Supplementary Information

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

Contents

A Python Implementation of Models

A.1 Off-the-Shelf Cross-Encoders

It is very simple and straightforward to apply off-the-shelf cross-encoders to measure the similarity of texts. The Python codes are provided below:

```python
from itertools import combinations
import pandas as pd
from sentence_transformers import CrossEncoder

# Step 1: IMPORT THE DATA
data = pd.read_csv("data.csv", encoding= "unicode_escape")
text = data["post_text"] # Get the column of text

# Step 2: PREPARE THE DATA FOR INPUTS
# Make pairs of texts for comparisons
pairs = list(combinations(text , 2))

# Step 3: APPLY AN OFF-THE-SHELF MODEL
# Download the model
model = CrossEncoder("cross-encoder/stsb-roberta-base")
# Predict the similarity
scores = model.predict(pairs)

# Step 4: STORE THE RESULTS
# Create a new dataframe
data_pair = pd.DataFrame(pairs , columns=["Text1", "Text2"])
# Add a new column of similarity scores
data_pair["scores"] = scores
```

A.2 Customized Cross-Encoders

Customized cross-encoders are supervised models. In this paper, I propose a k-fold cross-validation to create an ensemble model of cross-encoders that achieve the best performance on each validation set and apply this ensemble model to make predictions for the similarity of all pairs of news headlines in the second application. I visualize the workflow in Figure S1.

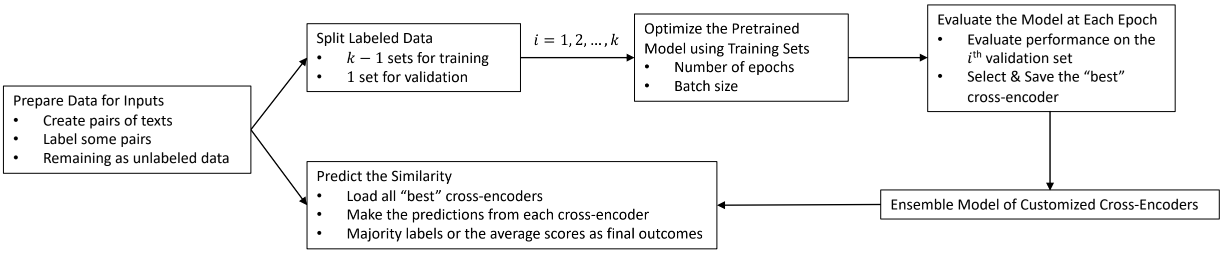

Prepare Data for Inputs
• Create pairs of texts
• Label some pairs
• Remaining as unlabeled data
Split Labeled Data
• $k-1$ sets for training
• 1 set for validation
$i = 1, 2, \ldots, k$
Optimize the Pretrained Model using Training Sets
• Number of epochs
• Batch size
Evaluate the Model at Each Epoch
• Evaluate performance on the i^{th} validation set
• Select & Save the "best" cross-encoder
Ensemble Model of Customized Cross-Encoders
Predict the Similarity
• Load all "best" cross-encoders
• Make the predictions from each cross-encoder
• Majority labels or the average scores as final outcomes

Figure S1: Workflow of Training Customized Cross-Encoders

The Python codes are provided below:

```python
import pandas as pd
import torch
from transformers import *
from torch.utils.data import DataLoader
from sentence_transformers import InputExample
from sentence_transformers.cross_encoder import CrossEncoder
from sentence_transformers.cross_encoder.evaluation import CESoftmaxAccuracyEvaluator

# CHOOSE PRETRAINED MODEL AND HYPERPARAMETERS
model_name = "roberta-base"
train_batch_size = 32
num_epochs = 10
k = 5

# IMPORT THE DATA (each row contains a pair of texts and their label)
labeled_set = pd.read_csv("labeled_data.csv")
# Randomly split to k folds
labeled_set = labeled_set.sample(frac=1, random_state=1)
N = round(len(labeled_set)/k)
labeled_set["fold"] = [1]*N + [2]*N + [3]*N + [4]*N + [5]*(len(labeled_set)-N*(k-1))
# Create a function to prepare data for inputs
def make_samples(df, test=False):
    samples = []
    for anchor, target, score in labeled_set[["Sen1", "Sen2", "Label"]].values:
        samples.append(
            InputExample(texts=[anchor, target], label=score),
        )
    return samples

# TRAIN THE MODEL
for i in range(1,k+1):
    print(i)
    train_set = labeled_set[labeled_set["fold"] != i]
    train_data = make_samples(train_set)
    validation_set = labeled_set[labeled_set["fold"] == i]
    validation_data = make_samples(test_set)
    train_dataloader = DataLoader(train_data, shuffle=True, batch_size=train_batch_size)
    evaluator_accuracy = CESoftmaxAccuracyEvaluator.from_input_examples(validation_data)
    model = CrossEncoder(model_name, num_labels=2)
    save_path = "checkpoint_fold"+ str(i)
    model.fit(train_dataloader=train_dataloader,
              evaluator=evaluator_accuracy,
              epochs=num_epochs,
              evaluation_steps=10000,
              warmup_steps=round(len(labeled_set)*0.05),
              output_path=save_path,
              save_best_model=True)
    model_best = CrossEncoder(save_path)
```

B Applications in Political Science

B.1 Application One: Information Distortion During Social Transmission

B.1.1 The Reuters Article

The whole Reuters article is as follows:

