Political Discussion at Scale: Argument Mining with Deep Learning

Isaac D. Mehlhaff*

Abstract

Social scientists are frequently interested in how individuals discuss controversial issues, but annotating text or speech data on a large scale is expensive and time-consuming. Argument mining, a subfield of natural language processing, uses computational models to extract argumentative structure and reasoning from text. I introduce the field of argument mining into political science and show how a deep learning approach, in particular, holds promise for understanding argumentation in real-world political discussion, even on challenging tasks like detecting argument quality and distinguishing between emotion- and fact-based arguments. Across nine tasks, deep learning models afford substantial improvements over more common feature extraction strategies, and they achieve state-of-the-art results on three tasks with previously set metrics. The strong results attained by the classifiers developed in this paper suggest that argument mining offers significant potential for facilitating the study of persuasion and interpersonal discussion.

*The University of North Carolina at Chapel Hill; mehlhaff@live.unc.edu; word count: 9,705.

1 Introduction

Persuasion is a central feature of politics, and political language is replete with argumentative tactics, persuasive appeals, and diverse lines of reasoning (Druckman 2022). Grounded in a speaker-message-receiver-channel framework (Lau 2020), scholars of political behavior have made valuable advances in the study of how speaker characteristics affect message uptake (Carlson 2019), how a receiver's existing knowledge and opinions shape their acceptance of persuasive messages (Chong and Druckman 2007), and how different communication venues mold attitudes and behaviors (Druckman, Levendusky, and McLain 2018). The "message" part of this framework, however, is less well-developed. Outside the extensive literature on framing effects (e.g. Nelson, Oxley, and Clawson 1997), we know relatively little about what, exactly, makes a message persuasive or an argument high-quality. Identifying high-quality arguments generally involves asking survey respondents or experiment subjects whether they perceive an argument as strong or weak (Eagly and Chaiken 1993), resulting in "an inadequate tautology" that limits scholars' and practitioners' ability to understand and influence public opinion (Druckman 2022, p. 73).

My central objective in this paper is to advance deep argument mining as a methodological approach to the study of persuasion in real-world political discussion. Argument mining—the automated extraction of argumentative structure and reasoning from text—encompasses a rapidly expanding set of tasks in natural language processing (NLP). Fields such as computational linguistics and legal studies have been pushing the boundaries of argument mining for over a decade (Habernal and Gurevych 2016; Moens et al. 2007). The social sciences, however, have been slow to adopt these methods despite clear relevance to a wide range of topics like interpersonal discussion, social media, legislative speeches, and other political phenomena (cf. Beauchamp 2012; Joseph et al. 2021).

To enhance the value of argument mining to social scientists, I make two main contributions. First, the range of tasks for which argument mining researchers have achieved strong accuracy metrics is, at present, relatively limited. Scholars have primarily placed emphasis on identifying

syntactical structure (Feng and Hirst 2011), parsing argument components (Lawrence and Reed 2017), and stance classification (Walker, Anand, Abbott, et al. 2012). Yet it is more holistic properties of argumentation that hold the most value for social scientists. For example, recent substantive work has concentrated on identifying disagreement in political discussions (Carlson and Settle 2016), measuring online political hostility (Bor and Petersen 2022), and understanding the strategic use of facts in elite rhetoric about international affairs (Demasi 2019). Classifying or measuring these types of qualitative properties are difficult tasks, and they have garnered more limited work in NLP. I explicitly turn my attention to these types of uses, focusing on nine such classification tasks.

Second, common feature extraction methods in argument mining (e.g. n-grams, vocabulary complexity, or part-of-speech tags) are not ideal for social scientific uses. Many such methods impose substantial researcher degrees of freedom, are non-exhaustive, and are often endogenous to relevant covariates like the author or speaker's language proficiency, level of education, or medium of communication. Further, the types of data many social scientists use (e.g. social media posts or open-ended survey responses) are often brief and do not conform to standard grammatical conventions, complicating the use of structural or lexical features. Instead, I argue that a deep learning approach can bypass many of these concerns.

In particular, I test six classifier architectures with features extracted from pre-trained bidirectional transformers (Devlin et al. 2019). I train and test these classifiers on two corpora of online discussion and debate forums: The Internet Argument Corpus contributes up to 28,000 utterances with gold-standard annotations on a variety of argumentative characteristics such as whether an utterance expresses disagreement, uses emotion- or fact-based arguments, or is addressed at an individual or a broader audience (Abbott, Ecker, et al. 2016; Walker, Anand, Fox Tree, et al. 2012), and the Argument Extraction Corpus contributes just under 4,000 utterances with annotations of argument quality (Swanson, Ecker, and Walker 2015).

As usual, results suggest there is no free lunch: Different architectures perform well on different tasks, though fine-tuned transformer neural networks and random forests with extreme gradient

boosting tend to perform best across all tasks. Deep learning architectures achieve F1 scores between 0.59 and 0.828. These scores represent improvements of 5.1 to 233.3 percent over a lexical baseline and 12.2 to 60 percent over previous state-of-the-art metrics from lexical models. Argument mining on social scientific data is a challenging task, but results suggest it can be a powerful tool in the social scientific arsenal.

2 Deep Argument Mining for the Social Sciences

Argumentation comprises a significant portion of humans' cognitive processing. Constructing, delivering, receiving, and evaluating arguments enables humans to clarify and express their opinions, influence the opinions of others, and solve novel problems (Mercier and Sperber 2017). Argument mining is a particular approach to NLP that aims to systematize our identification and understanding of argumentation in text. In this sense, argument mining is distinct from other, more popular NLP tasks such as sentiment analysis (Liu 2020), which detects emotion and mood; topic modeling (Roberts et al. 2014), which classifies texts based on subject matter; or semantic similarity modeling (Callaghan, Karch, and Kroeger 2020), which evaluates the degree to which different texts resemble each other. Theoretically grounded in philosophy and linguistics (van Eemeren and Grootendorst 2003; Walton 1998), argument mining has been primarily concerned with structural tasks. For example, early work in argument mining sought to detect components of argumentation (e.g. premises and conclusion) within a text as well as the connective tissue (e.g. evidence) that allowed components to interact (Palau and Moens 2009; Stab and Gurevych 2014a). More recently, scholars and practitioners have expanded the playing field to include tasks such as stance classification—determining an individual's opinion on an issue (Joseph et al. 2021)—and discussion-based argument detection, which extends earlier structural approaches to account for the reciprocal nature of argumentation (Chakrabarty et al. 2019).

