Using Cross-Encoders to Measure the Similarity of Short Texts in Political Science

Gechun Lin*

April 2024

Abstract

In many settings, scholars wish to estimate the similarity of political texts. However, the most commonly used methods in political science struggle to identify when two texts convey the same *meaning* as they rely too heavily on identifying words that appear in both documents. This limitation is especially salient when the underlying documents are short, an increasingly prevalent form of textual data in modern political research. Building on recent advances in computer science, I introduce to political science cross-encoders for precise estimates of *semantic* similarity in short texts. Scholars can use either off-the-shelf versions or build a customized model. I illustrate this approach in three examples applied to social messages generated in a telephone game, news headlines about US Supreme Court decisions, and Facebook posts from members of Congress. I show that cross-encoders, which utilize *pair*-level embeddings, offer superior performance across tasks relative to word-based and sentence-level embedding approaches.

*Graduate student in Political Science, Washington University in St. Louis (lingechun@wustl.edu).

I am deeply indebted to Jacob Montgomery, Christopher Lucas, Betsy Sinclair, and Ted Enamorado for their expertise and support on this project. Many thanks to Adeline Lo, Bruce Desmarais, Kevin Esterling, conference discussants of this paper, and reviewers for the AJPS who provided exceptionally helpful comments. I also thank my cohort and participants at the Washington University Political Data Science Lab who made many useful suggestions on much earlier versions.

Estimating the similarity of texts is an increasingly common task in political science research. For example, scholars of legislative studies compare the similarity of bills and laws to trace the flow of policy ideas (Wilkerson, Smith and Stramp, 2015), quantify legislative effectiveness (Casas, Denny and Wilkerson, 2020), or investigate the adoption and diffusion of policies (Linder et al., 2020; Hansen and Jansa, 2021; Hinkle, 2015). Scholars of political communication analyze text similarity to understand information distortion on social media (Anspach and Carlson, 2020) and how information changes as it flows (Carlson, 2019). Blumenau (2021) measures the influence of female members in the UK parliament using the similarity of legislators' speeches in political debate. Hager and Hilbig (2020) study the responsiveness of the German government based on the similarity of public opinion reports and government speeches.

The most widely used methods for measuring text similarity in political science research rely primarily on (almost) identical text segments or overlap of word sets in a pair of documents. Although words are the building blocks that transmit ideas, similar word usage in both documents is not always useful to identify when two texts convey the same *meaning*. This limitation is especially salient when the underlying documents are short, simply because there are not many words; therefore, word overlap is infrequent and noisy.

In this article, I build on recent advances in natural language processing to propose a meaning-based approach to estimate the similarity of short texts. The model, called cross-encoder, leverages state-of-the-art transformer-based neural networks, such as bidirectional encoder representations from transformers (BERT) (Devlin et al., 2018). In previous research, it was shown to do a better job of capturing the meaning of political texts compared to other embedding approaches (Wankmüller, 2019; Widmann and Wich, 2023). The particular aspect of cross-encoder models that makes them effective lies in processing a *pair* of documents together to extract feature representations. These representations are then fed forward into a task-specific neural network to classify documents in terms of similarity. Scholars can use off-the-shelf versions or customize their own cross-encoder models. Customization can include training with self-annotated data

and adjusting portions of the embedding model itself. All of these variations can be easily implemented with existing open-sourced software tools. Sample Python scripts are provided in Appendix A.

I illustrate this new approach to measuring the similarity of short texts with three applications in political science. In the first one, I compare the off-the-shelf cross-encoder estimations of information distortion with a standard unsupervised approach—cosine similarity—used by Carlson (2019). The results show that the cross-encoder is better at capturing the partisan biases presented in distorted social messages (subjects' summaries of a news article they read in a telephone-game experiment) than cosine similarity. In the second application, I collect original data from news headlines related to US Supreme Court (SCOTUS) cases to train a customized model. The purpose is to predict the semantic relationship between two news headlines as a way to identify different news frames. The customized model demonstrates that cross-encoders can be supervised to improve text comparisons. Not only does it outperform traditional word-based approaches, alternative embedding techniques, and crowdsourced human coders, but the resulting more accurate measurement of heterogeneity of news coverage enables me to detect that court decisions accompanied by dissenting opinions tend to receive more diverse media portrayals. In the third application, I demonstrate a new way of measuring elite polarization using the similarity of social media content from US senators' Facebook pages. Compared to using cosine similarity scores of posts represented by various embedding models (including bag-of-words, doc2vec, and BERT), the off-the-shelf cross-encoder estimations of polarization degree yield conclusions that are more aligned with established theories, which state that American politics is polarized regarding foreign affairs but the disagreement is less pronounced than in domestic policy.

The remainder of the paper proceeds as follows. I first describe the challenges that short texts pose to standard approaches of text similarity measurement in political science research. I then propose a meaning-based approach to compare short texts. Its implementation relies on a transformer-based language model, called cross-encoder (De-

vlin et al., 2018; Rosa et al., 2022), the most recent innovation by computer scientists for information retrieval. Next, I offer an intuitive explanation of the cross-encoder model, including how the underlying transformer extracts features from text pairs and utilizes a neural network classifier to estimate their similarity based on *pair*-level embeddings. I further demonstrate this approach in three political science studies: information distortion during social transmission, competing media framing of SCOTUS decisions, and elite polarization. Finally, I conclude with a discussion about the prospects and limitations of using cross-encoders to measure the similarity of short texts in political science research.

Measuring Text Similarity in Political Science

Measuring text similarity is an important task in a wide range of research areas in political science. A prominent example is examining policy diffusion. Wilkerson, Smith and Stramp (2015), for instance, detect similar sections of final laws and previous bills to trace the flow of policy ideas in the US Congress. Through comparing bills from all US state legislatures across policy domains, Linder et al. (2020) investigate the asymmetric effects of state partisanship on policy adoption and Hansen and Jansa (2021) find that low-resource state legislatures copy bills more frequently than their high-resource counterparts, except for complex policies. Hinkle (2015) studies whether the court ruling of constitutionality on one state's statute can affect other states' legislation. Casas, Denny and Wilkerson (2020) measure legislative effectiveness by identifying the texts of failed bills that later become provisions of other bills enacted into law. Düpont and Rachuj (2022) conduct cross-boarder analyses of party policy diffusion by comparing party manifestos in nineteen European countries.

In addition to legislative studies, scholars have measured the similarity of other types of documents, such as political speeches, social media posts and comments, and news and summary messages. To name a few, Carlson (2019) examines how information changes during media and social transmission by comparing the contents of an official report,

news articles, and interpersonal messages. Anspach and Carlson (2020) investigate how comments on news stories differ from the actual content of the articles being shared on social media. Also, Hager and Hilbig (2020) study the German government's responsiveness by estimating the similarity between cabinet members' speeches and public opinion reports. Giavazzi et al. (2023) find increasing language similarity between tweets posted by German constituencies and the radical-right party, AfD, after terrorist attacks.

Nevertheless, one caveat to the study of text similarity is that standard methods struggle to perform adequately when the underlying documents are short. As I discuss below, political scientists have primarily relied on methods to either identify segments of (almost) equivalent texts or compare the word sets used in any pair of documents. In short texts, similar word usage (with or without retaining the order of words), which indicates lexical similarity, does not readily translate into a shared meaning, which refers to semantic similarity. Social media posts (tweets), news headlines, and the like simply do not contain enough words to make word-based methods a reliable indicator of semantic similarity. This is particularly problematic given the increasing prevalence of these forms of textual data in modern political science research.

The three examples in Table 1 illustrate that lexical similarity can be an unreliable indicator of semantic similarity. Each news headline concerns a SCOTUS decision, and each pair covers the same case. Words that appear in both headlines are in bold.

Table 1 shows that the relationship between lexical and semantic similarity of news headlines is, at best, mixed. News headlines that are similar in terms of words may or may not differ in meaning. For example, the first pair of headlines (row one) is almost identical in both words and meaning. However, the second pair consists of similar words while conveying opposite meanings. The final pair conveys the same meaning without word overlap.

Table 1: News Headlines of SCOTUS Decisions

	Headline A	Headline B
Similar Words, Similar Meanings	**Supreme Court strikes down Louisiana** law on **abortion** clinic restrictions	**Supreme Court strikes down Louisiana abortion** limits
Similar Words, Different Meanings	**Supreme Court** hands **Trump** wins on tax returns, **financial records**	**Supreme Court** satisfies neither **Trump** nor his enemies in **financial records** cases
Different Words, Similar Meanings	SCOTUS Extends Title VII Protections to LGBT Employees	Employers Can't Discriminate Against Gay and Transgender Individuals, Supreme Court Rules

The issues highlighted in Table 1 reveal the need to go beyond a focus on lexical similarity and engage with the meanings of text pairs instead. To do so, I draw on recent work in computer science, which shows that deep-learning methods can perform well in measuring semantic similarity (Chandrasekaran and Mago, 2021). Specifically, I introduce to political science cross-encoders (Devlin et al., 2018), a deep-learning method that utilizes a state-of-the-art transformer-based language model to extract features of text pairs and estimates text similarity through a neural network classifier. Although the advantages of transformer-based language models have been demonstrated in classifying and generating politically relevant texts (Bestvater and Monroe, 2023; Widmann and Wich, 2023; Licht, 2023; Argyle et al., 2023), I carefully examine the validity of cross-encoders in estimating text similarity to obtain quantities of interest in this paper.