> The U.S. economy slowed less sharply in the first quarter than initially estimated due to unexpectedly higher consumer spending and a bigger jump in exports.
>
> Gross domestic product increased at a 1.4 percent annual rate instead of the 1.2 percent pace reported last month, the Commerce Department said in its final assessment on Thursday.
>
> It was still the slowest growth rate since the second quarter of last year. Economists polled by Reuters had expected GDP growth to remain unchanged at a 1.2 percent rate.
>
> GDP for the January-March period tends to underperform relative to the rest of the year due to perennial issues with the calculation of the data the government has said it is working to resolve.
>
> First-quarter economic growth was boosted by an upward revision to consumer spending, which accounts for more than two-thirds of U.S. economic activity. Consumer spending rose at a 1.1 percent rate instead of the previously reported 0.6 percent pace. It was still the slowest pace since the second quarter of 2013.
>
> Despite the upward revision, the Trump administration's stated target of swiftly boosting U.S. growth to 3 percent remains a challenge.
>
> A sustained average of 3 percent growth has not been seen since the 1990s. Since 2000, the U.S. economy has grown at an average 2 percent rate. The U.S. economy expanded 1.6 percent in 2016, the lowest rate in five years.
>
> President Donald Trump's economic program of tax cuts, regulatory rollbacks and infrastructure spending has yet to get off the ground five months into his presidency.
>
> Initial signs that economic growth re-accelerated sharply in the second quarter have also faltered with recent disappointing data on retail sales, manufacturing production and inflation. Housing data has also been mixed. The Atlanta Federal Reserve currently forecasts annualized GDP growth of 2.9 percent in the second quarter.
>
> Exports in the first quarter were revised to show a gain of 7.0 percent from the previously reported 5.8 percent.
>
> Business spending on equipment was revised to show it increasing at a 7.8 percent rate in the January-March period rather than the 7.2 percent previously estimated.
>
> Businesses accumulated inventories at a rate of $2.6 billion in the first quarter, rather than the $4.3 billion reported last month. Inventory

investment rose at a $49.6 billion rate in the fourth quarter of last year.

Inventories subtracted 1.11 percentage point from GDP growth instead of the 1.07 percentage point previously reported.

The government also reported that corporate profits after tax with inventory valuation and capital consumption adjustments fell at an annual rate of 2.7 percent in the first quarter after rising at a 2.3 percent pace in the prior three months.

B.1.2 Comparisons to Other Similarity Measures

In the Appendix of Carlson (2019), the author supplemented a new approach to measuring cosine similarity between social messages and the Reuters article using text2vec,[1] which is an R package that implements a word embedding model GloVe (Pennington and Manning, 2014). Table S1 shows that the text similarity measured by this word embedding approach is not significantly correlated to the hand-coded amount of message distortion.

In addition, I have locally fit a doc2vec model with the corpus of social messages. Then, I calculated cosine similarity between the social messages and the Reuters article using the document vectors generated by doc2vec. Table S1 indicates that the correlation is significant and negative, but is weaker than using cross-encoder.

Table S1: Associations between Message Distortion and Text Similarity

	Similarity between the Reuters Article and Social Messages		
	Cross-Encoder	Cosine (text2vec)	Cosine (doc2vec)
Intercept	0.405^{***}	0.525^{***}	0.900^{***}
	(0.010)	(0.012)	(0.005)
Message Distortion	-0.022^{***}	-0.007	-0.007^{*}
	(0.006)	(0.008)	(0.003)
Num. obs.	399	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].

[1] Note that Carlson (2019) called this approach doc2vec.

B.2 Application Two: Competing Headline Framing of SCOTUS Decisions

B.2.1 Data Collection

Through LexisNexis, I searched relevant news articles using the case names such as ALLEN v. COOPER. The search was restricted by the one-month window after SCOTUS rendered the decision of a specific case. To prevent the inclusion of irrelevant news articles,[2] I manually checked each document and delete them if deemed to be peripheral. Among the articles I collected, 395 news headlines are uninformative, while 979 news headlines contain substantive information about the cases.[3] Since uninformative news headlines do not highlight the content of case decisions, I excluded them from the similarity analysis.

B.2.2 MTurk Workers' Screening

Reliability is always the main concern when using crowdsourced workers to label data. Following the instructions of Ying, Montgomery and Stewart (2022), I designed a training module to educate workers about the task structure and coding rules. It requires workers hired from the Amazon Mechanical Turk to read five example HITs with answers and discussion and finish the test HITs. In order to be granted qualification, they have to answer at least 9 out of 10 questions correctly.

Training Module The following part presents the training module used to screen MTurkers. The training module contains five example HITs and ten test HITs. The answers and discussion of example HITs are viewable for MTurk workers who take the qualification test, while the answers of test HITs are not available during the real test.

> **Example 1:** Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
> Text 1: Alliance for Justice: Supreme Court Threatens Students and First Amendment
> Text 2: IJ Releases New Educational Choice Guide To State Constitutions After Espinoza
> Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.
> ○ Yes
> ○ No

[2] For example, the case name may appear at the end of an article where contains a long list of links directing to other articles. Some news articles are of advertisement nature celebrating particular lawyers won the case. News articles like this are inevitable to be collected when using the case names as search terms.

[3] Note that totally uninformative headlines, such as "The Supreme Court—April 20, 2020", "Daily Media Links 7/28", and "Supreme Court Decides Chiafalo v. Washington", cannot be coded based on my definition of similarity, as it is unlikely to infer what frames are used without context.

Answer: The correct answer is “No”. The meaning of one news headline is irrelevant to the other one. More specifically, knowing that “Supreme Court Threatens Students and First Amendment” cannot tell us that “IJ Releases New Educational Choice Guide To State Constitutions After Espinoza” and vice versa.

