the words that come before the target word. Being bidirectional, BERT improves upon these approaches by drawing context from both sides of each target word.

In addition to achieving state-of-the-art results in eleven common NLP tasks (Devlin et al. 2019), BERT is used in a wide variety of high-profile products such as Google Search, and it served as a springboard for even more advanced large language models like LaMDA. Scholars working on argument mining have also begun exploring the potential of BERT (Chakrabarty et al. 2019; Zhang, Lillis, and Nulty 2021). Huning, Mechtenberg, and Wang (2021) compare BERT to structural features on the task of argumentation detection and find that BERT offers the best performance.

S1.4 Classifier Architectures and Training Details

There is no universal classifier in machine learning; no one model will perform best across all tasks (Grimmer, Roberts, and Stewart 2021; Grimmer and Stewart 2013). I therefore test six classifiers in the deep learning pipeline. Each is detailed in this subsection along with details on any necessary hyperparameter tuning. All classifiers incorporate threshold tuning—calculated as the threshold which maximizes

the difference between the true and false positive rates—and early stopping after one iteration with no improvement.

The first and perhaps simplest classifier is a logistic regression with no regularization. The second is a support vector machine with stochastic gradient descent and a logistic loss function. Next is a series of tree-based classifiers, which have been used to great effect in recent political science work (Kaufman, Kraft, and Sen 2019; Montgomery and Olivella 2018). I test three such models: a random forest with 100 trees of unlimited depth, a Gini loss function, and no pruning; an extra-randomized trees classifier, also with 100 trees of unlimited depth, a Gini loss function, and no pruning; and a random forest with extreme gradient boosting, ten trees with a maximum depth of ten nodes, a logistic objective function, and L2 regularization.[2]

The final classifier is a fully connected sigmoid layer appended to the end of the BERT model, with all weights fine-tuned on each task.[3] The full neural network is then trained with binary cross-entropy loss, a fully connected dropout layer prior to the sigmoid layer, a learning rate with scheduled linear decay, and an AdamW optimizer (Loshchilov and Hutter 2019).

Table S2 presents the results of a grid search over three hyperparameters in the random forests with extreme gradient boosting: the learning rate, the proportion of data sampled in each tree, and the proportion of data sampled at each node. All three hyperparameters were allowed to take values in {0.2, 0.4, 0.6, 0.8}. Table S3 presents the results of a grid search over five hyperparameters in the fine-tuned BERT neural networks: the initial learning rate in {0.00005, 0.00001, 0.00015}, the weight decay rate in {0.01, 0.05, 0.1}, the proportion of the training data used for warm-up in {0.05, 0.1, 0.2}, the proportion of nodes dropped by the dropout layer in {0.2, 0.3, 0.4}, and the batch size in {32, 64, 128}.

Table S2: Results of Hyperparameter Tuning in Random Forests with Extreme Gradient Boosting

Task	Learning Rate	Tree Subsample	Node Subsample
Disagreement	0.4	0.4	0.8
Object of Address	0.2	0.6	0.6
Scope of Argument	0.2	0.2	0.8
Counterargue vs. Rebut	0.2	0.4	0.4
Question vs. Assert	0.8	0.8	0.6
Quality of Argument	0.8	0.8	0.6

[2]As a matter of computational resource constraints, the random forest with extreme gradient boosting contains fewer and more shallow trees compared to the other two tree-based classifiers. Results presented below may therefore be conservative, and could potentially be pushed higher by using larger forests.

[3]Fully training a neural network is much more computationally expensive than simply using it for inference and passing extracted features to a separate classifier. For this fine-tuned classifier, I therefore use the small BERT model (two hidden layers and two attention heads). Although this choice could result in more conservative performance metrics, it does not seem likely based on past work. Turc et al. (2019) show that small BERT models perform comparably to base BERT models and Kovaleva et al. (2019) show that base BERT is overparameterized, suggesting that decreasing the number and size of its hidden layers is not likely to have a drastic effect.

Table S3: Results of Hyperparameter Tuning in Fine-Tuned BERT Neural Networks

Task	Initial Learning Rate	Weight Decay Rate	Warm-Up Partition	Dropout Proportion	Batch Size
Disagree / Agree	0.00015	0.05	0.05	0.4	32
Object of Address	0.00005	0.05	0.2	0.4	128
Scope of Argument	0.00005	0.05	0.2	0.4	128
Counterargue vs. Rebut	0.00015	0.1	0.05	0.4	32
Question vs. Assert	0.00015	0.01	0.2	0.2	128
Quality of Argument	0.0001	0.1	0.1	0.2	32

To benchmark the performance of these classifiers, I use two baselines, one naïve and one lexical. The naïve baseline uses no feature extraction or model at all, and merely reports performance metrics that result from randomly guessing class labels. The lexical baseline performs feature extraction with unigrams, a standard method of extracting information from text and perhaps the most popular approach in political science applications (Grimmer and Stewart 2013; Monroe, Colaresi, and Quinn 2008; Quinn et al. 2010). To extract unigrams, I follow the standard practice of removing common stop words (i.e. words like "or," "the," or "is" that appear throughout documents of all types and carry little to no meaning)[4] and implementing word stemming, which reduces the total number of unique tokens by shortening each word in the corpus to its root (i.e. collapsing "legislative," "legislation," and "legislator" under the common stem "legislat"). I then convert each document in the corpus to a sparse vector of binary token indicators, indicating whether or not each word occurs in each document. The lexical baseline uses a support vector machine with stochastic gradient descent and a logistic loss function.