Methods for Measuring Text Similarity in Political Science

To begin, it is helpful to discuss how text similarity is measured in political science literature. Broadly, text similarity measurement involves identifying whether, or to what extent, texts are similar through pairwise comparison. Previous work relied on two categories of methods to measure text similarity: word sequencing and document vector.

Here, I discuss both along with some of their applications in political science before explaining cross-encoders.

Text-as-Sequence

Sequence methods represent texts as sequences of words. Words and their order and even punctuation are retained. This category of methods estimates similarity based on the degree to which parts of sequences match across documents. Smith-Waterman (SW) local alignment algorithm (Waterman, Smith and Beyer, 1976) is a well-known sequencing method, which has been applied to compare legislative texts (Wilkerson, Smith and Stramp, 2015; Linder et al., 2020; Gava, Jaquet and Sciarini, 2021). This approach performs all possible alignments of two sequences and identifies matched and mismatched sequences of texts as well as gaps between them. The algorithm scores text similarity higher when the length of matched sequences is longer and penalizes mismatched sequences and gaps.

A related innovation is from Casas, Denny and Wilkerson (2020). It adopts n-gram (a contiguous sequence of n words) matching to find blocks of shared texts in any pair of documents and constructs similarity statistics, including the longest matching sequence, average matching sequence length, and number of unique matching blocks. Once identified, these features are used in a supervised machine learning model to predict text similarity.

The limitation of sequencing methods is that they require documents to be relatively stable in their terminology and syntax. To date, the applications of these methods in political science are mostly restricted to legislative documents, which are written in formal and standard language. Sequencing methods are not effective in identifying the similarity of texts when applied to materials with higher variations of word choices and syntaxes. Short texts comprise exactly those features as they are easy to rephrase using synonyms and informal expressions.

Text-as-Vector

In the context of measuring text similarity, text-as-vector methods exploit the capability that documents can be represented as vectors using a number of methodologies. The most common approach, called bag-of-words (BoW) representation, discards the order of words in texts and usually requires some pre-processing to remove punctuation, infrequent terms, and stemming before converting documents into term-frequency vectors[1] (Carlson, 2019; Anspach and Carlson, 2020; Hager and Hilbig, 2020; Düpont and Rachuj, 2022).

Once accomplished, it is possible to measure the similarity between every two documents, usually in terms of the angle between their vectors, which is known as cosine similarity. Due to the relatively flexible vector representation of documents, cosine similarity can assess text similarity based on word usage without discounting the differences in syntax. Therefore, cosine similarity is often used to compare two sets of documents with distinct features, such as transcripts of government speech and research reports on public policy preferences (Hager and Hilbig, 2020), political debates among Congress members who may have diverse speaking styles (Blumenau, 2021), or texts from varying information sources—media outlets or the grapevine (Carlson, 2019).

Despite its popularity, the BoW approach has two main disadvantages that could lead to inaccurate measurement of semantic similarity. It treats each feature (unique words in the corpus) independently and does not sufficiently leverage the surrounding words because of discarding word order.[2] Consequently, synonyms are considered distinct features and polysemous words—terms that may have different meanings in different contexts—are indistinguishable.

In contrast, word embedding models, such as GloVe (Pennington and Manning, 2014), Word2Vec (Mikolov et al., 2013), and their extension doc2vec (Le and Mikolov, 2014)[3], are trained in neural networks that output numeric vectors for each unique word in the

[1] There are variations of bag-of-words representation, such as n-gram and TF-IDF.

[2] Even though bag-of-n-grams considers the word order in a short window, it suffers from data sparsity and high dimensionality.

[3] Doc2vec is an extension of Word2Vec that incorporates document vectors along with word vectors in the training process to predict the next word in a document.

corpus, ensuring that words of similar meanings are close to each other in the hypothetical space. Although word embeddings are arguably better at capturing semantic closeness between words, they have two fundamental limitations. First, they do not solve the problem of polysemy—each word is associated with a single static vector, which is not context-dependent. Thus, the word "strike" is assumed to have the same connotation in texts about the Supreme Court, labor disputes, and baseball scores. Second, it is unclear how effective the word embedding technique is at discerning the similarities or differences in meanings between two documents if each piece is simply the aggregation (e.g., average or weighted sum) of static word vectors. Perhaps for this reason, political scientists usually exploit the ability of word embeddings to find semantic relatedness between words (Rheault et al., 2016; Rodman, 2020), but find no substantive improvement when comparing documents relative to BoW approaches (Ziegler, 2022).

Instead, I propose a new method built on the most recent and advanced embedding technique—contextualized text representation—which is capable of extracting features of texts at the sentence (hereinafter, sentence embedding) and even *pair* level (hereinafter, *pair* embedding).

Moving Toward Contextualized Text Representation

One breakthrough in large language model development was the use of transformers (Vaswani et al., 2017). Transformers are a type of neural network architecture that employs self-attention mechanisms, allowing the model to weigh the importance of different words of the input text when generating text representations. Such text representations are contextualized since the model represents words in a way that accounts for the context in which they appear. Since the introduction of BERT (Devlin et al., 2018), computer scientists have broadened the scope of transformer-based language models by experimenting with various data, tweaking model hyperparameters, and even devising new model architectures. This work expanded the family of such models to include RoBERTa, DistilBERT, and ELECTRA (Liu et al., 2019; Sanh et al., 2019; Clark et al., 2020).

Transformer-based language models can provide sentence embeddings when the input is a single piece of text. Recent work in political science demonstrates the superiority of sentence embeddings in mining meaning from political texts. Bestvater and Monroe (2023) built a classifier to detect binary stances from tweets—approving or opposing the Women's March movement—based on BERT sentence embeddings. Widmann and Wich (2023) demonstrate that transformer-based models outperform dictionary and word embedding approaches to classify discrete emotions. Licht (2023) relies on multilingual sentence embeddings to categorize the topics and ideological positions of party manifestos. However, an important limitation here is that each text is processed in isolation.

Transformer models also can process a pair of texts simultaneously and represent them using one single vector, called *pair* embeddings. This enables the model to capture the relationships between two texts by interpreting their meanings within the context of each other. To the best of my knowledge, the potential of transformer-based models to embed text pairs in a vector space that can be trained for specific similarity tasks has not been explored in the political science literature. This paper aims to bridge that gap by introducing the *pair* embedding technique offered by a specific type of language model, cross-encoders, to more accurately capture the semantic similarity of short texts based on research inquiries.

I conclude this section with a comprehensive comparison of each method mentioned above (see Table 2). To summarize, word-based approaches cannot fully address the challenges of measuring the similarity of short texts. They would result in underestimation of semantic similarity between texts with shared meanings but different words, as well as overestimation of semantic similarity between texts with significant word overlap but distinct meanings. Cross-encoders model the interactions between two texts using pair embeddings, which are more capable of capturing subtle relationships that might be missed by sentence embeddings, which are context-aware within each individual text, let alone word embeddings, which only consider semantic closeness between individual words. More technical discussions of cross-encoder models follow in the next section.

Table 2: Overview of Methods for Measuring Text Similarity

Model	Advantages	Limitations	Usages
word-based approach (BoW, local alignment)	• easy to understand • no training corpus needed	• treat each word as an independent feature • reliance on word overlap can lead to over-inclusive or under-inclusive measures	• identify direct emulation
word embeddings (GloVe, Word2Vec)	• capture the semantic closeness of words so can recognize synonyms well • pre-trained models can provide word embedding vectors for commonly-used vocabulary • locally fit with own corpus is feasible in personal computer	• static word vectors are not context-aware • aggregated word vectors may oversimplify and misrepresent sentence-level meaning	• identify the similarity of texts that use different words but share meanings (broadly)
sentence embeddings (doc2vec, BERT)	• consider the order and context of words in sentences • recognize both synonyms and polysemous words • pre-trained models available	• most of the models are not specifically pre-trained for measuring text similarity • embedding each sentence separately may miss the nuanced differences that contribute to the relationship between two sentences	• identify the similarity of texts that use different words but share meanings (broadly) • distinguish between texts that contain similar words but convey different meanings (to some extent)
pair embeddings (cross-encoder)	• consider the entire context of both texts by processing pairs of texts together • capture the subtle and implicit relationships that might be missed when texts are embedded separately • off-the-shelf cross-encoders are available and can be fine-tuned for specific similarity tasks using self-annotated data	• customized training could be computationally intensive and require labeling text similarity • the decision to use off-the-shelf or customized cross-encoders is not trivial	• identify the similarity of texts that use different words but share meanings (broadly or specifically, depending on how the similarity of texts is annotated in training data) • distinguish between texts that contain similar words but convey different meanings

An Introduction to Cross-Encoders

In state-of-the-art Natural Language Processing (NLP) models, such as Long Short-Term Memory Recurrent Neural Networks (Staudemeyer and Morris, 2019) and transformers (Vaswani et al., 2017), text is analyzed as a series of tokens. Tokens are fundamental semantic units, typically consisting of words in most languages. These tokens and their order contain information about the meaning of texts. To quantify such information, NLP models map them to numeric values that computers can understand. This process is called feature extraction. The models make inferences from these features to specific tasks, like text classification.