Example 2: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
Text 1: US Supreme Court rejects Trump administration’s Clean Water Act interpretation
Text 2: Hardly Ever? Permitting of Indirect Discharges Under the Clean Water Act After County of Maui, Hawaii v. Hawaii Wildlife Fund
Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes
○ No

Answer: The correct answer is “No”. Without extra information about “Trump administration’s Clean Water Act interpretation”, we cannot know whether rejecting it means “permitting indirect discharges”. Therefore, knowing “US Supreme Court rejects Trump administration’s Clean Water Act interpretation” cannot infer that “Hardly Ever? Permitting of Indirect Discharges Under the Clean Water Act After County of Maui, Hawaii v. Hawaii Wildlife Fund” and vice versa.

Example 3: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
Text 1: Protecting Pride: Supreme Court Holds Title VII Prohibits Workplace Discrimination On The Basis Of Sexual Orientation And Gender Identity
Text 2: Supreme Court issues landmark Title VII ruling protecting sexual orientation and gender identity
Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes
○ No

Answer: The correct answer is “Yes”. The meaning of one news headline is basically identical to the other one. They both refer to the Court’s decision regarding Title VII which prohibits discrimination on sexual orientation and gender identity, in other words, protecting them.

Example 4: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
Text 1: Supreme Court allows punitive damages award against Sudan for 1998 embassy bombings
Text 2: The Potential Impact of Terrorism Lawsuits Under the Antiterrorism Act on Ordinary Corporate, Banking and Sovereign Enterprises

Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes

○ No

Answer: The correct answer is “No”. Knowing that “Supreme Court allows punitive damages award against Sudan for 1998 embassy bombings” cannot infer “The Potential Impact of Terrorism Lawsuits Under the Antiterrorism Act on Ordinary Corporate, Banking and Sovereign Enterprises”. Also, the information given in Text 2 is not enough to conclude that the Court “allows punitive damages award against Sudan for 1998 embassy bombings.”

Example 5: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
Text 1: In Key Ruling for Public Access, SCOTUS Says No Copyright In Georgia Code Annotations
Text 2: Supreme Court Expands Penumbra of Gov't Edicts Doctrine: Official Annotations to Code Not Copyrightable
Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes

○ No

Answer: The correct answer is “Yes”. Although the meaning of these two news headlines is not equivalent, one can be used to infer the other. More specifically, Text 2 tells us that the Court holds that “Official Annotations to Code” are not copyrightable, which implies that no copyright is granted for “Georgia Code Annotations” too, since “Georgia Code Annotations” are a subset of “Official Annotations to Code”.

Test 1: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
Text 1: Foxx: Educational Choice is Powerful
Text 2: School Choice Champions Earn Major Victory in SCOTUS' Espinoza Decision
Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes

○ No

Answer: [Not Available in the Real Test] “No”

Test 2: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
Text 1: SCOTUS Upholds CFPB but not its Singular Director Structure
Text 2: Supreme Court Divided on Trump's Power to Fire Head of Consumer Bureau

Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes
○ No
Answer: [Not Available in the Real Test] "No"

Test 3: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
Text 1: The Supreme Court Holds That a Showing of Willfulness is Not a Precondition to Recover Profits for Federal Trademark Infringement
Text 2: U.S. Supreme Court Rules that Profits Available Even from Non-Willful Trademark Infringers
Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes
○ No
Answer: [Not Available in the Real Test] "Yes"

Test 4: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
Text 1: Supreme Court Clarifies Race Discrimination Claims Under 42 U.S.C. 1981 Must Meet More Stringent 'But-For' Causation Standard
Text 2: Supreme Court confirms race discrimination claims under section 1981 require 'but-for' causation
Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes
○ No
Answer: [Not Available in the Real Test] "Yes"

Test 5: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
Text 1: Lori Loughlin Should Test Whether Barr's DOJ Will Give Her the Same Treatment as Michael Flynn
Text 2: Supreme Court 'Bridgegate' Ruling Is Great News for Lori Loughlin
Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes
○ No
Answer: [Not Available in the Real Test] "No"

Test 6: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:

Text 1: Is John Roberts a Judicial Minimalist, a Coward, or a Strategic Maximizer?
Text 2: Supreme Court Rules for School Choice, Religious Liberty, Cites Okla. A.G. Hunter Brief
Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes
○ No

Answer: [Not Available in the Real Test] "No"

Test 7: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
Text 1: Supreme Court Strikes Down Government-debt Exception to TCPA Ban on Autodialed and Prerecorded Calls to Cell Phones
Text 2: TCPA Class Actions Supreme Court Severs Government Debt Collection Exception
Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes
○ No

Answer: [Not Available in the Real Test] "Yes"

Test 8: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
Text 1: A.G. Hill: U.S. Supreme Court Decision on Unauthorized Aliens, ID Theft is Win for Indiana
Text 2: Justice Breyer: Conservative Majority's Decision in Immigration Case Created 'Colossal Loophole'
Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes
○ No

Answer: [Not Available in the Real Test] "No"

Test 9: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
Text 1: Supreme Court Overturns Rule that Only Signatories Can Compel International Arbitration
Text 2: U.S. Supreme Court Holds That New York Convention Does Not Bar Nonsignatory From Compelling International Arbitration
Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes
○ No

Answer: [Not Available in the Real Test] “Yes”

Test 10: Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:
Text 1: Project on Government Oversight: SCOTUS Decision Spells Uncertainty for Congressional Oversight of a President
Text 2: Supreme Court denies Congress access to Trump finance records but rules Manhattan DA can pursue them
Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