S1.5 Performance Metrics

This section evaluates the performance of each model in each task's test set. I begin by examining precision and recall scores. High precision indicates a low false-positive rate, while high recall indicates a low false-negative rate. A model with high precision but low recall is therefore correct most of the time when it predicts a positive label, but it predicts too few of them relative to the true labels. A model with high recall but low precision, on the other hand, predicts many positive labels, but most of those predictions are wrong. An ideal model would have both high precision and high recall, indicating that it captures most of the true positive labels, and those predictions are mostly accurate.

Recall from Table S1 that most tasks have a relatively unbalanced distribution of observations between classes. This class imbalance can make some performance metrics misleading (Williams 2021), so I report

[4] I preserve a range of stop words that would normally be removed but have been shown to be important for identifying disagreement and other relevant concepts in argument mining (Walker et al. 2012). These include words like "because," "then," and "so."

the weighted versions of both precision and recall, which combine scores from both classes and weight them by their support, as opposed to reporting only the positive-class metric. Weighted recall is mathematically equivalent to the overall accuracy of the classifier, so I do not present accuracy as a separate metric.

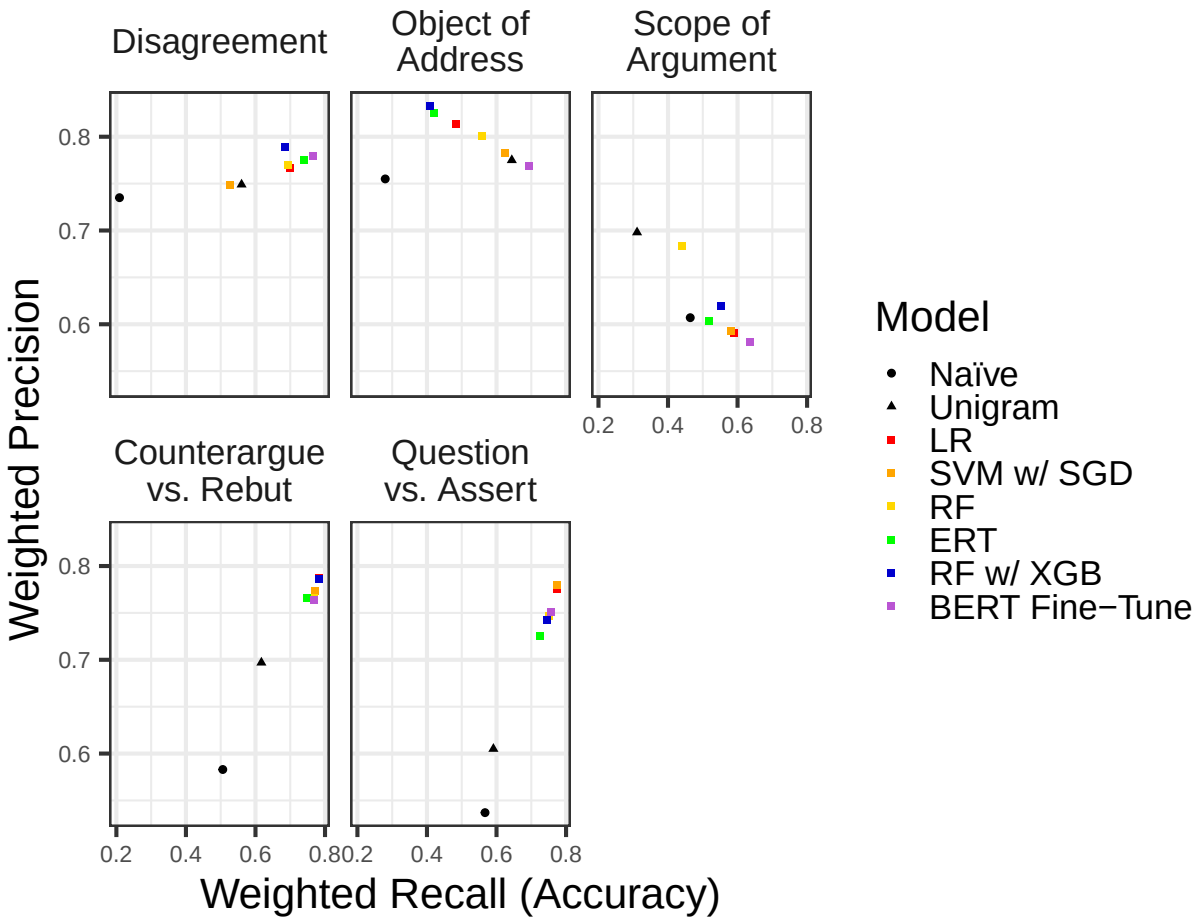

Figure S1: Weighted Precision and Recall. Weighted recall is mathematically equivalent to accuracy. Black points represent baselines, colored points represent deep learning models.

Figure S1 plots each model's precision against its recall on each task. Black points represent the two baselines, while colored points represent deep learning models. A better-performing model will appear closer to the upper-right corner of the plot, as this would indicate a model that both correctly identifies all relevant cases and avoids identifying irrelevant ones. On most tasks, the deep learning models outperform both baselines, achieving both higher precision and higher recall (and, by extension, higher accuracy). Across all tasks, the random forest with extreme gradient boosting and the fine-tuned BERT neural network generally perform best.