Cross-encoders follow the same basic routine when estimating text similarity. As shown in Figure 1, there are two main parts: feature representation (tokenization and embedding) and classification.

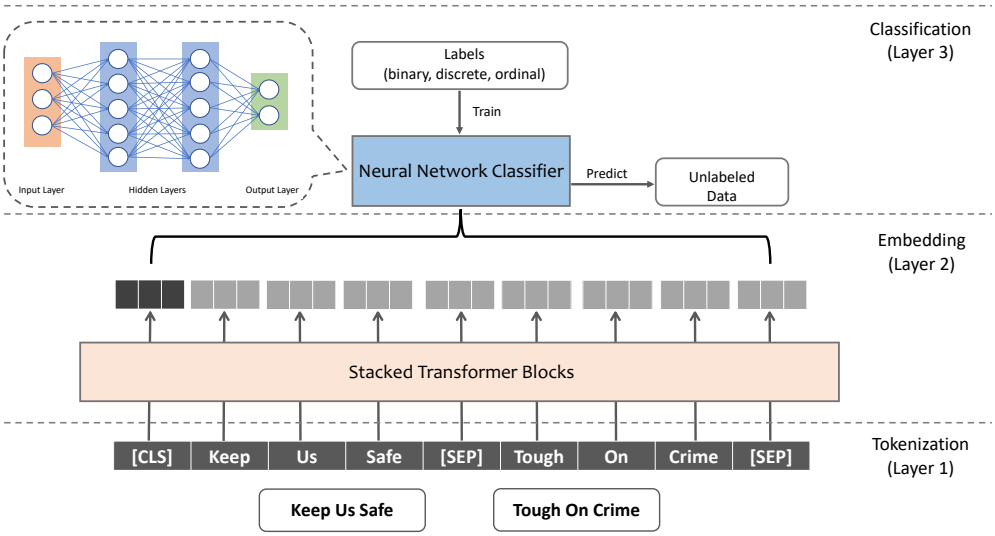

Figure 1: Cross-Encoder Model Architecture. The figure presents three layers of a cross-encoder model. In Layer 1 (Tokenization), the input pair of texts are processed as a sequence of words and punctuations. In Layer 2 (Embedding), the transformer converts each token to a vector of numbers. In Layer 3 (Classification), a neural network classifier takes the vector of a special [CLS] token (or alternatively, the average of all token vectors) as input and outputs the similarity of text pairs.

Concatenate, Split, and Tokenize

Assume that we want to compare the similarity of these two campaign slogans: "Keep Us Safe" and "Tough On Crime." Cross-encoders will process two sentences[4] together (see Figure 1, Layer 1). To that end, two sentences are concatenated and split into a sequence of tokens, where each token represents a word/subword/punctuation unit. Additionally, two types of special tokens are inserted into the sequence: (1) one [CLS] token is prepended to the beginning of the sequence; (2) the [SEP] token marks the end of each sentence. Tokenization prepares the input pair of sentences for entering the transformer, which consists of a stack of dense neural network layers that are used to encode texts with numeric vectors.

[4]Here, sentence is defined as an arbitrary span of contiguous text, rather than an actual linguistic sentence *per se*.

Extract Features from Text Pairs

Different transformer models, mainly pre-trained large language models, have been used in processing the sequence of tokens and embedding them in a hypothetical vector space (see Figure 1, Layer 2). Here, we are interested in extracting features from the pair of texts to estimate their similarity. During processing, each token interacts with other tokens in the pair and is finally converted to a numeric vector. This allows tokens to be interpreted within their specific context. The complete transformer model consists of a dense network where this process is repeated in multiple layers, with each layer generating more refined representation of the input tokens. Finally, the vector of the special [CLS] token generated in the last transformer layer is used as a comprehensive feature representation of the whole pair of sentences. That is, the vector associated with the [CLS] token represents an embedding for the *pair* of documents in some latent space that can be used for downstream classification as similar or not similar.

A key insight is that each token does not contribute equally to the feature representation due to a self-attention mechanism. For example, when embedding the [CLS] token, the self-attention mechanism assigns different scores to each token of the input sequence. A higher score between two tokens indicates that the information contained in one token is deemed more relevant when processing or interpreting the other token, rather than direct measures of similarity between tokens. This mechanism allows each token to attend to all other tokens within the input sequence to compute a representation incorporating information about the token itself and other tokens in its own sentence and in the paired sentence. As a result, the numerical representations of each token are aware of the context of underlying texts.[5]

[5]This is helpful for the model to learn more information about similarity that can only be obtained by knowing the context of both sentences. For example, consider these two headlines: (1) *Conservative Justices Deny Accountability to Family After Cross-Border Killing of Their Son* (2) *Supreme Court Rules Against Family Of Teen Killed In Mexico By Border Patrol.* Bag-of-word approaches can only detect that they contain three common words ("Family," "Border," "Kill"); sentence embedding and word embedding models will further understand that "Justices" and "Supreme Court" are semantically close due to their ability of capturing meanings of texts at the sentence or word level. However, *pair* embeddings allow the model to leverage knowledge from both texts, e.g., "Their Son" and "Family of Teen" refer to each other, "Cross-Border Killing" and "Killed In Mexico By Border Patrol" are the

Cross-encoders utilize the self-attention mechanism of transformers to generate contextualized representations of text pairs by concatenating two sentences into a single sequence as input. The resulting *pair* embedding better reflects the degree to which the two sentences are mutually informative. As I will demonstrate in three empirical applications, this results in superior performance to word-based approaches that treat words as independent units stripped of context, or simply embedding each sentence separately.

Train a Task-Specific Neural Network Classifier

To estimate text similarity, a cross-encoder must build a neural network model on top of the last transformer layer to learn the complex patterns between input features and output labels. The *pair* embeddings are used as input features and text pairs are usually labeled using binary categories indicating whether two sentences convey the same message, discrete classes that reflect the relationships (such as entailment, neutrality, or contradiction) between sentences, or ordinal categories representing the degree of similarity.

A neural network model consists of a large number of interconnected nodes organized into multiple layers (see Figure 1, Layer 3). The input layer of a neural network first receives the *pair* embedding—a high-dimensional feature vector—and each node represents an element of the vector. The final layer outputs the predicted probabilities over each label, such as "similar" or "dissimilar." The class with the highest probability is typically considered the model's prediction.

Between the input and output layers are hidden layers. The connections from nodes in one layer to nodes in another layer, which are visualized as lines, are model parameters (weights and bias) that the neural network learns during training.[6] The optimization

same event, and "Deny Accountability" means "Rule Against" the family since both headlines concern a judicial decision.

[6] Lin and Lucas (2023) build from a statistical context familiar to social scientists—logistic regression—to introduce neural networks. A simple neural network comprising an input layer and an output layer with a sigmoid activation function (equivalent to the inverse logit) that maps linear combinations of predictors, weights, and bias to probabilities of binary classes is identical to a logistic regression model. Adding hidden layers to neural networks allows modeling more complex relationships in data. By replacing the

process, known as backpropagation, includes calculating the gradient with respect to each model parameter and adjusting them to minimize errors between the predictions and the true labels.[7]

The deep-learning community has published several versions of cross-encoder models, which are trained on benchmark datasets that label the similarity of texts from image captions, news headlines, and user forums. These off-the-shelf cross-encoders have performed well in paraphrase mining and question answering. They are cost-effective when the task does not examine a particular definition of text similarity and the corpus does not contain many domain-specific expressions (exemplified in the first and third applications) Researchers also can create customizing cross-encoders by training on domain-specific data annotated based on their inquiries (exemplified in the second application of the next section).

In general, cross-encoders do not directly utilize information from the entire corpus to estimate text similarity since they are primarily designed to consider the input from two documents in a pair at a time.[8] However, customizing cross-encoders involves fine-tuning the underlying transformer model, a process of updating the parameters[9] relying on feedback obtained from prediction errors of the neural network classifier. To some extent, fine-tuning leverages information from all training documents to refine the feature representations of text pairs generated from the transformer model and creates an embedding space that adapts to specific similarity tasks.

sigmoid with softmax function, neural networks can be easily extended to handle multiclass classification tasks.

[7]The updating process of neural network weights and bias involves complex backpropagation algorithms. For more information, refer to Rumelhart, Hinton and Williams (1986).

[8]This is not inherently a limitation. Both off-the-shelf and customized cross-encoders are built on transformers that were pre-trained on large corpora, ensuring the models have a robust foundation of language, including recognizing synonyms and polysemous words (Garí Soler and Apidianaki, 2021).

[9]The parameters include the weights and biases of the feedforward neural networks, the weights for matrices in the self-attention mechanism, and so on.