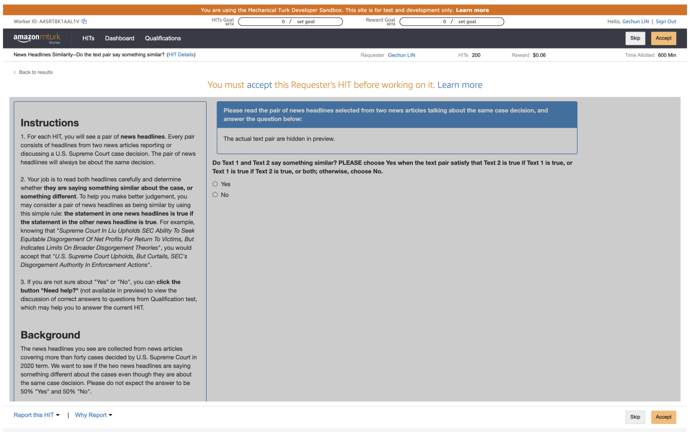
○ Yes
○ No

Answer: [Not Available in the Real Test] “No”

Only workers who finish the training module and pass the test HITs are qualified to do the paid HITs later. For each HIT, the workers viewed a pair of news headlines. Every pair consists of headlines from two news articles discussing a U.S. Supreme Court case decision. The pair of news headlines are always about the same decision. I asked the workers to code whether the pair of news headlines are saying something similar about the case, or something different. I used the entailment relationship as the rule to code the semantic similarity of news headlines while explained it in plain language: a pair of news headlines would be considered as being similar if the statement in one news headline is true given that the statement in the other one is true. Once accepting the tasks, on the left side of the screen, the worker sees the instructions, background, and attention about the tasks; on the right side, one pair of news headlines is followed by a question asking about the similarity with two options, “Yes” or “No”. If confused, the worker can click the “Need help?” button to review the answers and discussion of example HITs and test HITs that are from the training module. All of these designs are to improve workers’ performance as much as possible.

HITs From Workers’ View Before accepting the tasks, the workers can only preview the HIT in which the text pair are hidden. There is also no “Need help?” button on the preview version. Figure S2 shows the screenshot of preview HIT. Once the workers accept the tasks, their view of each HIT is as presented in Figure S3 without clicking the “Need help?” button and Figure S4 with clicking the “Need help?” button. The workers can scroll down to view all the instructions, background, and attention, as well as answers and discussion of example HITs and test HITs from the training module.

Evaluating Workers Using data labeled by crowdsourced coders has two potential threats for machine learning models. First, the quality of their work is likely to be varying. For example, workers may randomly make mistakes in coding similar pairs as not similar or dissimilar pairs as similar. Second, there may be inconsistency among workers’ coding rules. Some workers might operate similarity strictly such that they tend to code more data as dissimilar, while some workers might behave in the opposite direction. Although the MTurk workers I recruited all passed the quality screening described above,



You are using the Mechanical Turk Developer Sandbox. This site is for test and development only. **Learn more**

Worker ID: A4SRT8K1AAL1V

News Headlines Similarity--Do the text pair say something similar? (HIT Details)

Requester Gechun LIN HITs 200 Reward $0.06 Time Allotted 600 Min

‹ Back to results

You must accept this Requester's HIT before working on it. Learn more

Instructions

1. For each HIT, you will see a pair of **news headlines**. Every pair consists of headlines from two news articles reporting or discussing a U.S. Supreme Court case decision. The pair of news headliines will always be about the same decision.

2. Your job is to read both headlines carefully and determine whether **they are saying something similar about the case, or something different**. To help you make better judgement, you may consider a pair of news headlines as being similar by using this simple rule: **the statement in one news headlines is true if the statement in the other news headline is true**. For example, knowing that "*Supreme Court In Liu Upholds SEC Ability To Seek Equitable Disgorgement Of Net Profits For Return To Victims, But Indicates Limits On Broader Disgorgement Theories*", you would accept that "*U.S. Supreme Court Upholds, But Curtails, SEC's Disgorgement Authority In Enforcement Actions*".

3. If you are not sure about "Yes" or "No", you can **click the button "Need help?"** (not available in preview) to view the discussion of correct answers to questions from Qualification test, which may help you to answer the current HIT.

Background

The news headlines you see are collected from news articles covering more than forty cases decided by U.S. Supreme Court in 2020 term. We want to see if the two news headlines are saying something different about the cases even though they are about the same case decision. Please do not expect the answer to be 50% "Yes" and 50% "No".

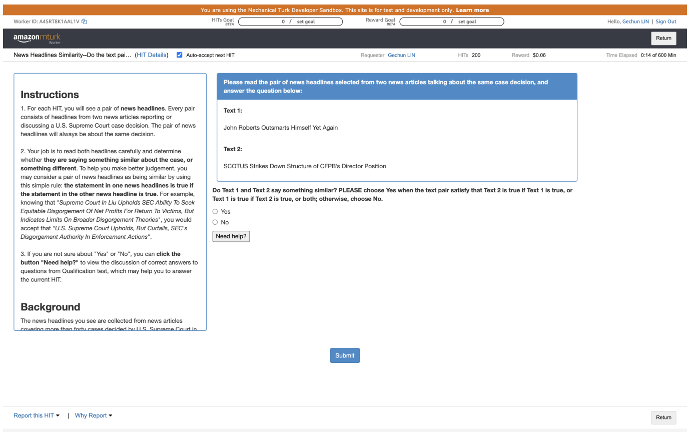
Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:

The actual text pair are hidden in preview.

Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes
○ No

Figure S2: Screenshot of Preview HIT



You are using the Mechanical Turk Developer Sandbox. This site is for test and development only. **Learn more**

Worker ID: A4SRT8K1AAL1V

News Headlines Similarity--Do the text pai… (HIT Details) ☑ Auto-accept next HIT

Requester Gechun LIN HITs 200 Reward $0.06 Time Elapsed 0:14 of 600 Min

Instructions

1. For each HIT, you will see a pair of **news headlines**. Every pair consists of headlines from two news articles reporting or discussing a U.S. Supreme Court case decision. The pair of news headliines will always be about the same decision.