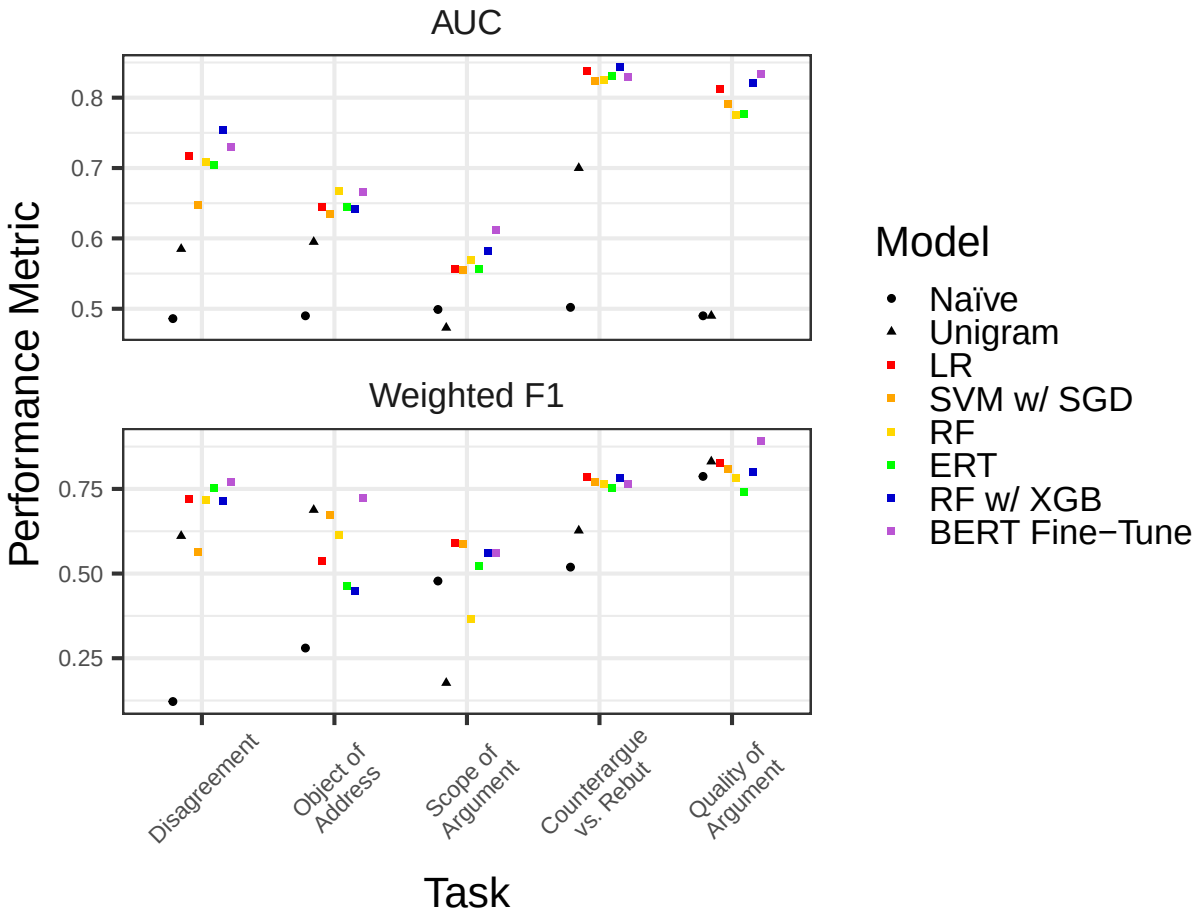

Figure S2: Area Under the ROC Curve and Weighted F1 Score. Black points represent baselines, colored points represent deep learning models.

Precision and recall offer important information for understanding classifier performance, but it is often difficult to evaluate the tradeoff between them. Two alternate metrics provide more easily evaluable summary measures of classifier performance. Receiver operating characteristic (ROC) curves plot the true positive rate against the false positive rate, showing the predictive skill of a classifier as the classification threshold varies. The area under the ROC curve (AUC) then acts as a metric of how well the classifier can distinguish between class labels, with a higher AUC indicating better performance. The second metric, the F1 score, is the harmonic mean of precision and recall.[5] Raw accuracy metrics can be misleading when class labels are unbalanced, so the F1 score is often used to provide a more robust measure of a classifier's accuracy.

These two metrics are shown in Figure S2, with the AUC plotted in the top facet and the F1 score in the bottom facet. Again, black points represent the two baselines, while colored points represent the deep

[5]For the same reasons as above, I also present the weighted F1 score.

learning models. Results largely mirror those in Figure S1, but the benefit of deep learning over a lexical model is much clearer here.

S2 Class Frequencies by Commenter Type

Figure S3 displays the proportion of each type of commenter's posts that are coded with each class label. This figure corresponds to Figure 3 in the main text. Participants are defined as commenters who post at least twice in the comment forest, while lurkers post only once, typically to award a delta.