Empirical Performance of Cross-Encoders in Political Science Applications

To illustrate the benefits of using cross-encoders, I present evidence from three empirical applications. First, using data from Carlson (2019), I apply a cross-encoder to measure information distortion during social transmission. I find that compared to the BoW cosine similarity approach used in Carlson (2019), the cross-encoder captures message distortion more precisely. The second application is a novel study, examining the heterogeneity of news headlines about SCOTUS decisions. For this example, I manually code the (dis)similarity of news headlines to train cross-encoders. This allows me to compare the performance of a wide range of approaches (including cosine similarity, local alignment, and even alternative embedding techniques, such as doc2vec and BERT) to measuring text similarity with the cross-encoder method. I find that cross-encoders are more accurate and they uncover patterns that otherwise would be missed, e.g., that cases with published dissents receive more diverse and politicized coverage than unanimous decisions. The third application presents a more challenging task where I estimate the similarity of social media posts from US senators that a topic model has already identified to be on the same subject. I apply a cross-encoder to measure the similarity of discussions on domestic and international topics from US senators and test the conventional wisdom that political disagreement in American politics "stops at the water's edge."

Application One: Information Distortion During Social Transmission

This application is a reanalysis of Carlson (2019) in which the author studies information distortion based on text similarity. In this experiment, subjects played a "telephone-game" where they read a Reuters article about the US economic performance (shown in Appendix B.1) and shared their summary with another hypothetical person. Carlson

(2019) argues that "if information is changing as it flows from one source to the next, we should expect fewer words to be the same between documents at each stage." Therefore, Carlson (2019) calculated BoW cosine similarity between the original article and "social messages," summaries of the original article provided by participants of the experiment, to quantify information distortion. By doing so, the author assumes that higher cosine similarity scores mean there are more words in common between two documents and less information distortion. However, such an assumption does not hold if the messages share some words with the Reuters article and still invent information that is not originally present. In this experiment, BoW cosine similarity easily overestimates the similarity between texts that overlap in words but convey different meanings; thus, does not sufficiently capture information distortion. To mitigate this issue, I apply an off-the-shelf cross-encoder, which has fine-tuned the RoBERTa model with a dataset of sentence pairs annotated with different degrees of similarity ranging from no overlap to equivalence in meaning (Cer et al., 2017), to measure the similarity between social messages and the Reuters article. This off-the-shelf cross-encoder is a good fit here because the more social messages preserve the meaning of the original article, the less information is distorted.

The quantile-quantile (QQ) plot (Figure 2) compares the distributions of normalized similarity scores estimated by these two methods. The BoW cosine approach estimates a much higher degree of similarity of social messages than the cross-encoder across a broad spectrum of medium scores, but both methods remain relatively consistent for messages with extremely low or high levels of similarity.

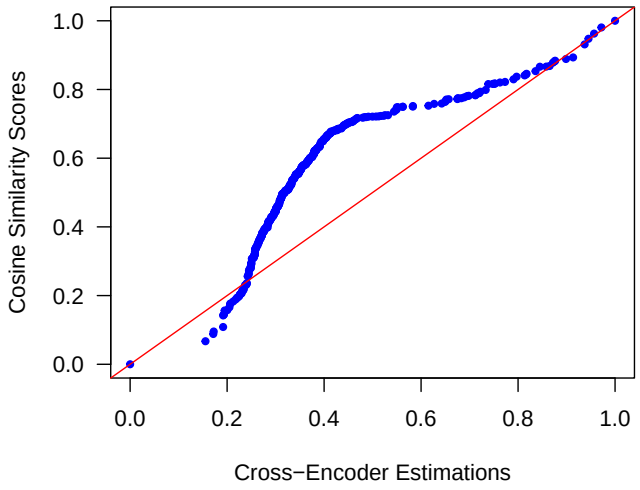

Figure 2: QQ Plot between the Distributions of Cross-Encoder Estimations and Cosine Similarity Scores. Both scores are normalized ($X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$) to the same scale [0, 1]. The correlation between them is 0.45 (modest).

After qualitatively examining instances where the two measures disagree, this pattern is not surprising, since experimental subjects usually summarized the Reuters article followed by their opinions and interpretation. However, BoW cosine similarity does not effectively discount the level of text similarity in cases where messages use similar terms while producing information that deviates from the original article. Table 3 shows the three top instances where the cosine similarity scores indicate similarity to the Reuters article, but the cross-encoder does not. The comments in the right column discuss aspects of the article where the meaning has been distorted in important ways (highlighted in the messages) that are obviously missed by the BoW cosine approach.

To further examine which model more accurately estimates proxies for information distortion, I test the association between the manually-coded amount of partisan bias and the automated text similarities between social messages and the Reuters article. In a book project that extends her work, Carlson (2024) identifies partisan bias as the main source of information distortion during social transmission. Since "Reuters is an objectively neutral news source" (Carlson, 2019), as subjects' summaries contained more partisan opinions, I expect the messages would become less similar to the original article, indicating higher levels of information distortion. Table 4 shows the results of univariate

Table 3: Top three instances of social messages where cosine similarity classifies them as similar to the Reuters article but the cross-encoder does not, using a threshold of 0.6 for both measures.

Social Messages	Comments
GDP growth has been the slowest since the second quarter of last year. Economic growth is also at the slowest pace since the second quarter of 2013. The new presidency is bad for the economy so far.	The original article does not comment on the new presidency.
According to the article, we are seeing some growth in the GDP in the first quarter of this year. President Trump is giving tax cuts for major businesses, which have been exporting a lot more. Consumers have also affected the GDP by spending more than usual. Although [there] is some growth, it is much smaller than what the US has done in the best and these trade deals could prove to be valuable for the future of our economy.	The original article talks about slow spending while the message talks about increased spending. The article also does not mention trade deals.
You need to start leaning to one side of the spectrum, at least on very specific issues. This article I read clearly indicates that the US economy is showing a nearly standstill growth. This is largely due to the do-nothing Republicans in the house & senate. The Democratic way is the only way to progress this country forward. The growth rate is the slowest it has been in 5 years. Seems like the Obama administration was doing something right, eh?	The message praises Democrats and contains criticism of Republicans that is not founded in the original article.

regressions, where the explanatory variable is *message distortion*—the number of units of information that favored or opposed Democrats or Republicans expressed in the social messages—and the dependent variable is *similarity* scores between the Reuters article and the social messages.[10] As is evident, only when *similarity* is measured by cross-encoder, the coefficient of *message distortion* is negative and statistically significant. In other words, the more distortion there is, the less similarity between the original article and social messages. This demonstrates that cross-encoder can sufficiently capture information distortion based on accurate measures of text similarity.[11]

[10]The purpose is to examine the direction and significance of the association between them rather than establishing any causality. Since message distortion is a count variable, I use it as an independent variable and continuous similarity scores as the outcome variable so they fit simple linear regressions. Alternatively, I calculate the Pearson correlation between these two variables, showing that the correlation between cross-encoder estimates and message distortion is -0.17 ($p < 0.001$); between cosine similarity scores and message distortion is -0.03 ($p > 0.5$).

[11]In the main text, I focus on the advantages of cross-encoder compared to BoW cosine similarity,

Table 4: Associations between Message Distortion and Text Similarity

	Similarity between the Reuters Article and Social Messages	
	Cross-Encoder	Cosine (BoW)
Intercept	0.405^{***}	0.541^{***}
	(0.010)	(0.011)
Message Distortion	-0.022^{***}	-0.004
	(0.006)	(0.007)
Num. obs.	399	399

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.
a) Message distortion is manually coded by a team of research assistants hired by Carlson, which was operated as counting the units of partisan information in each social message.
b) Similarity scores are normalized to the same scale [0, 1].

Application Two:

Competing Headline Framing of SCOTUS Decisions

In this application, I conduct a novel analysis of media framing of US Supreme Court decisions with the goal of measuring the heterogeneity of media content using text similarity. Media outlets usually have alternative emphases on an event, known as competing framing (Chong and Druckman, 2007; Druckman et al., 2010; Druckman, Peterson and Slothuus, 2013), and news coverage of court decisions is no exception. How court decisions are framed is particularly important because previous experimental work shows that it can affect public attitudes toward the judicial institution (Clawson and Waltenburg, 2003; Baird and Gangl, 2006; Nicholson and Howard, 2003; Hitt and Searles, 2018). However, prior work has not been able to characterize or study the extent to which media outlets adopt different frames at scale due to challenges in automated recognition of varying news frames within the same case.

To automatically identify heterogeneous frames, I created an ensemble of customized cross-encoders to predict the semantic relationship of all headline pairs written about the same Court ruling. I then use the (dis)similarity of headlines as an indicator of competing framing on SCOTUS decisions to test the intuitive hypothesis that *case decisions with*

which is the main approach used by Carlson (2019). In Appendix B.1.2, I present results using less computationally intensive models, including GloVe and doc2vec. Still, cross-encoder more effectively captures information distortions.

published dissents from the Court are associated with more heterogeneous frames. Non-unanimous decisions suggest that the issues behind the cases are more contentious and newsworthy. Therefore, the public may be more interested in them, and news outlets are more incentivized to tailor their frames to their audiences, leading to diverse depictions of the same decisions.