2. Your job is to read both headlines carefully and determine whether **they are saying something similar about the case, or something different**. To help you make better judgement, you may consider a pair of news headlines as being similar by using this simple rule: **the statement in one news headlines is true if the statement in the other news headline is true**. For example, knowing that "*Supreme Court In Liu Upholds SEC Ability To Seek Equitable Disgorgement Of Net Profits For Return To Victims, But Indicates Limits On Broader Disgorgement Theories*", you would accept that "*U.S. Supreme Court Upholds, But Curtails, SEC's Disgorgement Authority In Enforcement Actions*".

3. If you are not sure about "Yes" or "No", you can **click the button "Need help?"** to view the discussion of correct answers to questions from Qualification test, which may help you to answer the current HIT.

Background

The news headlines you see are collected from news articles covering more than forty cases decided by U.S. Supreme Court in

Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:

Text 1:

John Roberts Outsmarts Himself Yet Again

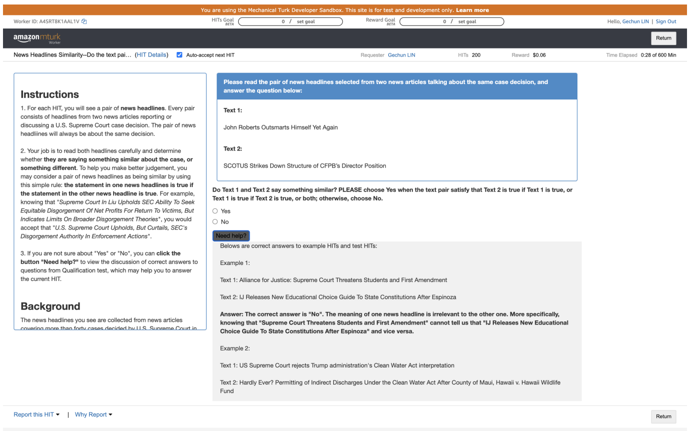
Text 2:

SCOTUS Strikes Down Structure of CFPB's Director Position

Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes
○ No

Figure S3: Screenshot of HIT (without clicking the help button)



You are using the Mechanical Turk Developer Sandbox. This site is for test and development only. **Learn more**

Worker ID: A4SRT8K1AAL1V

News Headlines Similarity--Do the text pai… (HIT Details) ☑ Auto-accept next HIT

Requester Gechun LIN HITs 200 Reward $0.06 Time Elapsed 0:28 of 600 Min

Instructions

1. For each HIT, you will see a pair of **news headlines**. Every pair consists of headlines from two news articles reporting or discussing a U.S. Supreme Court case decision. The pair of news headliines will always be about the same decision.

2. Your job is to read both headlines carefully and determine whether **they are saying something similar about the case, or something different**. To help you make better judgement, you may consider a pair of news headlines as being similar by using this simple rule: **the statement in one news headlines is true if the statement in the other news headline is true**. For example, knowing that "*Supreme Court In Liu Upholds SEC Ability To Seek Equitable Disgorgement Of Net Profits For Return To Victims, But Indicates Limits On Broader Disgorgement Theories*", you would accept that "*U.S. Supreme Court Upholds, But Curtails, SEC's Disgorgement Authority In Enforcement Actions*".

3. If you are not sure about "Yes" or "No", you can **click the button "Need help?"** to view the discussion of correct answers to questions from Qualification test, which may help you to answer the current HIT.

Background

The news headlines you see are collected from news articles covering more than forty cases decided by U.S. Supreme Court in

Please read the pair of news headlines selected from two news articles talking about the same case decision, and answer the question below:

Text 1:

John Roberts Outsmarts Himself Yet Again

Text 2:

SCOTUS Strikes Down Structure of CFPB's Director Position

Do Text 1 and Text 2 say something similar? PLEASE choose Yes when the text pair satisfy that Text 2 is true if Text 1 is true, or Text 1 is true if Text 2 is true, or both; otherwise, choose No.

○ Yes
○ No

Belows are correct answers to example HITs and test HITs:

Example 1:

Text 1: Alliance for Justice: Supreme Court Threatens Students and First Amendment

Text 2: IJ Releases New Educational Choice Guide To State Constitutions After Espinoza

Answer: The correct answer is "No". The meaning of one news headline is irrelevant to the other one. More specifically, knowing that "Supreme Court Threatens Students and First Amendment" cannot tell us that "IJ Releases New Educational Choice Guide To State Constitutions After Espinoza" and vice versa.

Example 2:

Text 1: US Supreme Court rejects Trump administration's Clean Water Act interpretation

Text 2: Hardly Ever? Permitting of Indirect Discharges Under the Clean Water Act After County of Maui, Hawaii v. Hawaii Wildlife Fund

Figure S4: Screenshot of HIT

unsystematic and systematic biases both occurred, which is documented in Table S2 showing the comparison results of workers' answers and expert-coded labels. As the workers' performances are not ideal, I decided not to train cross-encoder models with data labeled by them. Instead, I trained with expert-coded data and compare the model predictions to crowdsourced coders' work.

Table S2: Comparison of Workers' Answers and Expert-Coded Labels

Worker ID[1]	***T6U4		***JM6V		***N28F		***SQXP		***H400		***IIFM		***1LSR		***35HG	
Answer / Label	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Y	12	12	31	3	22	6	6	10	25	8	6	1	2	0	0	0
N	16	51	4	66	22	48	2	70	11	50	0	8	1	6	0	1
Accuracy Rate	0.70		0.93		0.71		0.86		0.80		0.93		0.89		1.00	
Mistake Type[2]	I		–		II$_Y$		II$_N$		I		–		–		–	

Note 1. Workers who answered too few HITs are excluded from the table because there are not enough expert-coded HITs mixed in their tasks for evaluation.
Note 2. "I" means the worker made random mistakes, which would create unsystematic bias; "II$_Y$" means the worker tended to code dissimilar pairs as similar, which would create systematic bias; "II$_N$" means the worker tended to code similar pairs as not similar, which would create systematic bias; "–" means the worker made few mistakes.