2.1 Two Problems

These structural tasks are important advancements, but they limit the usefulness of argument mining as a social scientific method. Save for some applications in legal studies (e.g. Moens et al. 2007), social scientists are not often interested in the precise syntactical design of texts or speeches.[1] More frequently, they are concerned with understanding more holistic properties of texts, such as whether an utterance is directed at an individual or a broader audience (Westwood 2015) or the degree to which a conversation lives up to deliberative ideals (Bächtiger and Hangartner 2010). Tasks designed to uncover such properties have garnered increased attention in recent years. Some research teams have pursued the detection of disagreement in political argumentation (Abbott, Walker, et al. 2011; Misra and Walker 2013), while others have devoted resources to scoring texts on diverse properties such as argument quality (Gretz et al. 2020; Ng et al. 2020), tolerance (Mukherjee et al. 2013), and specificity (Ko, Durrett, and Li 2019). Despite the gradual turn to holistic classification and regression tasks, these types of approaches still represent a relatively small segment of the argument mining literature.

Common feature extraction methods also limit the use of argument mining in social scientific resesarch. Huning, Mechtenberg, and Wang (2021) identify and test five types of features common in NLP generally and argument mining in particular:[2] *structural* features such as the number of question marks, exclamation points, or other punctuation; *lexical* features such as which words appear in the document (n-grams); *statistical* features such as the length of the document, lengths of words within the document, or the number of syllables per word; *syntactical* features such as part-of-speech tags; and *morphological* features such as tense, verb form, or verb reflexivity. For the types of tasks and corpora common in argument mining, these features are likely appropriate; training and target corpora are often comprised of highly structured documents like legal decisions (Walker, Pillaipakkamnatt, et al. 2019) or carefully written and edited documents like Wikipedia pages (Aker et al. 2017).

[1] Wilkerson and Casas (2017) provide an overview of text analysis in political science.

[2] See also Aker et al. (2017), Habernal and Gurevych (2017), and Stab and Gurevych (2014b).

The types of data gathered by behavioral scientists, however, are often neither structured nor carefully written. Consider the type of text that might be hastily entered into survey free-response boxes, squished into a 280-character tweet, or typed in a text message or online chat. These texts are likely to contain spelling errors, imperfect grammar, emojis, and other irregularities. For example, Rosenthal and McKeown (2015) give this example of an utterance they collected from an online debate forum: "KILLING A INNOCENT BABY ISN'T GONNA JUST GO AWAY YOU WILL HAVE TO LIVE WITH THE GUILT FOREVER!!!!!!!" This is clearly not representative of most utterances one might gather, but it does demonstrate several issues that are both common to behavioral data and seem likely to undermine the typical feature extraction strategies listed above; the writer uses "a innocent" instead of "an innocent," collapses "going to" into the colloquial "gonna," fails to break up the two sentences with a period, and uses seven exclamation points. A model trained on structural features, therefore, may identify this utterance as conveying a strong argument when, in fact, it is merely conveying a strong emotional reaction.

2.2 Two Solutions

In this paper, I push past these limitations and demonstrate the promise of argument mining in the social sciences. First, I move beyond the focus on structural tasks that has previously dominated argument mining and instead focus on nine classification tasks resembling the types of holistic properties in which behavioral scholars are often interested. I describe these tasks in detail in a subsequent section. Second, I forego the types of lexical features outlined above in favor of a deep learning approach. The vast majority of work in argument mining employs "shallow" learning, the more traditional approach to machine learning wherein features are extracted and chosen manually and fed directly to a model that learns how each feature contributes to an output. For example, a typical approach to argument mining with shallow learning might be to count the number of times each word appears in a document, how many question marks the document uses, and the average number of syllables in the document's words, and use this information in a logistic regression model to predict a class label. This approach has several benefits: It is less computa-

tionally expensive, it is conceptually simple, and it retains easily interpretable features. However, shallow learning techniques struggle to process raw data, they often require careful manipulation and deep domain expertise to effectively extract features, and their performance metrics are often overshadowed by those achieved through deep learning (LeCun, Bengio, and Hinton 2015).

Developed in earnest beginning in the mid-2000s (Schmidhuber 2015), deep learning approaches enable a machine to take raw data as inputs (e.g. the actual text of a tweet rather than some derived characteristics of it) and learn to represent that data in a way that allows it to conduct classification or regression (Bengio, Courville, and Vincent 2014). Deep learning is based on a simple premise: Humans perceive the world in multiple levels of abstraction, and this hierarchy of abstraction can be represented with an arbitrary number of layers in an artificial neural network (Bengio 2009; Deng and Yu 2013). As you are reading this very paper, your brain is forming a hierarchy of judgments about it. At the most abstract level, you would likely classify this as a paper about understanding argumentation in political texts. The present section, however, is describing argument mining. Drilling down even further, this paragraph explains what deep learning is and does. You might even scrutinize my use of individual words to convey my thoughts, such as "abstraction." Each word in this and the surrounding sentences contributes to your judgment about this paragraph, your judgment about this paragraph helps you figure out the purpose of this section, and knowing the purpose of this section helps you understand the topic of the entire paper.

Deep learning architectures follow a similar process. While these models are still mostly black boxes and the precise mechanics behind why they work are generally unknown (e.g. Kovaleva et al. 2019), recent studies have uncovered insights that begin to explain how they are able to achieve accuracy levels approaching or surpassing those of humans. For example, large neural networks use the meanings of smaller units of speech (such as words) to gradually infer the meanings of larger units (such as sentences or paragraphs) (Li et al. 2016), and long short-term memory networks distinguish between close context (200 or so other words immediately surrounding each word) as a source of fine-grained meaning and distant context (preceding paragraphs) as a source of high-level meaning for the document as a whole (Khandelwal et al. 2018).