Using headlines covering SCOTUS decisions, I find that cross-encoders are able to uncover the association between heterogeneous media coverage and published dissents. Once again, this relationship is not evident using other standard approaches in the field. Moreover, I show that cross-encoders provide more accurate estimates of text similarity than a wide range of embedding models and even outperform crowdsourced coders on this task.

Accurate Predictions on Text Similarity

I collected news articles that reported SCOTUS cases decided during January to July 2020.[12] Since I am interested in framing, I examine the news headlines instead of the full stories. Headlines are designed to grab attention and usually highlight the particular angle given to a news story. I compare every two informative headlines about the same case, resulting in a dataset of 27,407 pairs of headlines to be analyzed. More details about the data collection are available in Appendix B.2.1.

Then, I label the similarity of 1,022 pairs,[13] drawing on the concept of *entailment* from linguistics: a pair of news headlines is considered to be similar if the statement in one news headline is true given that the statement in the other is true. The resulting similarity reliably indicates the use of different news frames on the same case decision. For instance, consider the three news headlines about BOSTOCK v. CLAYTON COUNTY, GEORGIA. Headline (a) "SCOTUS Extends Title VII Protections to LGBTQ Employees" and

[12] I collected and cleaned the news headlines dataset in the summer of 2020, so the most recently decided SCOTUS cases were in July of that year.

[13] To guarantee that sufficient cases with fewer news headlines could be selected to train the model, I oversampled headlines on infrequently reported cases while undersampling those received more media coverage.

Headline (b) "Employers Can't Discriminate Against Gay and Transgender Individuals, Supreme Court Rules" frame this case as a victory for LGBTQ workers and are related to each other. On the contrary, Headline (c) "Seventh-day Adventist Church Responds to U.S. Supreme Court Employment Decision Impacting Religious Liberty" emphasizes the impact on religious liberty, which is semantically irrelevant to (a) or (b).

With these 1,022 pairs of labeled news headlines in hand, I train models using the sentence-transformers library in Python. After the 5-fold cross-validation (CV), I created an ensemble of customized cross-encoders, consisting of five models that obtained the highest accuracy within 10 epochs.[14] Each model's performance is evaluated by several classification metrics. In this case, precision[15] indicates how effectively the model can identify the distinct meanings of texts, despite sharing common words; recall[16] measures the model's ability to capture semantic similarity between texts using different words. As shown in Table 5, cross-encoders have relatively balanced and high precision and recall rates, suggesting that the method performs well in both scenarios. However, word-based approaches, such as BoW cosine similarity and SW local alignment, usually face a trade-off between precision and recall. Predicting texts to be similar in meaning based solely on a high overlap of words can lead to an increase in false negatives, thereby reducing the recall; but, relaxing the criteria for word overlap to identify texts as similar could result in overlooking subtle differences in meaning, which can produce false positives and thus decrease the precision.[17]

[14]For each fold I saved the model parameters in the epoch having the best out-of-sample predictions on the hold-out validation set.

[15]Precision $= \frac{TP}{TP+FP}$, minimizing FP (false positive) increases precision.

[16]Recall $= \frac{TP}{TP+FN}$, minimizing FN (false negative) increases recall.

[17]I examine this problem by fitting logit regressions using similarity scores to predict whether the news headlines are similar. Word-based approaches have high precision (BoW cosine: 0.73; SW alignment: 0.69) and low recall rates (BoW cosine: 0.56; SW alignment: 0.50). This suggests that word-based approaches have underinclusive measures of the similarity of these news headlines, resulting in false negative predictions where texts have similar meanings even though they use different words.

Table 5: Model Performances of Customized Cross-Encoders

Hold-Out Fold	Best Epoch	Accuracy	Precision	Recall	F1
1th	5	0.90	0.81	0.89	0.84
2th	6	0.90	0.85	0.75	0.80
3th	6	0.94	0.98	0.83	0.90
4th	9	0.90	0.81	0.88	0.84
5th	9	0.91	0.86	0.87	0.86
Average		0.91	0.86	0.84	0.85

Additionally, I assess the accuracy of the customized cross-encoders with several benchmarks. First, the performance of cross-encoders exceeds qualified layman coders (except one superb worker) who were hired from Amazon Mechanical Turk. The training module and task description are available in Appendix B.2.2. They labeled 2,000 pairs of news headlines, mixed with 500 expert-coded pairs. The average accuracy rate of workers' answers to the 500 pairs is 0.81, lower than the model's own out-of-sample predictions (0.91). As shown in Figure 3, workers who demonstrated higher accuracy in answering these 500 pairs also exhibited stronger alignment between their coding for the remaining 1,500 pairs of news headlines and the labels predicted by the ensemble model of cross-encoders.

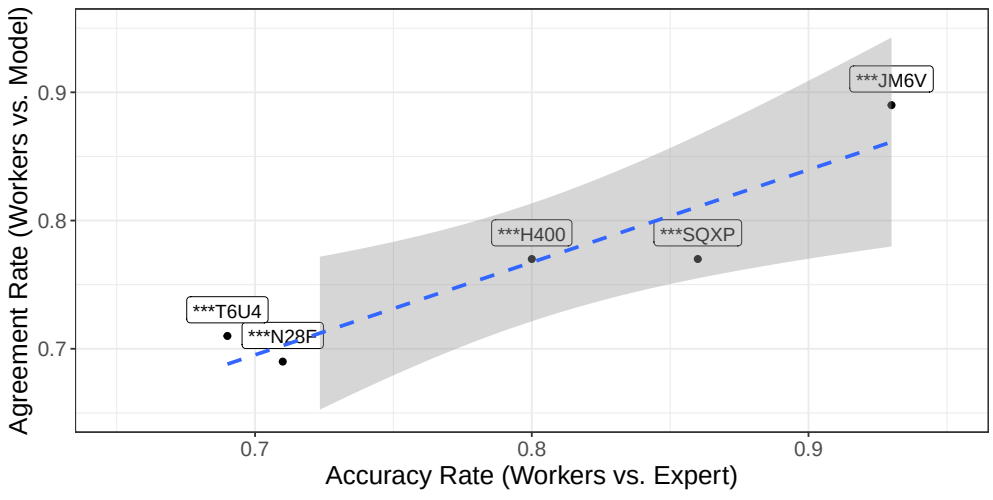

Figure 3: Better-Quality Workers Agree More with Model Predictions. The x-axis represents each worker's accuracy rate of their answers (using the labels of 500 expert-coded text pairs as ground truth). The y-axis is each worker's agreement rate between their answers and the model-predicted labels of the remaining 1500 text pairs. Only workers who have coded more than 30 pairs of news headlines are included in the analysis.

Second, using this same CV approach, I compare the performance of customized cross-encoders to a wide range of approaches, namely the off-the-shelf cross-encoder, cosine similarity (of different embedding models), and sequence alignment. Based on the area under the ROC curve (AUC) (see Figure 4), which measures the overall performance of a binary classifier, customized cross-encoders outperform all others.

There are several takeaways from this comparison. First, training with domain-specific data modestly improves the performance of a cross-encoder ($0.94 > 0.92$). Second, surprisingly, the predictive ability of locally-trained doc2vec, pre-trained BERT, or SRoBERTa[18] is no better than random guessing (each model has an AUC of around 0.5). It is possible that these models, which estimate text similarity based on sentence embedding, are not sensitive enough to identify the entailment relationship of news headlines that requires the context of both texts. Third, the two commonly used word-based methods in political science—BoW cosine similarity and SW local alignment—provide reasonable estimations (both AUCs > 0.8) of text similarity. This is likely thanks to those straightforward cases where the lexical overlap is indicative of semantic similarity.

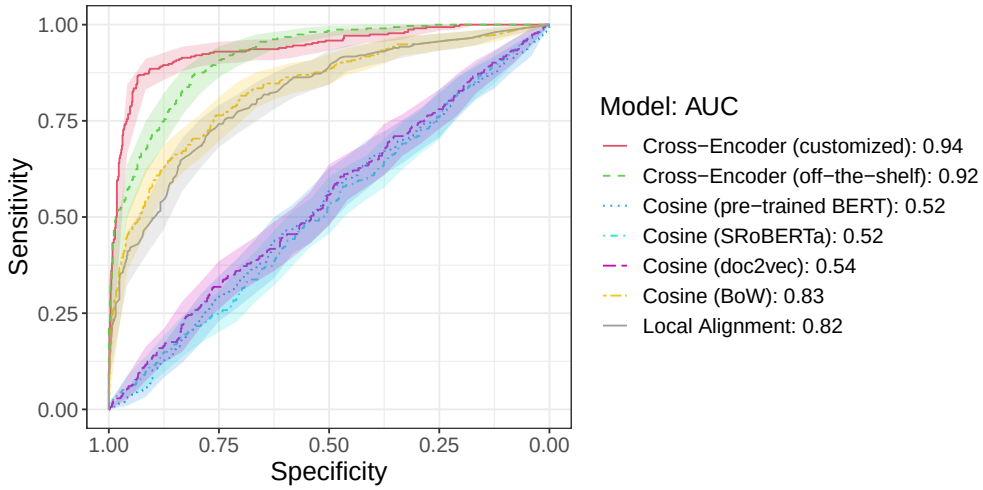

Figure 4: Area under the Curve (AUC). Similar to the customized cross-encoder, the other models are fit with logit regression on the labeled pairs. The calculation of AUC is based on the model responses on the validation set using a 5-fold CV. The 95% CIs are constructed by bootstrapping the predictions.