B.2.3 Computational Complexity and Scalability of Training Cross-Encoders

In total, there are 1,022 pairs of labeled news headlines. I gradually increased the training size from 10% to 80% when I trained customized cross-encoders. Table S3 contains different metrics to access the performance of each model in the test set achieved in the best epoch, as well as the training time. When the training data are too limited (about 100 pairs), one might fail to train the model. In this specific case, the customized cross-encoder has predicted all pairs in the test set to be "dissimilar," which is the majority category. As the training size was double (about 200 pairs, which is less than 1% of the whole dataset of 27,407 pairs), the training began to be effective (the customized cross-encoder stopped predicting all pairs to be "dissimilar."). This phenomenon suggests that there is a critical threshold in data volume for the model to begin learning meaningful distinctions between "similar" and "dissimilar" pairs. With more training data, the model performance—accuracy rates and the F1 scores—all increase.

Time is an important factor to consider in the scalability of training cross-encoders. Each reported time (seconds) is calculated from the initiation of a pre-trained RoBERT base model until the completion of training a customized cross-encoder over 10 epochs using a 24GB memory GPU (NVIDIA RTX A5000). Note that increasing the training size eightfold (from 0.1 to 0.8) merely doubles the training time. The fact that training time grows much more slowly than the expansion of training size suggests that cross-encoders have the potential to be trained at a large scale.

Table S3: Model Performance and Cost Across Training Sizes

Training Size	Time	Accuracy	Precision	Recall	F1
10%	14.41s	0.69	NA	0.00	NA
20%	16.80s	0.86	0.80	0.75	0.77
30%	19.02s	0.88	0.83	0.78	0.80
40%	21.13s	0.90	0.91	0.76	0.83
50%	23.11s	0.90	0.85	0.85	0.85
60%	25.40s	0.90	0.81	0.92	0.86
70%	27.66s	0.91	0.90	0.82	0.86
80%	28.90s	0.91	0.82	0.94	0.88

B.2.4 Detailed Explanations of Variables

Heterogeneity of News Coverage

In this application, I create an ensemble model of multiple cross-encoders, which are the ones that achieve the highest accuracy rate in each of the hold-out validation set, resulting in five best models after conducting a 5-fold cross-validation. Each model predictes whether a pair of news headlines is "similar" with some probability. If the probability is greater than 0.5, the cross-encoder classifies that pair as "similar"; otherwise, "dissimilar." The label of each pair of news headlines is determined using a majority rule. I measure the heterogeneity of news coverage using the average probabilities of being "dissimilar" predicted by the majority models for each pair of news headlines.

The similarity of news headlines has also been estimated by a wide range of alternative approaches—SW local alignment and cosine similarity of different embedding models. All methods give continuous scores between 0 and 1, representing the degree of similarity. I subtract each value from 1 to represent the heterogeneity of news coverage.

Unanimity of Case Decisions

The U.S. legal system allows individual judges to disagree with the majority opinion of the court which gives rise to its judgment. Judges may either write concurring opinions because they agree with the case outcomes but not the reasoning in the majority opinions, or have dissenting opinions when disagreeing with both. I define non-unanimous decisions as only those having dissenting opinions for two reasons. First, concurring opinions are very rare with only two occurrences in 2020 (January to July). Second, concurring opinions are not strong signals against the majority decisions. They usually contain subtly different or additional reasons as the legal basis supporting the majority decisions. Accordingly, cases with published dissents were coded as 1, 0 otherwise.

Salience

Previous research has considered different ways of measuring case salience. The dominant approach is Epstein and Segal (2000), which proposes to code a binary indicator of case salience by whether the case appeared on the front page of the New York Times on the day after the Court announced its decision. However, this after-decision measurement could introduce post-treatment bias in estimating the effect of unanimous decisions on heterogeneity of media coverage. Instead, I follow Maltzman and Wahlbeck (1996), which determines the salience of a case by the number of amicus briefs filed for each case.[4] I took the log transformation of these numbers to obtain a normal distribution of amicus participation.

Issue Area

The Supreme Court database[5] identified the subject matter of controversy discussed in each case into fourteen areas: civil rights, criminal procedure, First Amendment, due process, privacy, attorneys' or governmental officials' fees or compensation, unions, economic activity, judicial power, federalism, interstate relation, federal taxation, miscellaneous, and private law. Certain issue areas are represented by only one or two cases in my data collection. Adding each of fourteen issue areas to the regressions is harmful for model identification as there are not enough variations in the main explanatory variable *unanimity*—all decisions were split within some issue areas. It is also unnecessary, since my purpose is to control the differences between cases determining fundamental rights and those do not. Therefore, I split them into two categories. Cases falling under the first four issue areas are coded as 1, 0 otherwise.