This approach seems to offer clear advantages over more traditional methods of shallow learning. First, it more closely mimics the cognitive processes used to comprehend argumentation or discussion. Representing each word as an isolated entity—as do most feature extraction strategies in argument mining—throws away a great deal of meaning, such as the order in which words commonly appear in a document. Moreover, the persuasive effect of a post, document, or conversation comes not from the individual words the speaker uses, but how those words coalesce to create a cohesive thought. Second, because it takes raw text as an input, deep learning obviates the need for manual feature extraction, as it constructs and learns to represent features automatically as part of the learning process (Grimmer, Roberts, and Stewart 2021). Deep learning architectures nevertheless impose two main drawbacks. First, they translate raw inputs into numeric representations that are effectively meaningless, making substantive interpretation of features nearly impossible. However, for applications—like argument mining—in which the objective is high-accuracy classification that can be used to answer other research questions, scholars are typically less concerned with which features, exactly, allow a model to perform at a high level. Second, deep learning architectures typically require large amounts of training data, but transfer learning and data augmentation—both of which I employ here—offer opportunities to circumvent the problem of scarce data.

3 Data and Methods

3.1 Data Sources and Tasks

I rely on two corpora to train the classifiers in this paper. For eight of the nine tasks, I use the Internet Argument Corpus (IAC), a collection of posts extracted from several online debate and discussion forums (Abbott, Ecker, et al. 2016; Walker, Anand, Fox Tree, et al. 2012). Given the imperfect nature of behavioral data as described above, it may be particularly important in this case that I use a corpus gathered from sources similar to those on which the model is likely to be used for inference. 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 a variety of characteristics: whether the post expresses disagreement, uses an emotion- or fact-based argument, attacks or is respectful toward an interlocutor, uses a nasty or nice tone, directs its argument toward a specific individual or a broader audience, employs a defeater (evidence to directly contradict the entirety of someone's argument) or undercutter (an argument targeted at a specific piece of evidence), merely attacks another argument or provides an argument of its own, and questions an interlocutor or asserts an original idea. 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.[3] 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 from the Argument Extraction Corpus, which is largely composed of posts from the IAC (Swanson, Ecker, and Walker 2015). This corpus focuses in particular on the topics of same-sex marriage, gun control, the death penalty, and evolution. Similar to the IAC, each post is annotated by seven human coders on the continuous scale $[0,1]$ according to how difficult or easy the argument is to interpret. Posts making no argument at all are given a score of zero. All coders' scores are then averaged to get the final real-valued argument quality score. Across all topics, the authors report an average intraclass correlation coefficient of 0.415, a value indicating moderate agreement (Cicchetti 1994). Additional information on data manipulation is included in the following subsection.

A wide array of studies have used the IAC and Argument Extraction Corpus to construct unique tasks (Galitsky, Ilvovsky, and Pisarevskaya 2018; Hartmann et al. 2019; Misra, Ecker, and Walker

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

2016) and train models (Lukin, Anand, et al. 2017; Misra and Walker 2013; Oraby, Harrison, et al. 2016). Three of the tasks I pursue here have 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. On the emotional or factual argument classification task, Oraby, Reed, et al. (2015) achieve an F1 score of 0.462. Finally, on the nasty or nice tone classification task, Lukin and Walker (2013) achieve an F1 score of 0.69. All past state-of-the-art metrics were set using some combination of the structural or lexical features identified in the previous section.

3.2 Data Manipulation

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 1 provides descriptive statistics of the data used for all nine tasks, with the total N and class balance representing the final, dichotomized corpora. Each of the eight IAC tasks is described with two labels; the first of these labels represents the positive class (i.e. coded as one) and the second label represents the negative class (coded as zero). 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 1: Descriptive Statistics of Document Annotations

Task	N	Range	Mean	SD	Class Balance
Disagree / Agree	28,171	[-5, 5]	-0.916	1.689	0.802 / 0.198
Emotion / Fact	23,057	[-5, 5]	0.065	1.536	0.468 / 0.532
Attacking / Respectful	24,819	[-5, 5]	0.628	1.485	0.242 / 0.758
Nasty / Nice	26,612	[-5, 5]	0.895	1.439	0.159 / 0.841
Individual / Audience	5,997	[-5, 5]	-1.271	2.09	0.816 / 0.184
Defeater / Undercutter	5,603	[-5, 5]	-0.671	2.245	0.673 / 0.327
Counterargue / Attack	5,831	[-5, 5]	-0.479	2.293	0.625 / 0.375
Question / Assert	6,216	[-5, 5]	0.717	2.368	0.319 / 0.681
Argument Quality	3,895	[0, 1]	0.54	0.277	0.595 / 0.405

3.2.1 Data Augmentation

From here, there is one final step prior to model training. Just like humans, machines learn best when they have lots of repetition and access to diverse examples. I therefore conduct data augmentation on the training set before performing feature extraction. Data augmentation is frequently used in computer vision (Torres and Cantú 2022) and NLP (Shorten, Khoshgoftaar, and Furht 2021) to help guard against model overfitting that might be brought about by poor data availability and variation. This procedure involves taking the observed documents in the training set and applying various transformations to produce new documents that retain the same basic meaning as the originals but with different words, phrases, or syntax. I perform four different types of augmentation. In back-translation, I translate the text into a different language, then translate

it back to the original language. I choose German as the translation language for its high lexical similarity to English. In contextual word embedding, I randomly choose thirty percent of tokens, feed the surrounding words to a pre-trained bidirectional transformer model trained on next-word prediction (Devlin et al. 2019), and substitute the predicted word in for the original. In synonym augmentation, I randomly choose thirty percent of tokens and substitute the most similar word from the WordNet lexical database (Fellbaum 1998). Finally, in random cropping, I randomly delete thirty percent of tokens. To illustrate how augmentation can diversify characteristics of text, Table 2 displays two examples of documents before and after augmentation. Final training sets range in size from 15,580 documents for the argument quality task to 112,684 documents for the disagreement identification task.

Table 2: Examples of Documents Before and After Augmentation

	Evolution	**Gun Control**
Original	Biological evolution does not involve the formation of life nor is it a rule that life can only come from life it is a tendency because it is vastly more efficient	Firearm purchases not requiring background checks allow felons to purchase weapons from law abiding citizens
Back-Translation	Biological evolution does not involve the formation of life, nor is it a rule that life can only arise from life, it is a tendency because it is much more efficient.	Gun purchases without background checks allow criminals to buy guns from law-abiding citizens
Contextual Embeddings	biological evolution does not exclusively involve just the formation of life nor is it a firm rule that life can eventually only come from life as it is finding a new tendency because therefore it is already vastly becoming more efficient	special firearm license purchases not requiring uniform background checks allow certified felons to purchase certain weapons from a law abiding citizens
Synonyms	Biological phylogenesis act non demand the organization of life nor is it a normal that life can only come from life it is a tendency because information technology is vastly more effective	Firearm purchases not requiring screen background check let outlaw to buy weapons from law abiding citizens
Random Cropping	Biological evolution does not involve the formation of life nor is it a rule that life can is vastly more efficient	Firearm allow felons to purchase weapons from law abiding citizens

3.3 Feature Extraction

3.3.1 Lexical Feature Extraction

In order to assess whether and to what degree deep learning offers improvement over more common structural approaches, I need to compare its performance to that of a lexical baseline. Features for this baseline take the form of 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.