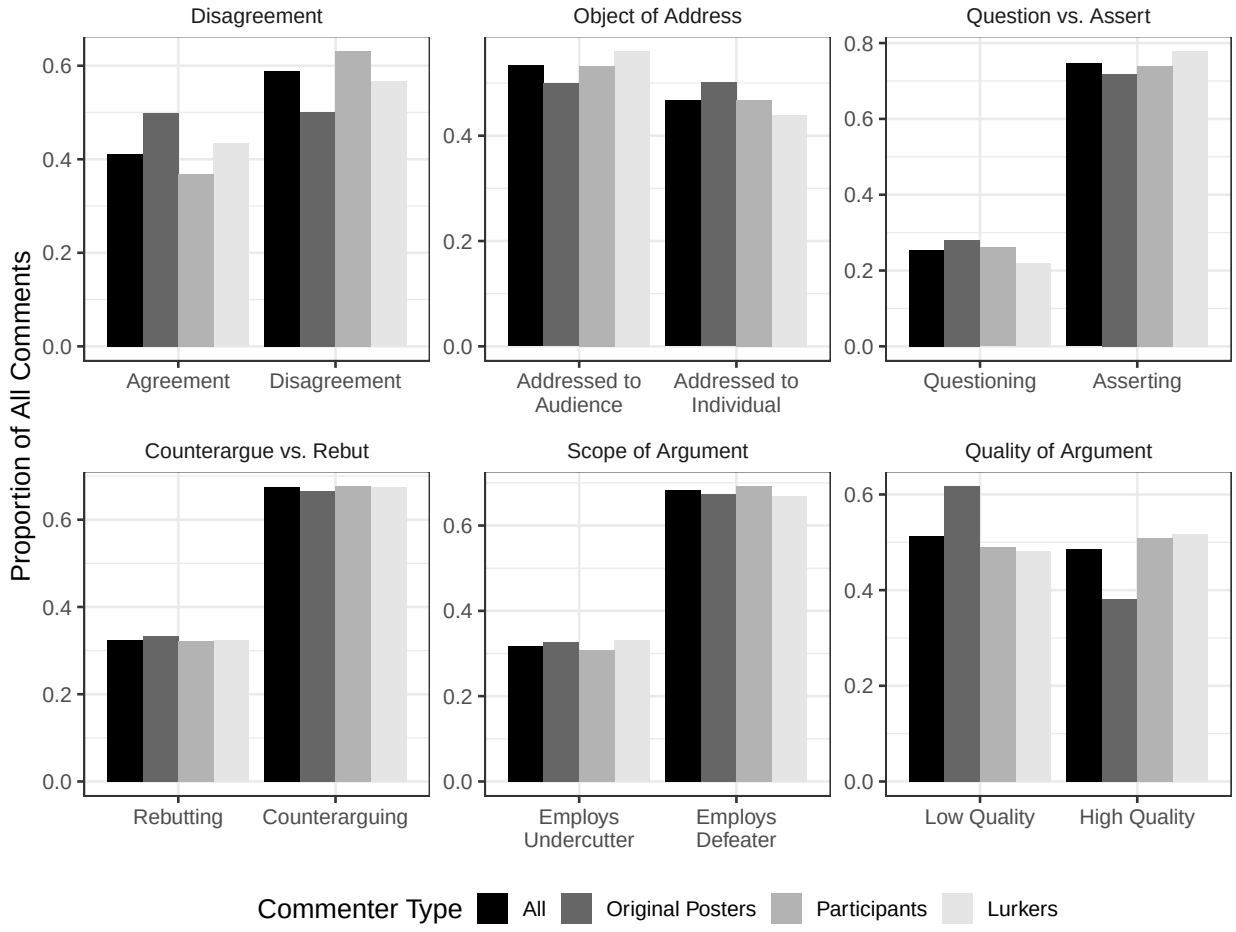

Figure S3: Class Frequencies of Argumentation Characteristics. Frequencies are broken down by commenter type.

S3 Full Model Results

This section presents full model results in tabular format. Those in section S3.2 correspond to results presented graphically in the main text. Results in section S3.1 use the full dataset and results in section S3.3, like those in the main text, also use a subsample of the data to guard against the possibility that the full-data results may be overpowered, but these models use data subsampled at the level of the post instead of the comment. In this sampling schema, entire posts are therefore either included or excluded as opposed to most posts being incomplete, as in the comment-level sampling schema.

S3.1 Results with Full Data

Table S4: Effect of Disagreement on Attitude Change (Full Sample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.145*	0.159*	−0.079	−0.032
	(0.016)	(0.016)	(0.069)	(0.099)
Disagreement	0.257*	0.263*	0.136*	0.213*
	(0.004)	(0.004)	(0.024)	(0.027)
Author Deltas	−0.686*	−0.752*	0.081*	−0.962*
	(0.011)	(0.011)	(0.033)	(0.081)
Depth	0.133*	0.113*	0.081*	0.096*
	(0.003)	(0.003)	(0.005)	(0.004)
Score	−4.376*	−4.491*	−7.066*	−8.203*
	(0.014)	(0.014)	(0.053)	(0.092)
Observations	1,025,857	1,025,857	1,025,857	1,025,857
Log Likelihood	−85,172.290	−80,256.050	−6,876.590	−3,521.771
Akaike Inf. Crit.	170,354.600	160,522.100	13,763.180	7,053.542

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S5: Effect of Directly Addressing an Individual on Attitude Change (Full Sample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.385*	0.401*	0.234*	0.308*
	(0.019)	(0.019)	(0.086)	(0.121)
Direct Address	0.248*	0.254*	0.126*	0.204*
	(0.004)	(0.004)	(0.025)	(0.027)
Author Deltas	−0.695*	−0.763*	0.075*	−0.971*
	(0.011)	(0.011)	(0.033)	(0.081)
Depth	0.132*	0.112*	0.080*	0.096*
	(0.003)	(0.003)	(0.005)	(0.004)
Score	−4.363*	−4.473*	−7.154*	−8.280*
	(0.010)	(0.011)	(0.039)	(0.076)
Observations	1,025,857	1,025,857	1,025,857	1,025,857
Log Likelihood	−85,012.490	−80,097.750	−6,873.714	−3,518.855
Akaike Inf. Crit.	170,035.000	160,205.500	13,757.430	7,047.709