B.2.5 The Association between Case Factors and Media Coverage

Table S4 provides the results from OLS regressions. I regressed *heterogeneity* on *unanimity*, controlling both *salience* and *issue area*. Formally, the models are expressed as:

$$\textbf{Heterogeneity} = \gamma_0 + \gamma_1\textbf{Unanimity} + \gamma_2\textbf{Salience} + \gamma_3\textbf{Issue Area} + \varepsilon$$

Table S4: Effect of Unanimity on the Heterogeneity of News Coverage

	Heterogeneity of News Coverage						
	Cross-Encoder (Customized)	Cross-Encoder (Off-the-Shelf)	Cosine (BERT)	Cosine (SRoBERTa)	Cosine (doc2vec)	Cosine (BoW)	Alignment
Intercept	1.048^{*}	0.485	-0.046	0.324	-0.733^{***}	-0.179	-0.096
	(0.486)	(0.418)	(0.247)	(0.396)	(0.175)	(0.194)	(0.189)
Unanimity	-0.738^{**}	-0.582^{*}	-0.115	-0.389	0.105	-0.418	-0.372
	(0.270)	(0.294)	(0.153)	(0.281)	(0.165)	(0.237)	(0.194)
Salience	-0.315^{*}	-0.175	-0.017	-0.138	0.165^{***}	0.011	-0.002
	(0.137)	(0.118)	(0.071)	(0.112)	(0.047)	(0.043)	(0.046)
Issue Area	0.159	0.216	0.134	0.231	0.152	0.195	0.149
	(0.167)	(0.153)	(0.094)	(0.136)	(0.102)	(0.136)	(0.117)
R^2	0.052	0.030	0.004	0.019	0.019	0.024	0.016
Adj. R^2	0.052	0.030	0.003	0.019	0.018	0.023	0.016
Num. obs.	27407	27407	27407	27407	27407	27407	27407

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.
Standard errors are clustered at the case level.

Additionally, I conduct one-sided z-tests to see whether the effect of *unanimity* is larger when the outcomes are measured by cross-encoders than other approaches. Table S5 shows the test results.

[4] I collected the number of amicus briefs through docket search for each case on https://www.supremecourt.gov/.

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

Table S5: Differences of the Coefficients of Unanimity (Customized Cross-Encoders vs. other Models

	difference of estimations	difference of variances	Z score	percentile
Cross-Encoder (Off-the-Shelf)	-0.19	0.14	-0.51	30.62%
Cosine (pre-trained BERT)	-0.76	0.09	-2.55	0.54%
Cosine (SRoBERTa)	-0.40	0.14	-1.09	13.72%
Cosine (doc2vec)	-0.94	0.10	-3.05	0.11%
Cosine (BoW)	-0.37	0.12	-1.08	14.01%
Local Alignment	-0.42	0.10	-1.33	9.12%

B.2.6 Robustness Check

The group of fundamental rights undergoes changes and developments over time, but they are primarily found in the Constitution, Amendments, and federal statutes such as Civil Right Act. To ease the concern that the empirical finding depends on the decision to split issue areas to two categories, I alternatively code whether a decision is about fundamental rights based on the legal provisions considered in the case. The variable *law type* equals to 1 if the Supreme Court database classifies the legal basis of the case into Constitution, Constitutional Amendment, or federal statutes, and 0, otherwise. Replacing the *issue area* with *law type*, the regression results still show that only cross-encoder (customized) can uncover the relationship between the unanimity of decisions and the heterogeneity of news coverage (see Table S6).

Table S6: Effect of Unanimity on the Heterogeneity of News Coverage

	Heterogeneity of News Coverage						
	Cross-Encoder (Customized)	Cross-Encoder (Off-the-Shelf)	Cosine (BERT)	Cosine (SRoBERTa)	Cosine (doc2vec)	Cosine (BoW)	Alignment
(Intercept)	0.954^{*}	0.399	-0.134	0.302	-0.582^{**}	-0.198	-0.144
	(0.470)	(0.402)	(0.237)	(0.389)	(0.208)	(0.189)	(0.173)
Unanimity	-0.648^{*}	-0.494	-0.032	-0.354	-0.009	-0.388	-0.321
	(0.258)	(0.279)	(0.146)	(0.269)	(0.151)	(0.212)	(0.172)
Law Type	0.261	0.304	0.230^{*}	0.239	-0.046	0.201	0.195^{*}
	(0.183)	(0.160)	(0.094)	(0.154)	(0.121)	(0.108)	(0.089)
Salience	-0.318^{*}	-0.178	-0.020	-0.137	0.173^{**}	0.012	-0.003
	(0.138)	(0.118)	(0.070)	(0.113)	(0.053)	(0.045)	(0.046)
R^2	0.053	0.030	0.004	0.017	0.017	0.022	0.016
Adj. R^2	0.053	0.030	0.004	0.017	0.016	0.022	0.016
Num. obs.	27407	27407	27407	27407	27407	27407	27407

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.
Standard errors are clustered at the case level.

B.3 Application Three: Does political disagreement end at the water's edge?

B.3.1 Measuring Issue-Specific Elite Polarization

Regarding US foreign policy, the main source of data examined in extant scholarship on elite polarization includes congressional voting records and bill sponsorship. For instance, Kupchan and Trubowitz (2007), Busby and Monten (2008), Chaudoin, Milner and Tingley (2010), and Bryan and Tama (2022) calculate the percentages of bipartisan votes or legislation cosponsored by members of different parties to analyze the yearly patterns of bipartisanship in US Congress. Jeong and Quirk (2019) estimates the foreign policy preferences of senators (Jeong, 2018) using their roll-call votes, inferring the degree of polarization as the difference between the respective means of the ideal points of Democratic and Republican senators (McCarty, Poole and Rosenthal, 2016). However, two issues are associated with this behavior-based measurement of polarization in foreign policy: strategic voting (Clinton, 2012) and unidimensionality in scaling member preference (Aldrich, Montgomery and Sparks, 2014). And it is unclear whether they would result in an upward or downward bias in the estimation.