[18] Also known as bi-encoder, which modifies the pre-trained BERT or RoBERTa model using siamese and triplet network architectures to generate semantically meaningful sentence embeddings that can be compared using cosine similarity (Reimers and Gurevych, 2019). An in-depth comparison between bi-encoder and cross-encoder is beyond the scope of this paper.

Politicized Coverage of Split Decisions

Next, I apply these competing methods to test the intuitive association between split court decisions and heterogeneous media coverage. Each observation is a pair of news headlines from articles reporting the same decision. The main explanatory variable *unanimity* indicates whether a Court decision is accompanied by a dissenting opinion. The outcome variable *heterogeneity* is constructed based on the similarity scores of the pair of news headlines about the same case, subtracting the value from 1 to represent the heterogeneity of news coverage.[19] To facilitate comparable analyses, I standardized the scores.

I also include both *salience* and *issue area* as control variables to account for potential confounding effects. Salient and fundamental right cases may not only lead to the Court's most controversial opinions but also prompt media to cater to their audiences' policy preferences when reporting the Court's decisions. I measure the degree of salience by the logged number of amicus briefs filed for each case. I also determine whether a case is about fundamental rights based on the fourteen issue areas (the subject matter of the controversy discussed in each case) identified in the Supreme Court database.[20] The following four issue areas—criminal procedure, civil rights, First Amendment, and due process—usually cover cases pertaining to fundamental rights that protect individual liberty (e.g., freedom of speech, freedom of religion, and protection against self-incrimination, unreasonable searches and seizures) from government encroachment. Accordingly, I divide all issue areas into two categories: the above four and others.[21]

To model the effect of *unanimity* on *heterogeneity*, I apply the ordinary least squares (OLS) regression.[22] Estimations of the standardized coefficient of *unanimity* are presented in Figure 5, showing that both versions of cross-encoders provide high-quality measures

[19] More details on the coding of these variables are available in Appendix B.2.4.

[20] The database is available in http://scdb.wustl.edu/index.php.

[21] Although I believe that issue areas are informative enough to indicate whether the questions before the Court pertain to fundamental rights, I provide an alternative approach in Appendix B.2.6 that looks at the legal basis considered in the case. The main conclusion that only cross-encoders can detect the significant and negative association between unanimous decisions and heterogeneous media coverage holds.

[22] Regression results can be found in Appendix B.2.5. Standard errors are clustered at the case level. The model is specified as $\texttt{Heterogeneity} = \gamma_0 + \gamma_1 \texttt{Unanimity} + \gamma_2 \texttt{Salience} + \gamma_3 \texttt{Issue Area} + \varepsilon$.

of the heterogeneity of headlines, allowing the regression model to uncover significant and negative effects of unanimity on the heterogeneity of news coverage. This finding indicates that split decisions are more likely to be portrayed in a politicized manner by the media, leading to the use of different news frames. In contrast, neither word-based methods nor sentence embedding approaches detect such a relationship with statistical significance.

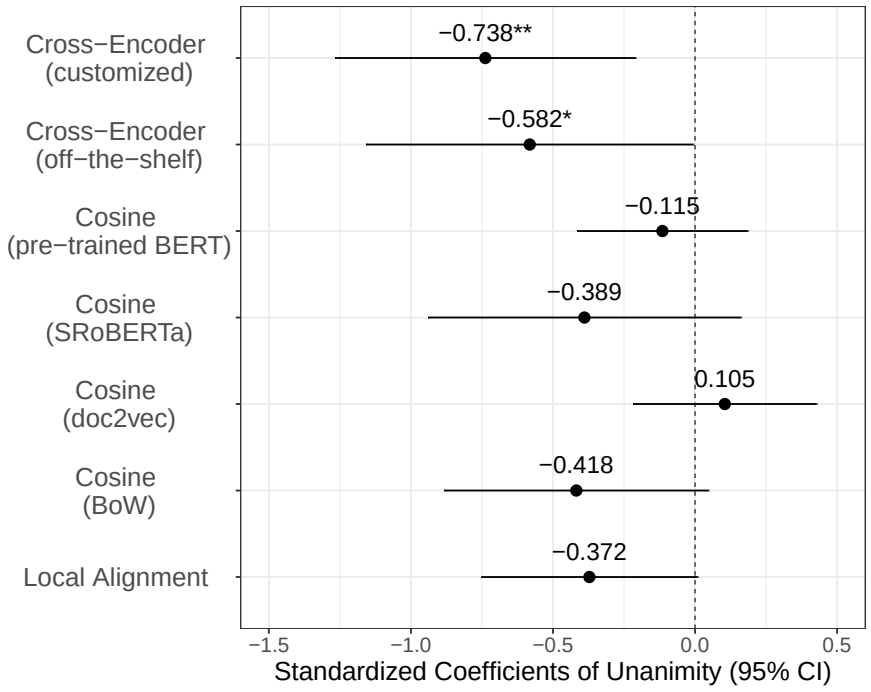

Figure 5: Effects of Unanimity on the Heterogeneity of News Coverage. The x-axis represents the standardized effect sizes ($^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$) with 95% CI. Effects are estimated by the same OLS model while the outcomes are measured by different methods. Correspondingly, the y-axis lists the name of each method.

Comparing these coefficients clearly shows that, while the customized training of cross-encoders may yield only marginal improvements in model performance (as measured by AUC in Figure 4), the increment translates into a notable increase of statistical significance in the second-stage empirical test. This suggests that customized training with domain-specific labeled data can be beneficial in both model performance and empirical testing. Also, we can be more confident about the empirical results, knowing that the validated machine predictions are good proxies for the variable-of-interest. To be clear,

all of the coefficients in Figure 5 have negative signs, except the Cosine(doc2vec) model. And cross-encoders do not always lead to significantly different estimates of the effect of unanimity.[23] Still, the improvement of cross-encoders in the accuracy of measuring text similarity enables the empirical model to identify effects significantly different from zero, which is important for hypothesis testing in social science.

Application Three:

Does political disagreement end at the water's edge?

Finally, I examine political polarization among inter- and intra-party legislators based on the text similarity of their social media posts. To test the conventional wisdom that "politics stops at the water's edge," I measure the degree of elite polarization in domestic and international issues using a corpus collected by Ying, Montgomery and Stewart (2022) from Facebook pages of senators who served in the 115th US Congress. I subset the corpus to posts published in 2017 when all members were in office. Furthermore, Ying, Montgomery and Stewart (2022) identify ten domestic topics and ten international topics from the STM estimate of that corpus. For each topic, I select relevant posts based on whether they have the highest proportion of discussion on that topic. I paired every two relevant posts within the same topic published in the same week by different senators, resulting in 100,999 observations.[24]

This is a challenging test because I need to examine the nuanced meaning of posts that discuss the same policy issue and highly overlap in word usage. Here, I do not intend to train a customized cross-encoder as the comparison of policy views is high-dimensional, which results in difficult manual labeling. Instead, I apply the off-the-shelf cross-encoder model, which estimates continuous scores between 0 and 1 representing

[23] See Appendix B.2.5. I conduct one-sided z-tests, showing that the effect of *unanimity* is larger when using customized cross-encoder than pre-trained BERT ($p < 0.01$), doc2vec ($p < 0.01$), and local alignment ($p < 0.1$).

[24] I only compare posts within the same topic because I am interested in the extent that politicians have different opinions on the same policy. I also set the time frame to be one week, since conversations usually take time and there may be a delay between responses from different senators.

semantic similarity of sentences, to measure the degree of polarization. This approach is similar to Myrick (2021), which relies on the partisan differences in language to measure polarization but differs in that I focus on meaning rather than word usage.[25] Additionally, each pair of posts is coded as *inter-party* if they were published by senators from different parties.

For comparisons, I calculate the cosine similarity of these posts represented by different embedding techniques and standardize all the resulting scores. The range of embedding techniques spans from the traditional and most commonly-used BoW approach to cutting-edge deep-learning methods—shallow neural network (doc2vec) and very deep transformer-based language models (BERT)—which recently appeared in political science journals.

My aim is to address two views promoted at the intersection of International Relations and American Politics. One suggests that US politicians fail to maintain bipartisan cooperation on US foreign affairs due to the absence of external threats after the Cold War, as well as increasing domestic ideological division and partisan electoral competition (Kupchan and Trubowitz, 2007; Busby and Monten, 2008; Trubowitz and Mellow, 2011; Milner and Tingley, 2015; Jeong and Quirk, 2019). This implies that intra-party statements should be more similar than inter-party statements within international topics. Another view states that political disagreement on international issues is less pronounced compared to domestic ones (Bryan and Tama, 2022) and the American public generally perceives weaker partisan types in foreign policy (Kertzer, Brooks and Brooks, 2021). Prominent explanations include the existence of information asymmetries in foreign affairs enabling executive discretion and suppression of opposition criticism (Canes-Wrone, Howell and Lewis, 2008), and the incentive for politicians to prioritize domestic policy due to greater voter attention on domestic issues than foreign affairs during elections (Heaney and Rojas, 2015). This suggests that the inter-party differences should be lower for international topics relative to domestic topics.