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S6: Effect of Asserting an Idea on Attitude Change (Full Sample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	−0.945*	−1.008*	−0.325*	−0.768*
	(0.025)	(0.026)	(0.086)	(0.151)
Assert	0.253*	0.258*	0.131*	0.206*
	(0.004)	(0.004)	(0.024)	(0.027)
Author Deltas	−0.653*	−0.717*	0.088*	−0.928*
	(0.011)	(0.011)	(0.033)	(0.080)
Depth	0.131*	0.111*	0.080*	0.095*
	(0.003)	(0.003)	(0.005)	(0.004)
Score	−4.115*	−4.214*	−7.039*	−8.070*
	(0.010)	(0.011)	(0.039)	(0.075)
Observations	1,025,857	1,025,857	1,025,857	1,025,857
Log Likelihood	−84,287.960	−79,352.950	−6,869.578	−3,506.086
Akaike Inf. Crit.	168,585.900	158,715.900	13,749.160	7,022.172

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S7: Effect of Presenting a Counterargument on Attitude Change (Full Sample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.826*	0.885*	0.325*	0.708*
	(0.021)	(0.022)	(0.079)	(0.130)
Counterargue	0.250*	0.255*	0.129*	0.203*
	(0.004)	(0.004)	(0.024)	(0.027)
Author Deltas	−0.653*	−0.717*	0.089*	−0.928*
	(0.011)	(0.011)	(0.033)	(0.080)
Depth	0.130*	0.110*	0.080*	0.096*
	(0.003)	(0.003)	(0.005)	(0.004)
Score	−4.908*	−5.063*	−7.343*	−8.744*
	(0.020)	(0.021)	(0.068)	(0.127)
Observations	1,025,857	1,025,857	1,025,857	1,025,857
Log Likelihood	−84,285.530	−79,333.900	−6,868.202	−3,504.689
Akaike Inf. Crit.	168,581.100	158,677.800	13,746.400	7,019.377

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S8: Effect of Employing a Defeater on Attitude Change (Full Sample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.997*	1.044*	0.496*	0.850*
	(0.022)	(0.023)	(0.083)	(0.136)
Defeater	0.248*	0.253*	0.126*	0.201*
	(0.004)	(0.004)	(0.024)	(0.027)
Author Deltas	−0.662*	−0.726*	0.082*	−0.932*
	(0.011)	(0.011)	(0.033)	(0.080)
Depth	0.131*	0.110*	0.080*	0.095*
	(0.003)	(0.003)	(0.005)	(0.004)
Score	−5.055*	−5.201*	−7.476*	−8.862*
	(0.021)	(0.022)	(0.073)	(0.133)
Observations	1,025,857	1,025,857	1,025,857	1,025,857
Log Likelihood	−83,930.310	−79,016.500	−6,857.632	−3,498.175
Akaike Inf. Crit.	167,870.600	158,043.000	13,725.260	7,006.351

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S9: Effect of Argument Quality on Attitude Change (Full Sample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.578*	0.605*	0.252*	0.369*
	(0.016)	(0.017)	(0.069)	(0.101)
High Quality	0.247*	0.252*	0.129*	0.203*
	(0.004)	(0.004)	(0.024)	(0.027)
Author Deltas	−0.652*	−0.715*	0.088*	−0.935*
	(0.011)	(0.011)	(0.033)	(0.080)
Depth	0.132*	0.112*	0.080*	0.095*
	(0.003)	(0.003)	(0.005)	(0.004)
Score	−4.608*	−4.730*	−7.244*	−8.414*
	(0.014)	(0.015)	(0.052)	(0.092)
Observations	1,025,857	1,025,857	1,025,857	1,025,857
Log Likelihood	−84,545.310	−79,622.690	−6,870.528	−3,515.004
Akaike Inf. Crit.	169,100.600	159,255.400	13,751.060	7,040.009

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

S3.2 Results with Comment-Level Sampling

Table S10: Effect of Disagreement on Attitude Change (Comment Subsample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.142*	0.170*	−0.310	−0.074
	(0.050)	(0.052)	(0.204)	(0.296)
Disagreement	0.246*	0.256*	0.099	0.073
	(0.013)	(0.014)	(0.079)	(0.115)
Author Deltas	−0.700*	−0.817*	0.206*	−0.412*
	(0.034)	(0.037)	(0.095)	(0.179)
Depth	0.174*	0.133*	0.061*	0.097*
	(0.013)	(0.011)	(0.014)	(0.014)
Score	−4.373*	−4.535*	−6.817*	−7.739*
	(0.043)	(0.047)	(0.148)	(0.234)
Observations	102,600	102,600	102,600	102,600
Log Likelihood	−8,536.302	−7,957.950	−755.990	−390.020
Akaike Inf. Crit.	17,082.600	15,925.900	1,521.979	790.041

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S11: Effect of Directly Addressing an Individual on Attitude Change (Comment Subsample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.466*	0.475*	0.416	0.309
	(0.057)	(0.059)	(0.242)	(0.362)
Direct Address	0.234*	0.245*	0.079	0.061
	(0.014)	(0.014)	(0.081)	(0.117)
Author Deltas	−0.711*	−0.829*	0.192*	−0.418*
	(0.034)	(0.038)	(0.095)	(0.179)
Depth	0.172*	0.132*	0.061*	0.097*
	(0.013)	(0.011)	(0.013)	(0.014)
Score	−4.380*	−4.527*	−7.070*	−7.839*
	(0.033)	(0.036)	(0.117)	(0.177)
Observations	102,600	102,600	102,600	102,600
Log Likelihood	−8,509.887	−7,933.991	−755.878	−389.806
Akaike Inf. Crit.	17,029.770	15,877.980	1,521.757	789.612