3.3.2 Feature Extraction with BERT

In contrast to this "bag of words" approach, 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.

[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, Anand, Fox Tree, et al. 2012). These include words like "because," "then," and "so."

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.

To understand why this feature is crucial in NLP, consider how each of the three approaches (non-contextual, unidirectionality, and bidirectionality) deal with polysemy—the phenomenon in which a word or phrase can hold multiple potential meanings. Figure 1 uses two sets of exemplar sentences to demonstrate the importance of bidirectional context. The two sentences on the left lack context: "I went to the bank" and "I crashed into the bank." The word "bank," in these cases, is polysemous; it could refer to a financial institution, the bank of a river, a snow bank, or even other distinct objects.

No Context	Left Context	Right Context
I went to the **bank**.	I went to the *Context*(river) **bank**.	I went to the **bank** *Context*(to make a deposit).
I crashed into the **bank**.	I crashed into the *Context*(snow) **bank**.	I crashed into the **bank** *Context*(of snow).

Figure 1: Exemplar Sentences Demonstrating Importance of Bidirectional Context

One way humans differentiate between these meanings is by paying attention to context *before* the polysemous word, or "left context." For example, the speaker might specify that they went to the river bank or that they crashed into the snow bank, which would clear up any confusion about

which bank they went to or whether they recently drove their car through the wall of a financial institution. Word embedding models, while taking these nearby words into account when *learning* what words "mean," would not use this context when *interpreting* language, as they would assign a single value to the token representing "bank," regardless of the type of bank to which the speaker was referring. Both uni- and bidirectional models, however, could use this left context to infer different meanings for the work "bank."

Another way humans infer meaning, however, is by paying attention to context *after* the polysemous word, or "right context." For example, the speaker could say "I went to the bank to make a deposit," from which we would infer that they went to a financial institution. Or they could again refer to a snow bank, but by saying "bank of snow" instead. Only deeply bidirectional language models are capable of learning word meaning from right context; word embeddings and unidirectional models would fail to distinguish between these two meanings of "bank," and unidirectional models may even assign different meaning to the two sentences about snow banks. If a machine is to truly "know a word by the company it keeps" (Firth 1957, p. 11), it must rely on bidirectional language models.

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.

3.4 Classifier Architectures

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 and compare them to two baseline classifiers. 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.

3.4.1 Deep Learning Classifiers

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.[5] I use a grid search to tune three additional hyperparameters in the random forest 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 of which are allowed to take values in $\{0.2, 0.4, 0.6, 0.8\}$.

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.[6] 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). I also use a grid search to tune five hyperparameters in these 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

[5] 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.

[6] 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.

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 $\{32, 64, 128\}$. The Supplementary Information displays results of hyperparameter tuning in both the random forests with extreme gradient boosting and the fine-tuned BERT neural networks.

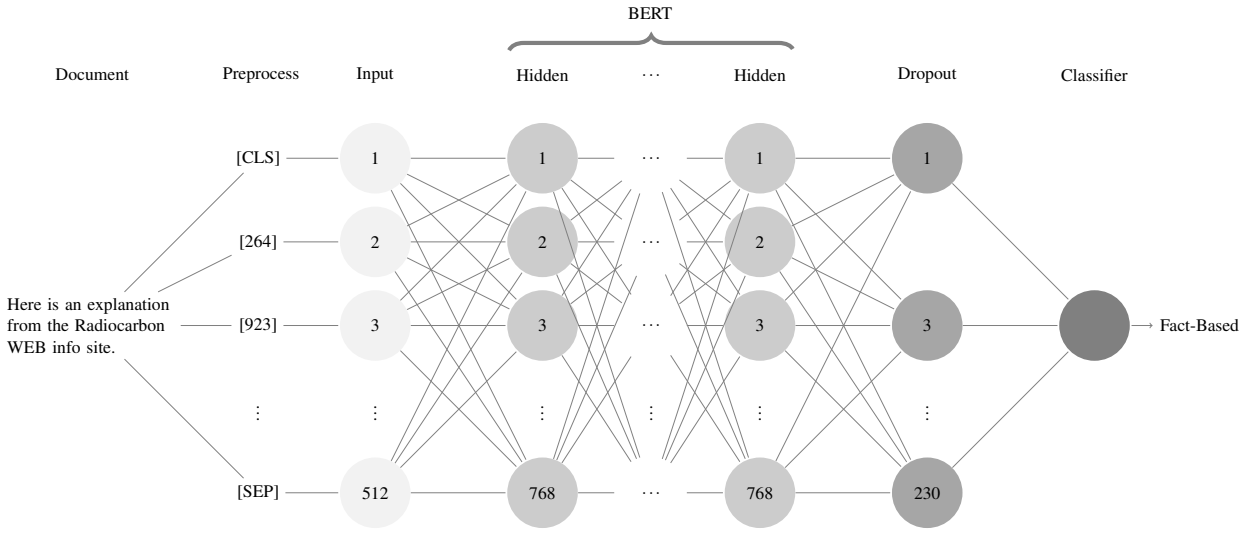

Figure 2: Visual Representation of Deep Learning Pipeline. Hidden layers are transformer blocks. Fine-tuned BERT models have two hidden layers with 128 neurons and two attention heads. Other classifiers using BERT feature extraction have twelve hidden layers with 768 neurons and twelve attention heads. Dropout and output layers are fully connected in fine-tuned BERT model.

The full deep learning pipeline is depicted in Figure 2. This diagram shows how an exemplar document is preprocessed into the format required by BERT. Those tokens are then taken directly by the input layer, which feeds forward into the hidden layers of BERT. A fully-connected dropout layer conducts regularization, and finally a classifier takes the output of either the dropout layer (in the fine-tuned neural network) or the final hidden layer (for all other classifier architectures) and makes a determination about which label should be assigned to the document.