Therefore, I move to a text-based approach, which relies on the partisan differences in language usage to measure polarization. A recent study Myrick (2021) utilizes a supervised machine-learning model to predict the legislators' parties based on their congressional speech regarding foreign adversaries—higher predictive accuracy means greater polarization (Peterson and Spirling, 2018). This approach may work poorly if two parties simply use distinct words to raise different issues rather than having opposing views on the same topic.

To avoid this pitfall, I invent a different approach, which combines topic modeling with text similarity to measure issue-specific polarization. I rely on the topic proportions estimated by the Structural Topic Model (STM) (Roberts, Stewart and Airoldi, 2016) to determine the main theme of each document. Then, I estimate the similarity of every two texts belonging to the same topic. I argue that the similarity of post contents could serve as a proxy for ideological distance. My approach incorporates two important features of polarization—intergroup heterogeneity and intragroup homogeneity (Druckman, Peterson and Slothuus, 2013)—through comparing the similarity scores of contents published by politicians with different party affiliations (inter-party posts) and those published by politicians from the same party (intra-party posts). Intuitively, the combination of higher similarity of intra-party posts with lower similarity of inter-party posts on the same topic indicates more polarized views on that policy issue.

B.3.2 Interaction Effects Between Domain of Policy Issue and Post Type

The regression results are presented in Table S7. Formally, the models are expressed as:
$\textbf{Similarity} = \beta_0 + \beta_1\textbf{Inter-Party} + \beta_2\textbf{International} + \beta_3\textbf{Inter-Party} \times \textbf{International} + \varepsilon$

Additionally, I conduct two-sided z-tests to see whether the cross-encoder model estimates coefficients that are different from the cosine approaches. Table S8 compares the coefficient *Inter-Party*, showing that cross-encoder provides estimates that are statistically different from those of all alternatives. Table S9 compares the coefficient *Interna-*

tional, showing that cross-encoder provides estimates that are statistically different from those of BoW and BERT. Table S10 compares the coefficient *Inter-Party* × *International*, showing that cross-encoder provides estimates that are statistically different from those of doc2vec and BERT.

Table S7: Moderating Effects of Policy Issue on the Relationship between Post Type and Similarity of Content

	Similarity of Social Media Content			
	Cross-Encoder	Cosine (BoW)	Cosine (Doc2Vec)	Cosine (BERT)
Intercept	-0.001	0.056^{***}	-0.043^{***}	-0.007
	(0.011)	(0.013)	(0.013)	(0.011)
Inter-Party Posts	-0.245^{***}	-0.178^{***}	-0.093^{***}	-0.010
	(0.020)	(0.026)	(0.021)	(0.019)
International Policy	0.402^{***}	-0.045^{*}	0.407^{***}	0.049^{**}
	(0.022)	(0.020)	(0.020)	(0.017)
Interaction	0.142^{***}	0.143^{***}	0.026	0.023
	(0.033)	(0.034)	(0.030)	(0.027)
R^2	0.038	0.006	0.026	0.001
Adj. R^2	0.038	0.006	0.025	0.001
Num. obs.	100999	100999	100999	100999

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$.
Standard errors are clustered at both the senator pair and topic levels.

Table S8: Differences of the Coefficients of Inter-Party (Cross-Encoder vs. alternatives)

	difference of estimations	difference of variances	Z score	percentile
Cosine (BoW)	-0.07	0.00	-2.02	4.3%
Cosine (doc2vec)	-0.15	0.00	-5.33	0%
Cosine (BERT)	-0.23	0.00	-8.54	0%

Table S9: Differences of the Coefficients of International (Cross-Encoder vs. alternatives)

	difference of estimations	difference of variances	Z score	percentile
Cosine (BoW)	0.45	0.00	15.02	0%
Cosine (doc2vec)	-0.01	0.00	-0.17	86.93%
Cosine (BERT)	0.35	0.00	12.64	0%

Table S10: Differences of the Coefficients of Interaction (Cross-Encoder vs. alternatives)

	difference of estimations	difference of variances	Z score	percentile
Cosine (BoW)	-0.00	0.00	-0.02	98.45%
Cosine (doc2vec)	0.12	0.00	2.60	0.93%
Cosine (BERT)	0.12	0.00	2.83	0.47%

References

Aldrich, John H, Jacob M Montgomery and David B Sparks. 2014. "Polarization and Ideology: Partisan Sources of Low Dimensionality in Scaled Roll Call Analyses." *Political Analysis* 22(4): 435–56.

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.

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

Chaudoin, Stephen, Helen V Milner and Dustin H Tingley. 2010. "The Center Still Holds: Liberal Internationalism Survives." *International Security* 35(1): 75–94.

Clinton, Joshua D. 2012. "Using Roll Call Estimates to Test Models of Politics." *Annual Review of Political Science* 15: 79–99.

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.

Epstein, Lee and Jeffrey A Segal. 2000. "Measuring Issue Salience." *American Journal of Political Science* pp. 66–83.

Jeong, Gyung-Ho. 2018. "Measuring Foreign Policy Positions of Members of The Us Congress." *Political Science Research and Methods* 6(1): 181–196.

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.

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.

Maltzman, Forrest and Paul J Wahlbeck. 1996. "May it Please the Chief? Opinion Assignments in the Rehnquist Court." *American Journal of Political Science* pp. 421–443.

McCarty, Nolan, Keith T Poole and Howard Rosenthal. 2016. *Polarized America: The Dance of Ideology and Unequal Riches.* MIT 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.

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.

Peterson, Andrew and Arthur Spirling. 2018. "Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems." *Political Analysis* 26(1): 120–128.

Roberts, Margaret E, Brandon M Stewart and Edoardo M Airoldi. 2016. "A Model of Text for Experimentation in the Social Sciences." *Journal of the American Statistical Association* 111(515): 988–1003.

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.