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S12: Effect of Asserting an Idea on Attitude Change (Comment Subsample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	−0.996*	−1.100*	−0.192	−0.406
	(0.079)	(0.086)	(0.244)	(0.380)
Assert	0.244*	0.254*	0.092	0.065
	(0.013)	(0.014)	(0.079)	(0.116)
Author Deltas	−0.668*	−0.782*	0.204*	−0.397*
	(0.034)	(0.037)	(0.095)	(0.178)
Depth	0.170*	0.130*	0.059*	0.097*
	(0.013)	(0.011)	(0.014)	(0.014)
Score	−4.108*	−4.240*	−6.945*	−7.693*
	(0.031)	(0.034)	(0.117)	(0.175)
Observations	102,600	102,600	102,600	102,600
Log Likelihood	−8,437.570	−7,853.554	−756.743	−389.315
Akaike Inf. Crit.	16,885.140	15,717.110	1,523.486	788.629

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S13: Effect of Presenting a Counterargument on Attitude Change (Comment Subsample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.855*	0.955*	0.100	0.260
	(0.066)	(0.071)	(0.221)	(0.332)
Counterargue	0.240*	0.250*	0.091	0.063
	(0.013)	(0.014)	(0.080)	(0.116)
Author Deltas	−0.669*	−0.782*	0.202*	−0.402*
	(0.034)	(0.037)	(0.095)	(0.179)
Depth	0.170*	0.131*	0.060*	0.098*
	(0.013)	(0.011)	(0.014)	(0.014)
Score	−4.933*	−5.160*	−7.058*	−7.965*
	(0.062)	(0.068)	(0.186)	(0.290)
Observations	102,600	102,600	102,600	102,600
Log Likelihood	−8,440.744	−7,853.264	−756.979	−389.678
Akaike Inf. Crit.	16,891.490	15,716.530	1,523.958	789.355

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S14: Effect of Employing a Defeater on Attitude Change (Comment Subsample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	1.081*	1.131*	0.467	1.065*
	(0.072)	(0.076)	(0.244)	(0.423)
Defeater	0.234*	0.244*	0.085	0.045
	(0.013)	(0.014)	(0.080)	(0.118)
Author Deltas	−0.674*	−0.788*	0.201*	−0.394*
	(0.033)	(0.037)	(0.094)	(0.176)
Depth	0.171*	0.130*	0.059*	0.099*
	(0.013)	(0.011)	(0.014)	(0.014)
Score	−5.127*	−5.312*	−7.331*	−8.614*
	(0.068)	(0.073)	(0.216)	(0.398)
Observations	102,600	102,600	102,600	102,600
Log Likelihood	−8,394.540	−7,819.448	−755.039	−385.758
Akaike Inf. Crit.	16,799.080	15,648.900	1,520.078	781.516

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S15: Effect of Argument Quality on Attitude Change (Comment Subsample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.608*	0.639*	0.460*	−0.300
	(0.051)	(0.054)	(0.208)	(0.294)
High Quality	0.235*	0.245*	0.082	0.082
	(0.013)	(0.014)	(0.080)	(0.114)
Author Deltas	−0.667*	−0.781*	0.213*	−0.427*
	(0.034)	(0.037)	(0.095)	(0.180)
Depth	0.171*	0.130*	0.058*	0.097*
	(0.013)	(0.012)	(0.014)	(0.014)
Score	−4.626*	−4.789*	−7.241*	−7.647*
	(0.044)	(0.047)	(0.163)	(0.201)
Observations	102,600	102,600	102,600	102,600
Log Likelihood	−8,466.101	−7,887.993	−754.630	−389.525
Akaike Inf. Crit.	16,942.200	15,785.990	1,519.259	789.050

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

S3.3 Results with Post-Level Sampling

Table S16: Effect of Disagreement on Attitude Change (Post Subsample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.109*	0.122*	0.040	0.176
	(0.048)	(0.050)	(0.203)	(0.294)
Disagreement	0.270*	0.276*	0.185*	0.075
	(0.012)	(0.013)	(0.062)	(0.115)
Author Deltas	−0.665*	−0.730*	0.045	−0.805*
	(0.032)	(0.034)	(0.098)	(0.221)
Depth	0.134*	0.115*	0.111*	0.144*
	(0.009)	(0.009)	(0.012)	(0.013)
Score	−4.327*	−4.440*	−7.038*	−8.144*
	(0.041)	(0.044)	(0.160)	(0.269)
Observations	108,656	108,656	108,656	108,656
Log Likelihood	−9,181.185	−8,643.778	−792.681	−390.629
Akaike Inf. Crit.	18,372.370	17,297.560	1,595.361	791.258

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S17: Effect of Directly Addressing an Individual on Attitude Change (Post Subsample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.378*	0.386*	0.270	0.407
	(0.056)	(0.058)	(0.243)	(0.350)
Direct Address	0.261*	0.267*	0.177*	0.063
	(0.012)	(0.013)	(0.062)	(0.117)
Author Deltas	−0.677*	−0.742*	0.040	−0.815*
	(0.032)	(0.034)	(0.099)	(0.223)
Depth	0.133*	0.114*	0.110*	0.144*
	(0.009)	(0.009)	(0.013)	(0.013)
Score	−4.333*	−4.441*	−7.063*	−8.115*
	(0.031)	(0.033)	(0.113)	(0.210)
Observations	108,656	108,656	108,656	108,656
Log Likelihood	−9,162.939	−8,626.477	−792.190	−390.328
Akaike Inf. Crit.	18,335.880	17,262.950	1,594.381	790.656