3.4.2 Baseline Classifiers

To benchmark the performance of these classifiers and evaluate the potential of deep learning in argument mining, I need baseline performance metrics to which they can be compared. 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, for which feature extraction was outlined in the previous section, uses a support vector machine with stochastic gradient descent and a logistic loss function.

4 Results

4.1 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 1 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 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 3 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.[7] On most tasks, the deep learning models outperform both baselines, achieving both higher precision and higher recall

[7] Full precision-recall curves are presented in the Supplementary Information.

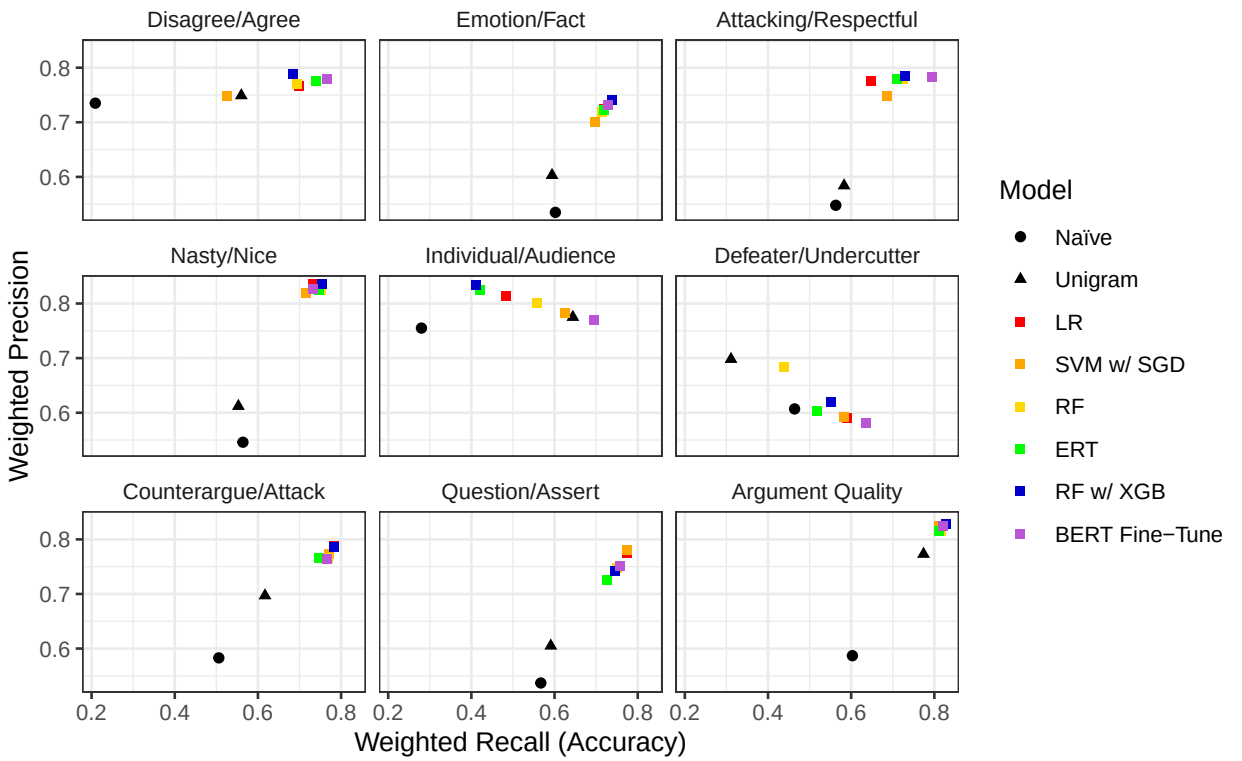

Figure 3: Weighted Precision and Recall. Weighted recall is mathematically equivalent to accuracy. Black points represent baselines, colored points represent deep learning models.

(and, by extension, higher accuracy). Their advantage is less obvious on the individual/audience and defeater/undercutter tasks, on which there is a clear tradeoff between precision and recall. Even so, deep learning is competitive with the lexical baseline on the individual/audience task and looks to have a slight advantage in the defeater/undercutter task. Across all tasks, the random forest with extreme gradient boosting and the fine-tuned BERT neural network generally perform best.