[25] I review the existing approaches to examining elite polarization regarding foreign policy in Appendix B.3.1.

To test these expectations, I fit interaction models[26] to explore whether the domain of policy issues moderates the effect of *inter-party* post on the *similarity* of social media content. Figure 6 visualizes the marginal effects of *inter-party* post conditional on international versus domestic issues.

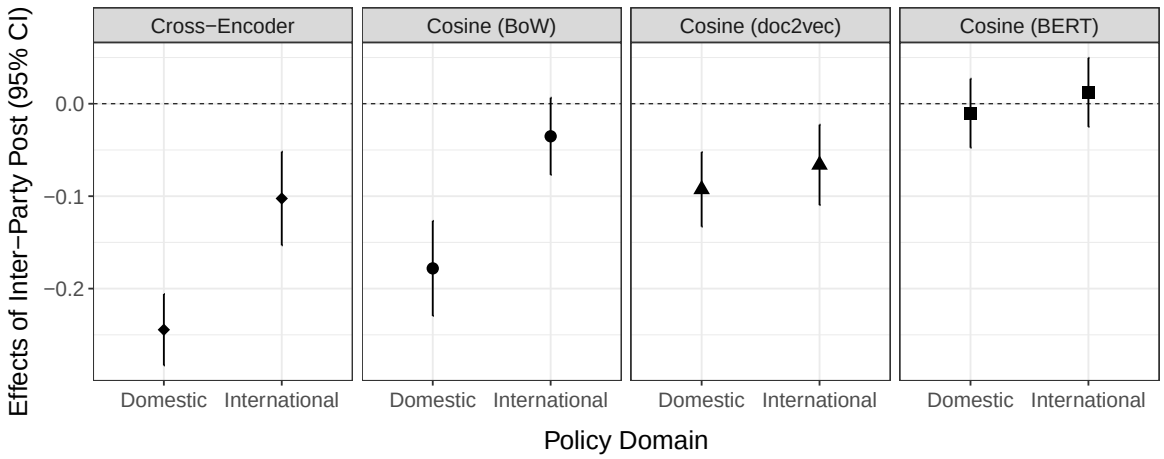

Figure 6: Conditional Marginal Effects of Inter-Party Post. The y-axis represents the effect sizes (with 95% CI) of inter-party post on the text similarity of social media content regarding the same policy issue. The x-axis displays the domain of policy issue (domestic vs. international), which conditions the effects of inter-party post.

In Figure 6, the cross-encoder model clearly shows that inter-party posts are generally less similar than intra-party ones and the effect amplifies when the underlying topics are domestic issues. However, the conclusions change when relying on cosine similarity scores. Specifically, the BoW model cannot detect the effect of inter-party posts on international topics, suggesting that bipartisanship still exists on foreign issues; the locally-trained doc2vec model indicates that the effects of inter-party posts on reducing content similarity remain the same across domestic and international topics; and the pre-trained BERT model suggests that there is no significant difference between the degree of similarity of inter-party posts and intra-party posts, implying that American politics is not polarized at all. This application demonstrates that cross-encoder estimations lead to conclusions aligned with prevailing views in the literature, whereas other methods do not.

[26]Regression results can be found in Appendix B.3.2. Standard errors are clustered at both the senator pair and topic levels. The model is specified as $\textbf{Similarity} = \beta_0 + \beta_1 \textbf{Inter-Party} + \beta_2 \textbf{International} + \beta_3 \textbf{Inter-Party} \times \textbf{International} + \varepsilon$.

Conclusion

Measuring text similarity is important for addressing various questions in areas such as legislative studies, political communication, and democratic representation. To overcome the difficulty of estimating the similarity of short texts, I introduce to political science an NLP model cross-encoder (Devlin et al., 2018; Rosa et al., 2022), which leverages *pair*-level embeddings to more precisely estimate the degree of similarity based on the context of both texts. Off-the-shelf cross-encoders are readily available for academic research[27] and applied scientists can train their own customized models using existing open-sourced tools.

In this paper, cross-encoders are shown to provide superior performance across tasks. The first application illustrates that the off-the-shelf cross-encoder's estimates of information distortion are better at capturing the amount of partisan bias in social messages than the standard BoW approach of cosine similarity. The third application also demonstrates that the off-the-shelf cross-encoder is capable of identifying the semantic (dis)similarity of language used by copartisans or outpartisans when they discuss the same topic. This allows me to measure the different degrees of elite polarization in international and domestic issues and test the conventional wisdom that "politics stops at the water's edge." Moreover, cross-encoders can be supervised to improve text comparisons when using specific notions of similarity. In the second application, the customized cross-encoder model not only predicts the semantic relationship of news headlines more accurately than alternative methods, ranging from word-based approaches to sentence-level embedding techniques, but also outperforms crowdsourced human coders. This allows me to detect statistically significant effects in the second-stage empirical inquiries.

Admittedly, the implementation of cross-encoders has some limitations. One constraint is that the maximum length of input sequences the model can accommodate is 512 tokens. Although the constraint of input length prevents using this method to directly

[27]Hugging Face, an online model repository, hosts cross-encoders, which are free to download from https://huggingface.co/cross-encoder.

estimate the similarity of long texts, such as congressional speeches and court opinions, these documents can be divided into paragraphs or sentences for more fine-grained analyses. They also can be converted to short texts using automatic summarization tools, but this may result in loss of information. In the near future, the maximum input of 512 tokens will likely be increased as computer scientists continue to develop larger transformer models capable of processing longer sequences.

The requirements of memory and hardware are another limitation of using this method. Like many other deep-learning models, cross-encoders prefer to process data on GPUs (graphics processing units) rather than CPUs (central processing units), which saves significant time. A GPU with a minimum of 12 to 16 GB of RAM is essential for training customized cross-encoders. Budget-constrained researchers may consider using computing resources offered by online platforms, such as Kaggle and Google Colab, which provide GPUs that are free of charge for tens of hours per week. The alternative is to use off-the-shelf cross-encoders. These models were trained on text similarity data for NLP applications such as information retrieval, detecting duplicate queries, and matching questions and answers. To obtain meaningful results, the task at hand should be comparable enough to these tasks and the corpus to be analyzed should not be too domain-specific.

Last, text similarity measurement can pose computational challenges due to the quadratic complexity of completing all pairwise comparisons, where the number of documents, denoted as n, leads to $O(n^2)$ operations. Two common strategies may be considered: sampling, which involves randomly selecting some pairs of documents to be compared, and blocking, which involves partitioning data into smaller sets for comparison. Blocking can be achieved using simpler algorithms, like clustering or keyword matching, to group similar documents, followed by more detailed similarity assessments using cross-encoders. Blocking rules can be devised based on theoretical considerations to determine which pairs are meaningful for comparison. For example, I focus on assessing the similarity of news headlines concerning the same SCOTUS decisions because the concept of interest, competing framing, revolves around the selective presentation of the same event.

Similarly, to test whether politicians exhibit less polarization in foreign affairs compared to domestic policy, I focus on comparing social media posts on the same topics to measure issue-specific polarization.

Nevertheless, the applications of cross-encoders in political science research are promising. The framework of pairwise comparisons can be useful to measure political concepts related to (dis)similarity. As exemplified here, I measure information distortion, heterogeneous frames, and polarization using text similarity. At the same time, we should be careful about using cross-encoders to identify direct emulation, such as diffusion of policy ideas and laws within the country. Still, as cross-encoders estimate text similarity without the constraint of lexical overlap, they have great potential for analyzing multilingual corpora (e.g., the Comparative Party Manifesto and European Parliament Speeches). For instance, an increasing line of research aims to scale the ideological positions of political parties in the world by comparing their manifestos. It is important to choose models like cross-encoders that can capture substantive similarity in policy stances in addition to exact policy emulation. Hopefully, the ability of cross-encoders to conduct semantic comparisons of short political texts will open a broader range of substantive research.

References

Anspach, Nicolas M. and Taylor N. Carlson. 2020. "What to Believe? Social Media Commentary and Belief in Misinformation." *Political Behavior* 42(3): 697–718.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting and David Wingate. 2023. "Out of One, Many: Using Language Models to Simulate Human Samples." *Political Analysis* 31(3): 337–51.

Baird, Vanessa A. and Amy Gangl. 2006. "Shattering the Myth of Legality: The Impact of the Media's Framing of Supreme Court Procedures on Perceptions of Fairness." *Political Psychology* 27(4): 597–614.

Bestvater, Samuel E. and Burt L Monroe. 2023. "Sentiment is Not Stance: Target-Aware Opinion Classification for Political Text Analysis." *Political Analysis* 31(2): 235–56.