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S18: Effect of Asserting an Idea on Attitude Change (Post Subsample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	−0.920*	−0.967*	−0.408	−0.568
	(0.074)	(0.079)	(0.258)	(0.411)
Assert	0.266*	0.272*	0.182*	0.071
	(0.012)	(0.013)	(0.062)	(0.116)
Author Deltas	−0.634*	−0.696*	0.058	−0.778*
	(0.032)	(0.034)	(0.098)	(0.221)
Depth	0.133*	0.114*	0.110*	0.143*
	(0.009)	(0.009)	(0.012)	(0.013)
Score	−4.092*	−4.192*	−6.925*	−7.916*
	(0.030)	(0.032)	(0.112)	(0.205)
Observations	108,656	108,656	108,656	108,656
Log Likelihood	−9,087.527	−8,550.302	−791.244	−389.618
Akaike Inf. Crit.	18,185.050	17,110.600	1,592.488	789.235

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S19: Effect of Presenting a Counterargument on Attitude Change (Post Subsample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.701*	0.773*	0.111	−0.008
	(0.061)	(0.065)	(0.218)	(0.319)
Counterargue	0.263*	0.269*	0.184*	0.077
	(0.012)	(0.013)	(0.062)	(0.115)
Author Deltas	−0.636*	−0.697*	0.050	−0.803*
	(0.032)	(0.034)	(0.099)	(0.222)
Depth	0.131*	0.112*	0.111*	0.144*
	(0.009)	(0.009)	(0.012)	(0.013)
Score	−4.781*	−4.945*	−7.090*	−8.029*
	(0.057)	(0.061)	(0.184)	(0.289)
Observations	108,656	108,656	108,656	108,656
Log Likelihood	−9,108.326	−8,563.769	−792.560	−390.830
Akaike Inf. Crit.	18,226.650	17,137.540	1,595.119	791.660

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S20: Effect of Employing a Defeater on Attitude Change (Post Subsample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.908*	0.937*	0.493*	0.626
	(0.065)	(0.068)	(0.243)	(0.369)
Defeater	0.261*	0.266*	0.178*	0.065
	(0.012)	(0.013)	(0.062)	(0.116)
Author Deltas	−0.644*	−0.706*	0.049	−0.785*
	(0.032)	(0.034)	(0.098)	(0.220)
Depth	0.134*	0.114*	0.110*	0.143*
	(0.010)	(0.009)	(0.012)	(0.013)
Score	−4.957*	−5.087*	−7.378*	−8.501*
	(0.062)	(0.065)	(0.216)	(0.353)
Observations	108,656	108,656	108,656	108,656
Log Likelihood	−9,066.030	−8,531.456	−790.354	−389.096
Akaike Inf. Crit.	18,142.060	17,072.910	1,590.707	788.191

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

Table S21: Effect of Argument Quality on Attitude Change (Post Subsample)

	Dependent variable:			
	Received a Delta			
	All	Original Posters	Participants	Lurkers
	(1)	(2)	(3)	(4)
Intercept	0.581*	0.611*	0.165	0.355
	(0.049)	(0.051)	(0.199)	(0.292)
High Quality	0.258*	0.263*	0.182*	0.065
	(0.012)	(0.013)	(0.062)	(0.117)
Author Deltas	−0.632*	−0.693*	0.052	−0.784*
	(0.032)	(0.034)	(0.098)	(0.221)
Depth	0.135*	0.116*	0.110*	0.144*
	(0.009)	(0.009)	(0.012)	(0.013)
Score	−4.580*	−4.704*	−7.097*	−8.221*
	(0.042)	(0.044)	(0.147)	(0.255)
Observations	108,656	108,656	108,656	108,656
Log Likelihood	−9,110.452	−8,571.424	−792.365	−390.050
Akaike Inf. Crit.	18,230.900	17,152.850	1,594.730	790.100

Note: *p<0.05. Models are binomial logits fit with penalized maximum-likelihood. All continuous variables are unit normalized.

References

Abbott, Rob, Brian Ecker, et al. (May 2016). "Internet Argument Corpus 2.0: An SQL Schema for Dialogic Social Media and the Corpora to Go With It". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. Portorož, Slovenia, pp. 4445–4452.

Abbott, Rob, Marilyn Walker, et al. (June 2011). "How Can You Say Such Things?!? Recognizing Disagreement in Informal Political Argument". In: *Proceedings of the Workshop on Language in Social Media*. Portland, OR: Association for Computational Linguistics, pp. 2–11.

Chakrabarty, Tuhin et al. (Nov. 2019). "AMPERSAND: Argument Mining for PERSuAsive oNline Discussions". In: *2019 EMNLP: Conference on Empirical Methods in Natural Language Processing*. arXiv:2004.14677. Hong Kong: Association for Computational Linguistics. arXiv: 2004.14677 [cs].

Devlin, Jacob et al. (May 2019). "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding". Pre-Print. Google AI Language. arXiv: 1810.04805.

Galitsky, Boris, Dmitry Ilvovsky, and Dina Pisarevskaya (Mar. 2018). "Argumentation in Text: Discourse Structure Matters". In: *19th International Conference on Computational Linguistics and Intelligent Text Processing*. Hanoi.

Gretz, Shai et al. (Apr. 2020). "A Large-Scale Dataset for Argument Quality Ranking: Construction and Analysis". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.5, pp. 7805–7813.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart (2021). "Machine Learning for Social Science: An Agnostic Approach". In: *Annual Review of Political Science* 24, pp. 395–419.