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

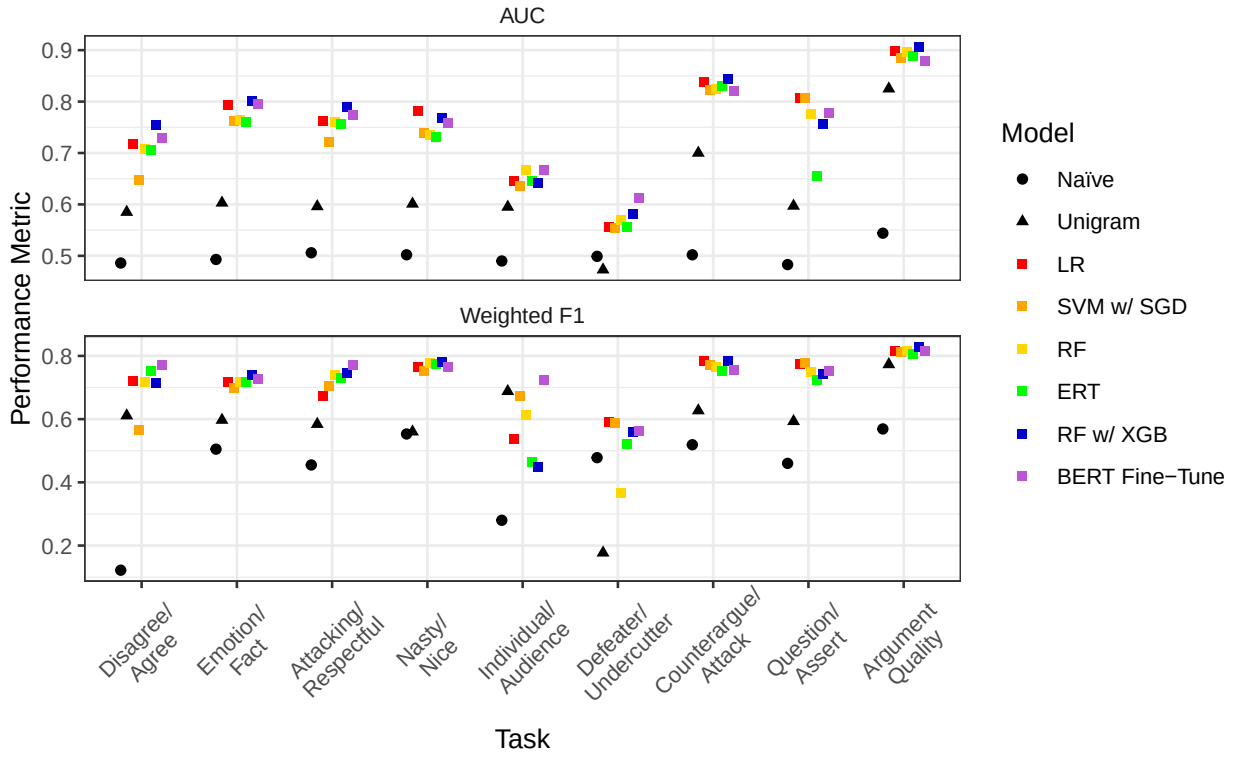

Figure 4: Area Under the ROC Curve and Weighted F1 Score. Black points represent baselines, colored points represent deep learning models.

and recall.[8] 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 4, with the AUC plotted in the top facet and the F1 score in the bottom facet.[9] Again, black points represent the two baselines, while colored points represent the deep learning models. Results largely mirror those in Figure 3, but the benefit of deep learning over a lexical model is much clearer here.

Three findings deserve emphasis. First, on every task, deep learning models achieve higher AUC metrics than either baseline. The same is largely true of F1 scores, though the lexical baseline performs remarkably well on the individual/audience task. Second, an even more important test is to compare the deep support vector machine to the lexical baseline, which also uses a support vector machine. Even here, the deep learning model bests the lexical model on every task when

[8]For the same reasons as above, I also present the weighted F1 score.
[9]Full ROC curves are presented in the Supplementary Information.

considering AUC metrics and on all but one task when looking at F1 scores. The advantage held by deep learning over the lexical baseline does not, therefore, appear to be an artifact of the particular classifier used to estimate the baseline. Finally, as usual, there is no free lunch; different classifier architectures are better-suited for different tasks. The fine-tuned neural network achieves the best F1 score on three tasks, the logistic regression and random forest with extreme gradient boosting on two tasks each, and the support vector machine on one task, though it is rare for one classifier to hold a substantial advantage over the others. These results suggest that researchers should test a wide variety of classifiers to identify the one that offers the best performance on each particular task.

4.2 Improvement Over Baselines and Previous State-of-the-Art Metrics

To more clearly summarize the potential for deep learning to uncover holistic properties of argumentative texts, Table 3 provides the F1 score achieved by the best-performing deep learning classifier, the absolute and relative gain over the naïve baseline, and the absolute and relative gain over the lexical baseline. The benefits from using deep learning are substantial, with most relative improvements over the lexical baseline between 5 and 39 percent. The F1 scores themselves are likewise appreciable, with most in the high 70 percent range. Although such metrics are not directly comparable across tasks, these numbers eclipse those of other recently published tasks in the social sciences (e.g. Brady et al. 2021).

State-of-the-art metrics have been previously set on three tasks: disagreement identification, emotional or factual argument classification, and nasty or nice tone classification. All of these metrics were achieved with various forms of lexical or structural feature extraction. Table 4 displays these state-of-the-art metrics and compares them to the performance metrics achieved here.[10] The deep learning models presented in this paper represent substantial advancements, with absolute gains ranging from 8.3 percent on disagreement accuracy to 27.7 percent on emotion/fact F1.

[10]Wang and Cardie (2014) and Oraby, Reed, et al. (2015) report results for each class separately. These per-class metrics have been combined to facilitate comparison to the metrics calculated for the models presented in this paper.

Table 3: Absolute and Relative Improvement Over Baselines

Task	Best F1	Improvement over Naïve Baseline		Improvement over Lexical Baseline	
		Absolute	Relative	Absolute	Relative
Disagree / Agree	77.1%	64.9%	531.967%	16%	26.187%
Emotion / Fact	73.9%	23.4%	46.336%	14.2%	23.786%
Attacking / Respectful	78.7%	33.2%	72.967%	20.3%	34.76%
Nasty / Nice	78.1%	22.8%	41.23%	22.1%	39.464%
Individual / Audience	72.3%	44.3%	158.214%	3.5%	5.087%
Defeater / Undercutter	59%	11.2%	23.431%	41.3%	233.333%
Counterargue / Attack	78.5%	26.6%	51.252%	15.8%	25.199%
Question / Assert	77.6%	31.6%	68.696%	18.3%	30.86%
Argument Quality	82.8%	25.9%	45.518%	5.5%	7.115%

These improvements correspond to relative gains of 12.17 percent on disagreement accuracy and 59.96 percent on emotion/fact F1.

Table 4: Absolute and Relative Improvement Over Previous State-of-the-Art Metrics

Task	Citation	Metric	Previous	New	Absolute Gain	Relative Gain
Disagree / Agree	Wang and Cardie (2014)	F1	63.57%	77.1%	13.53%	21.28%
Disagree / Agree	Abbott, Walker, et al. (2011)	Acc.	68.2%	76.5%	8.3%	12.17%
Emotion / Fact	Oraby, Reed, et al. (2015)	F1	46.2%	73.9%	27.7%	59.96%
Nasty / Nice	Lukin and Walker (2013)	F1	69%	78.1%	9.1%	13.19%

4.3 Example Classifications

Finally, I provide a series of exemplary documents from the test set and show in Table 4.3 how the models classify the documents on three tasks. For purposes of demonstration, examples were selected for variation on classes. All classifications were produced using the random forest classifier with extreme gradient boosting. Seeing these example classifications along with their companion documents highlights the difficulty of these tasks; the boundary between class labels is not always clear-cut. Even given a certain level of prediction error, however, the classifiers seem to return sensible results. For instance, the first document appears to be making an argument with appeals to emotion while the third document makes a factual argument and the fifth document

hardly makes an argument at all, merely attacking an interlocutor's argument with emotion-laden rhetorical questions. This small set of results also makes clear that the models do not view these categories as codetermined. For example, one can express agreement with an interlocutor but still provide an argument, as in the second document. Or one can state facts without providing an explicit argument, as in the fourth document.