Blumenau, Jack. 2021. "The Effects of Female Leadership on Women's Voice in Political Debate." *British Journal of Political Science* 51(2): 750–71.

Bryan, James D. and Jordan Tama. 2022. "The Prevalence of Bipartisanship in US Foreign Policy: An Analysis of Important Congressional Votes." *International Politics* 59(5): 874–97.

Busby, Joshua W. and Jonathan Monten. 2008. "Without Heirs? Assessing the Decline of Establishment Internationalism in US Foreign Policy." *Perspectives on Politics* 6(3): 451–72.

Canes-Wrone, Brandice, William G. Howell and David E. Lewis. 2008. "Toward a Broader Understanding of Presidential Power: A Reevaluation of the Two Presidencies Thesis." *The Journal of Politics* 70(1): 1–16.

Carlson, Taylor N. 2019. "Through the Grapevine: Informational Consequences of Interpersonal Political Communication." *American Political Science Review* 113(2): 325–39.

Carlson, Taylor N. 2024. *Through the Grapevine: Socially Transmitted Information and Distorted Democracy*. University of Chicago Press.

Casas, Andreu, Matthew J Denny and John Wilkerson. 2020. "More Effective than We Thought: Accounting for Legislative Hitchhikers Reveals a More Inclusive and Productive Lawmaking Process." *American Journal of Political Science* 64(1): 5–18.

Cer, Daniel, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio and Lucia Specia. 2017. "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Cross-lingual Focused Evaluation." *arXiv preprint arXiv:1708.00055* .

Chandrasekaran, Dhivya and Vijay Mago. 2021. "Evolution of Semantic Similarity—A Survey." *ACM Computing Surveys (CSUR)* 54(2): 1–37.

Chong, Dennis and James N. Druckman. 2007. "A Theory of Framing and Opinion Formation in Competitive Elite Environments." *Journal of Communication* 57(1): 99–118.

Clark, Kevin, Minh-Thang Luong, Quoc V Le and Christopher D Manning. 2020. "Electra: Pre-training Text Encoders as Discriminators Rather Than Generators." *arXiv preprint arXiv:2003.10555* .

Clawson, Rosalee A and Eric N Waltenburg. 2003. "Support for a Supreme Court Affirmative Action Decision: A Story in Black and White." *American Politics Research* 31(3): 251–79.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805* .

Druckman, James N., Cari Lynn Hennessy, Kristi St. Charles and Jonathan Webber. 2010. "Competing Rhetoric Over Time: Frames Versus Cues." *The Journal of Politics* 72(1): 136–48.

Druckman, James N., Erik Peterson and Rune Slothuus. 2013. "How Elite Partisan Polarization Affects Public Opinion Formation." *American Political Science Review* 107(1): 57–79.

Düpont, Nils and Martin Rachuj. 2022. "The Ties That Bind: Text Similarities and Conditional Diffusion among Parties." *British Journal of Political Science* 52(2): 613–30.

Garí Soler, Aina and Marianna Apidianaki. 2021. "Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses." *Transactions of the Association for Computational Linguistics* 9: 825–44.

Gava, Roy, Julien M. Jaquet and Pascal Sciarini. 2021. "Legislating or Rubber-Stamping? Assessing Parliament's Influence on Law-Making with Text Reuse." *European Journal of Political Research* 60(1): 175–98.

Giavazzi, Francesco, Felix Iglhaut, Giacomo Lemoli and Gaia Rubera. 2023. "Terrorist Attacks, Cultural Incidents, and the Vote for Radical Parties: Analyzing Text from Twitter." *American Journal of Political Science* .

Hager, Anselm and Hanno Hilbig. 2020. "Does Public Opinion Affect Political Speech?" *American Journal of Political Science* 64(4): 921–37.

Hansen, Eric R. and Joshua M. Jansa. 2021. "Complexity, Resources and Text Borrowing in State Legislatures." *Journal of Public Policy* 41(4): 752–75.

Heaney, Michael T. and Fabio Rojas. 2015. *Party in the Street: The Antiwar Movement and the Democratic Party after 9/11*. Cambridge University Press.

Hinkle, Rachael K. 2015. "Into the Words: Using Statutory Text to Explore the Impact of Federal Courts on State Policy Diffusion." *American Journal of Political Science* 59(4): 1002–21.

Hitt, Matthew P. and Kathleen Searles. 2018. "Media Coverage and Public Approval of the US Supreme Court." *Political Communication* 35(4): 566–86.

Jeong, Gyung-Ho and Paul J. Quirk. 2019. "Division at the Water's Edge: The Polarization of Foreign Policy." *American Politics Research* 47(1): 58–87.

Kertzer, Joshua D., Deborah Jordan Brooks and Stephen G. Brooks. 2021. "Do Partisan Types Stop at the Water's Edge?" *The Journal of Politics* 83(4): 1764–82.

Kupchan, Charles A. and Peter L. Trubowitz. 2007. "Dead Center: The Demise of Liberal Internationalism in the United States." *International Security* 32(2): 7–44.

Le, Quoc and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *International Conference on Machine Learning*. PMLR pp. 1188–96.

Licht, Hauke. 2023. "Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings." *Political Analysis* 31(3): 366–79.

Lin, Gechun and Christopher Lucas. 2023. An Introduction to Neural Networks for the Social Sciences. In *The Oxford Handbook of Engaged Methodological Pluralism in Political Science*. Oxford University Press.

Linder, Fridolin, Bruce Desmarais, Matthew Burgess and Eugenia Giraudy. 2020. "Text as Policy: Measuring Policy Similarity through Bill Text Reuse." *Policy Studies Journal* 48(2): 546–74.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer and Veselin Stoyanov. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." *arXiv preprint arXiv:1907.11692* .

Mikolov, Tomas, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv preprint arXiv:1301.3781* .

Milner, Helen V and Dustin Tingley. 2015. *Sailing the Water's Edge: The Domestic Politics of American Foreign Policy*. Princeton University Press.

Myrick, Rachel. 2021. "Do External Threats Unite or Divide? Security Crises, Rivalries, and Polarization in American Foreign Policy." *International Organization* 75(4): 921–58.

Nicholson, Stephen P. and Robert M. Howard. 2003. "Framing Support for the Supreme Court in the Aftermath of Bush v. Gore." *The Journal of Politics* 65(3): 676–95.

Pennington, Jeffrey, Richard Socher and Christopher D. Manning. 2014. Glove: Global Vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. EMNLP pp. 1532–43.

Reimers, Nils and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Betworks." *arXiv preprint arXiv:1908.10084* .

Rheault, Ludovic, Kaspar Beelen, Christopher Cochrane and Graeme Hirst. 2016. "Measuring Emotion in Parliamentary Debates with Automated Textual Analysis." *PLOS ONE* 11(12): e0168843.

Rodman, Emma. 2020. "A Timely Intervention: Tracking the Changing Meanings of Political Concepts with Word Vectors." *Political Analysis* 28(1): 87–111.

Rosa, Guilherme, Luiz Bonifacio, Vitor Jeronymo, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo and Rodrigo Nogueira. 2022. "In Defense if Cross-Encoders for Zero-Shot Retrieval." *arXiv preprint arXiv:2212.06121* .

Rumelhart, David E, Geoffrey E Hinton and Ronald J Williams. 1986. "Learning Representations by Back-propagating Errors." *Nature* 323(6088): 533–36.

Sanh, Victor, Lysandre Debut, Julien Chaumond and Thomas Wolf. 2019. "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter." *arXiv preprint arXiv:1910.01108* .

Staudemeyer, Ralf C. and Eric Rothstein Morris. 2019. "Understanding LSTM–A Tutorial into Long Short-Term Memory Recurrent Neural Networks." *arXiv preprint arXiv:1909.09586* .

Trubowitz, Peter and Nicole Mellow. 2011. "Foreign Policy, Bipartisanship and the Paradox of Post-September 11 America." *International Politics* 48: 164–87.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin. 2017. "Attention is All You Need." *Advances in Neural Information Processing Systems* 30.

Wankmüller, Sandra. 2019. "Introduction to Neural Transfer Learning with Transformers for Social Science Text Analysis." *Sociological Methods & Research* p. 00491241221134527.

Waterman, Michael S., Temple F. Smith and William A. Beyer. 1976. "Some Biological Sequence Metrics." *Advances in Mathematics* 20(3): 367–87.

Widmann, Tobias and Maximilian Wich. 2023. "Creating and Comparing Dictionary, Word Embedding, and Transformer-Based Models to Measure Discrete Emotions in German Political Text." *Political Analysis* 31(4): 626–41.

Wilkerson, John, David Smith and Nicholas Stramp. 2015. "Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach." *American Journal of Political Science* 59(4): 943–56.

Ying, Luwei, Jacob M. Montgomery and Brandon M. Stewart. 2022. "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures." *Political Analysis* 30(4): 570–89.

Ziegler, Jeffrey. 2022. "A Text-As-Data Approach for Using Open-Ended Responses as Manipulation Checks." *Political Analysis* 30(2): 289–97.