Grimmer, Justin and Brandon M. Stewart (2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". In: *Political Analysis* 21.3, pp. 267–297.

Hartmann, Mareike et al. (June 2019). "Issue Framing in Online Discussion Fora". In: *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. arXiv:1904.03969. Minneapolis, MN: Association for Computational Linguistics. arXiv: 1904.03969 [cs].

Huning, Hendrik, Lydia Mechtenberg, and Stephanie W. Wang (Mar. 2021). "Detecting Argumentative Discourse in Online Chat Experiments". Working Paper. Hamburg, Germany.

Kaufman, Aaron Russell, Peter Kraft, and Maya Sen (July 2019). "Improving Supreme Court Forecasting Using Boosted Decision Trees". In: *Political Analysis* 27.3, pp. 381–387.

Kovaleva, Olga et al. (Nov. 2019). "Revealing the Dark Secrets of BERT". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Hong Kong: Association for Computational Linguistics, pp. 4364–4373.

Landis, J. Richard and Gary G. Koch (Mar. 1977). "The Measurement of Observer Agreement for Categorical Data". In: *Biometrics* 33.1, p. 159.

Loshchilov, Ilya and Frank Hutter (May 2019). "Decoupled Weight Decay Regularization". In: *International Conference on Learning Representations*. New Orleans.

Lukin, Stephanie M. et al. (Aug. 2017). "Argument Strength Is in the Eye of the Beholder: Audience Effects in Persuasion". Pre-Print. University of California, Santa Cruz. arXiv: 1708.09085.

Mikolov, Tomas et al. (2013). "Distributed Representations of Words and Phrases and Their Compositionality". In: *Advances in Neural Information Processing Systems*, pp. 3111–3119. arXiv: 1310.4546.

Misra, Amita, Brian Ecker, and Marilyn A. Walker (Sept. 2016). "Measuring the Similarity of Sentential Arguments in Dialog". In: *Proceedings of the SIGDIAL 2016 Conference*. Los Angeles: Association for Computational Linguistics, pp. 276–287. arXiv: 1709.01887 [cs].

Misra, Amita and Marilyn Walker (Aug. 2013). "Topic Independent Identification of Agreement and Disagreement in Social Media Dialogue". In: *Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Metz, France: Association for Computational Linguistics, pp. 41–50.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn (2008). "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict". In: *Political Analysis* 16.4, pp. 372–403.

Montgomery, Jacob M. and Santiago Olivella (July 2018). "Tree-Based Models for Political Science Data". In: *American Journal of Political Science* 62.3, pp. 729–744.

Oraby, Shereen, Vrindavan Harrison, et al. (Sept. 2016). "Creating and Characterizing a Diverse Corpus of Sarcasm in Dialogue". In: *Proceedings of the SIGDIAL 2016 Conference*. arXiv:1709.05404. Los Angeles: Association for Computational Linguistics. arXiv: 1709.05404 [cs].

Oraby, Shereen, Lena Reed, et al. (May 2015). "And That's A Fact: Distinguishing Factual and Emotional Argumentation in Online Dialogue". In: *Proceedings of the 2nd Workshop on Argumentation Mining*. Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies. Denver.

Quinn, Kevin M. et al. (Jan. 2010). "How to Analyze Political Attention with Minimal Assumptions and Costs". In: *American Journal of Political Science* 54.1, pp. 209–228.

Radford, Alec et al. (2018). "Improving Language Understanding by Generative Pre-Training". Pre-Print. OpenAI.

Rodriguez, Pedro and Arthur Spirling (Jan. 2022). "Word Embeddings: What Works, What Doesn't, and How to Tell the Difference for Applied Research". In: *The Journal of Politics* 84.1, pp. 101–115.

Snow, Rion et al. (Oct. 2008). "Cheap and Fast - But Is It Good? Evaluating Non-Expert Annotations for Natural Language Tasks". In: *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, HI: Association for Computational Linguistics, pp. 254–263.

Stegmueller, Daniel (Aut. 2011). "Apples and Oranges? The Problem of Equivalence in Comparative Research". In: *Political Analysis* 19.4, pp. 471–487.

Turc, Iulia et al. (Sept. 2019). "Well-Read Students Learn Better: On the Importance of Pre-Training Compact Models". Pre-Print. Google Research. arXiv: 1908.08962.

Vaswani, Ashish et al. (Dec. 2017). "Attention Is All You Need". In: *31st Conference on Neural Information Processing Systems*. Long Beach, CA.

Walker, Marilyn A. et al. (May 2012). "A Corpus for Research on Deliberation and Debate". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation*. Istanbul, pp. 812–817.

Wang, Lu and Claire Cardie (June 2014). "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*. Baltimore: Association for Computational Linguistics, pp. 97–106.

Williams, Christopher K. I. (Apr. 2021). "The Effect of Class Imbalance on Precision-Recall Curves". In: *Neural Computation* 33.4, pp. 853–857.

Zhang, Gechuan, David Lillis, and Paul Nulty (Dec. 2021). "Can Domain Pre-Training Help Interdisciplinary Researchers from Data Annotation Poverty? A Case Study of Legal Argument Mining with BERT-Based Transformers". In: *Proceedings of the Workshop on Natural Language Processing for Digital Humanities*. Association for Computational Linguistics, pp. 121–130.

Zhu, Yukun et al. (Dec. 2015). "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books". In: *2015 IEEE International Conference on Computer Vision*. Santiago, Chile: Institute of Electrical and Electronics Engineers, pp. 19–27.