Table 5: Predicted Classifications of Exemplar Documents from Test Set

Document	Disagree	Agree	Emotion	Fact	Counterargue	Attack
So why do the elderly marry? Why do those who know they can never conceive children marry? Marriage is not simply a way of providing a stable home life for children at all. Marriage today is for companionship and love.	✓		✓		✓	
Cost would certainly be an issue. If the adoptive parents were willing to take on that cost, that would be best. I wonder, though, how much demand there really is out there for adopted babies.		✓	✓		✓	
Another critical error. Evolution does not say life is chaotic. Mutations are random, not natural selection or evolution. It is a heavily guided process. We have already shown that the primary building blocks of life can be simulated and created. They can even explain how the human eye developed—a recent discovery.	✓			✓	✓	
I know what you mean. I also wonder how they think their children or grandchildren will be paying for their medical bills. Note, by way of comparison, this Pittsburgh local TV clip of the September March for Jobs. See how clearly and intelligently the protestors articulate their objectives.		✓		✓		✓
What about the pain and death inflicted upon the innocent women and girls forced to give birth? Have you no regard for them?	✓		✓			✓

5 Conclusion

The overarching takeaway from the models developed and evaluated in this paper is that argument mining can be a valuable tool in the social scientific toolkit, but the degree of that potential

may depend on the methods used to implement it. Argument mining on data relevant to social scientists is a challenging task—utterances are often brief and do not conform to standard language structures, and aspects of argumentation can be highly subjective. For these reasons, lexical models and other popular methods of feature extraction are likely not appropriate for use on these data. Deep learning may present a more reliable path forward for understanding political discussion and argumentation. As always, however, there is no universal classifier; different models perform well on different tasks, though fine-tuned BERT neural networks and random forests with extreme gradient boosting tend to give consistently high performance relative to other classifier architectures. Researchers applying deep learning methods to their own problems should test a wide variety of classifiers to find the one best-suited for their task.

Finally, it bears mentioning that the exercises in this paper largely function as a proof of concept. Though they are complex relative to many common methods of text analysis in political science, the deep learning models employed here are still fairly elementary, and the fine-tuned neural networks use only the small BERT model. But if these models can nevertheless achieve substantial gains over lexical models, the ceiling for deep argument mining seems to be quite high. Researchers working in similar topic areas should continue to push the envelope with more complex network architectures, larger model sizes, and higher-quality corpora to achieve even better results.

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.

Aker, Ahmet et al. (2017). "What Works and What Does Not: Classifier and Feature Analysis for Argument Mining". In: *Proceedings of the 4th Workshop on Argument Mining*. EMNLP 2017: Conference on Empirical Methods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, pp. 91–96.

Bächtiger, André and Dominik Hangartner (Oct. 2010). "When Deliberative Theory Meets Empirical Political Science: Theoretical and Methodological Challenges in Political Deliberation". In: *Political Studies* 58.4, pp. 609–629.

Beauchamp, Nick (Sept. 2012). "A Bottom-Up Approach to Linguistic Persuasion in Advertising: Predicting and Explaining the Effects of TV Ads Using Automated Text Analysis". Working Paper. Columbia University.

Bengio, Yoshua (2009). "Learning Deep Architectures for AI". In: *Foundations and Trends in Machine Learning* 2.1, pp. 1–127.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (Apr. 2014). *Representation Learning: A Review and New Perspectives*. arXiv: `1206.5538 [cs]`.

Bor, Alexander and Michael Bang Petersen (Feb. 2022). "The Psychology of Online Political Hostility: A Comprehensive, Cross-National Test of the Mismatch Hypothesis". In: *American Political Science Review* 116.1, pp. 1–18.

Brady, William J. et al. (Aug. 2021). "How Social Learning Amplifies Moral Outrage Expression in Online Social Networks". In: *Science Advances* 7 (eabe5641).

Callaghan, Timothy, Andrew Karch, and Mary Kroeger (July 2020). "Model State Legislation and Intergovernmental Tensions over the Affordable Care Act, Common Core, and the Second Amendment". In: *Publius: The Journal of Federalism* 50.3, pp. 518–539.

Carlson, Taylor N. (May 2019). "Through the Grapevine: Informational Consequences of Interpersonal Political Communication". In: *American Political Science Review* 113.2, pp. 325–339.

Carlson, Taylor N. and Jaime E. Settle (Dec. 2016). "Political Chameleons: An Exploration of Conformity in Political Discussions". In: *Political Behavior* 38.4, pp. 817–859.

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].

Chong, Dennis and James N. Druckman (June 2007). "Framing Theory". In: *Annual Review of Political Science* 10.1, pp. 103–126.

Cicchetti, Domenic V. (Dec. 1994). "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology". In: *Psychological Assessment* 6.4, pp. 284–290.

Demasi, Mirko A. (Feb. 2019). "Facts as Social Action in Political Debates about the European Union: Facts as Social Action". In: *Political Psychology* 40.1, pp. 3–20.

Deng, Li and Dong Yu (2013). "Deep Learning: Methods and Applications". In: *Foundations and Trends in Signal Processing* 7.3, pp. 197–387.

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

Druckman, James N. (2022). "A Framework for the Study of Persuasion". In: *Annual Review of Political Science* 25.

Druckman, James N., Matthew S. Levendusky, and Audrey McLain (Jan. 2018). "No Need to Watch: How the Effects of Partisan Media Can Spread via Interpersonal Discussions". In: *American Journal of Political Science* 62.1, pp. 99–112.

Eagly, Alice H. and Shelly Chaiken (1993). *The Psychology of Attitudes*. Fort Worth, TX: Harcourt Brace Jovanovich.

Fellbaum, Christiane, ed. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: The MIT Press.

Feng, Vanessa Wei and Graeme Hirst (June 2011). "Classifying Arguments by Scheme". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*. Portland, OR, pp. 987–996.

Firth, John Rupert (1957). *Studies in Linguistic Analysis*. Oxford, UK: Blackwell Publishers.

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.

Habernal, Ivan and Iryna Gurevych (Nov. 2016). "What Makes a Convincing Argument? Empirical Analysis and Detecting Attributes of Convincingness in Web Argumentation". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, TX: Association for Computational Linguistics, pp. 1214–1223.

— (Apr. 2017). "Argumentation Mining in User-Generated Web Discourse". In: *Computational Linguistics* 43.1, pp. 125–179.

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.

Joseph, Kenneth et al. (Nov. 2021). "(Mis)Alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys". In: *Empirical Methods in Natural Language Processing*. Punta Cana, Dominican Republic. arXiv: 2109.01762.

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.

Khandelwal, Urvashi et al. (July 2018). "Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Vol. 1. Melbourne, Australia: Association for Computational Linguistics, pp. 284–294.

Ko, Wei-Jen, Greg Durrett, and Junyi Jessy Li (Jan. 2019). "Domain Agnostic Real-Valued Specificity Prediction". In: *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*. Honolulu, HI: Association for the Advancement of Artificial Intelligence, pp. 6610–6617.

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.

Lau, Richard R. (2020). "Classic Models of Persuasion". In: *The Oxford Handbook of Electoral Persuasion*. Ed. by Elizabeth Suhay, Bernard Grofman, and Alexander H. Trechsel. New York: Oxford University Press, pp. 27–50.

Lawrence, John and Chris Reed (Sept. 2017). "Mining Argumentative Structure from Natural Language Text Using Automatically Generated Premise-Conclusion Topic Models". In: *Proceedings of the 4th Workshop on Argument Mining*. EMNLP 2017: Conference on Empirical Meth-

ods in Natural Language Processing. Copenhagen: Association for Computational Linguistics, pp. 39–48.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (May 2015). "Deep Learning". In: *Nature* 521.7553, pp. 436–444.

Li, Jiwei et al. (June 2016). "Visualizing and Understanding Neural Models in NLP". In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA: Association for Computational Linguistics, pp. 681–691.

Liu, Bing (2020). *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. 2nd ed. Cambridge, UK: Cambridge University Press.

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

Lukin, Stephanie and Marilyn Walker (June 2013). "Really? Well. Apparently Bootstrapping Improves the Performance of Sarcasm and Nastiness Classifiers for Online Dialogue". In: *Proceedings of the Workshop on Language in Social Media*. Atlanta: Association for Computational Linguistics, pp. 3–40.

Lukin, Stephanie M., Pranav Anand, 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.

Mercier, Hugo and Dan Sperber (2017). *The Enigma of Reason*. Cambridge, MA: Harvard University Press.

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.

Moens, Marie-Francine et al. (June 2007). "Automatic Detection of Arguments in Legal Texts". In: *Proceedings of the 11th International Conference on Artificial Intelligence and Law*. Stanford, CA: ACM Press, pp. 225–230.

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.

Mukherjee, Arjun et al. (Aug. 2013). "Public Dialogue: Analysis of Tolerance in Online Discussions". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Vol. 1. Sofia, Bulgaria, pp. 1680–1690.

Nelson, Thomas E., Zoe M. Oxley, and Rosalee A. Clawson (Sept. 1997). "Toward a Psychology of Framing Effects". In: *Political Behavior* 19.3, pp. 221–246.

Ng, Lily et al. (Dec. 2020). "Creating a Domain-Diverse Corpus for Theory-Based Argument Quality Assessment". In: *7th Workshop on Argument Mining*. 28th International Conference on Computational Linguistics. arXiv:2011.01589. Barcelona. arXiv: 2011.01589 [cs].

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 Argu-*

mentation Mining. Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies. Denver.

Palau, Raquel Mochales and Marie-Francine Moens (June 2009). "Argumentation Mining: The Detection, Classification, and Structure of Arguments in Text". In: *Proceedings of the 12th International Conference on Artificial Intelligence and Law*. Barcelona: ACM Press, pp. 98–107.

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.

Roberts, Margaret E. et al. (Oct. 2014). "Structural Topic Models for Open-Ended Survey Responses". In: *American Journal of Political Science* 58.4, pp. 1064–1082.

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.

Rosenthal, Sara and Kathy McKeown (2015). "I Couldn't Agree More: The Role of Conversational Structure in Agreement and Disagreement Detection in Online Discussions". In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague: Association for Computational Linguistics, pp. 168–177.

Schmidhuber, Jürgen (2015). "Deep Learning in Neural Networks: An Overview". In: *Neural Networks* 61, pp. 85–117.

Shorten, Connor, Taghi M. Khoshgoftaar, and Borko Furht (Dec. 2021). "Text Data Augmentation for Deep Learning". In: *Journal of Big Data* 8.1, p. 101.

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.

Stab, Christian and Iryna Gurevych (Aug. 2014a). "Annotating Argument Components and Relations in Persuasive Essays". In: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin, pp. 1501–1510.

— (2014b). "Identifying Argumentative Discourse Structures in Persuasive Essays". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha: Association for Computational Linguistics, pp. 46–56.

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

Swanson, Reid, Brian Ecker, and Marilyn Walker (Sept. 2015). "Argument Mining: Extracting Arguments from Online Dialogue". In: *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Prague: Association for Computational Linguistics, pp. 217–226.

Torres, Michelle and Francisco Cantú (Jan. 2022). "Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data". In: *Political Analysis* 30.1, pp. 113–131.

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.

Van Eemeren, Frans H. and Rob Grootendorst (2003). *A Systematic Theory of Argumentation. The Pragma-Dialectic Approach*. Cambridge, UK: Cambridge University Press.

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., Pranav Anand, Rob Abbott, et al. (June 2012). "Stance Classification Using Dialogic Properties of Persuasion". In: *2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal: Association for Computational Linguistics, pp. 592–596.

Walker, Marilyn A., Pranav Anand, Jean E. Fox Tree, 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.

Walker, Vern R., Krishnan Pillaipakkamnatt, et al. (June 2019). "Automatic Classification of Rhetorical Roles for Sentences: Comparing Rule-Based Scripts with Machine Learning". In: *Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Text*. Montréal.

Walton, Douglas N. (1998). *The New Dialectic: Conversational Contexts of Argument*. Toronto: University of Toronto Press.

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.

Westwood, Sean J. (Oct. 2015). "The Role of Persuasion in Deliberative Opinion Change". In: *Political Communication* 32.4, pp. 509–528.

Wilkerson, John and Andreu Casas (2017). "Large-Scale Computerized Text Analysis in Political Science: Opportunities and Challenges". In: *Annual Review of Political Science* 20, pp. 529–544.

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.