Troubles in/with Text: Finetuning NLP to Analyze Declassified Rationalizations of Rights Violations

Sarah K. Dreier,* Sofia Serrano,† Emily K. Gade,‡ and Noah A. Smith§

July 14, 2020

Abstract

How do governments rationalize policies that violate the rights of their citizens? To answer this question, researchers might consult classified government correspondence documents that are later publicly released. Analyzing archival collections is a challenging task, requiring experts to systematically code distinct rationalization categories and then qualitatively, quantitatively, and/or computationally evaluate hypotheses. The challenges are compounded by the nature of the data, which in many cases is not only unstructured but also imperfectly digitized text. New advances in automation for text-related tasks, originating in the computing field of natural language processing (NLP), offer potential improvements for scaling qualitative analysis and dealing with "noisy" data. Here, we use a novel data source—recently declassified archives of the UK Prime Minister's correspondence during the "Troubles in Northern Ireland"—and assess the effectiveness of emerging techniques from NLP in identifying and categorizing rationalizations for human rights violations in that collection.

Introduction

In an effort to quell tensions in the early years of its "Troubles in Northern Ireland" (1969–1974), the British government interned 1,874 Northern Irish Catholics without trial. During times of real or perceived national insecurity, states often target certain citizens with violent repression or oppression, thereby selectively surrendering their commitments to human rights. However, liberal democracies must do so while claiming to maintain their legal commitments to protecting those living within their borders. Understanding how UK officials rationalized British internment policies to themselves and their colleagues provides critical insights about how a state constructs or maintains legal legitimacy while authorizing extralegal force (Section 1).

As a starting point for understanding how these policies were rationalized, we consider a novel collection of nearly 7,000 recently declassified documents from the British Prime Minister's security-related correspondence files between 1969–1974 (Section 2).[1] About 20 percent of these pages are directly relevant to internment. Our first contribution is to digitize and hand-code digitized versions

*Corresponding author: skdreier@uw.edu; *University of Washington*

†*University of Washington*

‡*Emory University*

§*University of Washington*

[1]The data were obtained manually from the UK National Archives at Kew (`https://www.nationalarchives.gov.uk/`).

of these documents at the sentence level for rationalizations for internment into twelve categories.[2] We expect that the categories and the accompanying data will inform future substantive research on government rationalization of rights violations.

Using this entirely hand-coded dataset, we explore the use of computational text-as-data methods to identify and classify government rationalizations. Computational extensions to such qualitative content analysis are not new in political science, which often analyzes concepts that are complex, specific, and understood in contrast to one another. Other examples include policy frames found in news media (Card et al., 2015) and rhetorical positioning in political speeches (Acree et al., 2020). A hallmark of these concepts is that they are cognitively identifiable to a human coder (with some training) but cannot be neatly tied to specific words or phrases. For this reason, we refer to these complex, specific ideas as 'know-when-I-see-it' (KWISI) concepts. Scholars face the challenging and laborious task of identifying and categorizing KWISI concepts within a wide range of of politically relevant text, including laws and treaties (Spirling, 2012), legislative debates, campaign rhetoric and constituent outreach, political news, citizens' political engagement, and archival records of classified government deliberation.

In the same way that computing enabled major advances in statistical analysis, methods from the computing field of natural language processing (NLP) offer promise for improving the efficiency, consistency, and scalability of text analysis in political science. Indeed, though "text as data" methods have successfully advanced many political science analyses (Grimmer and Stewart, 2013), we note that new developments in NLP research are neither driven by tasks relevant to political science (e.g., KWISI concept identification) nor routinely adopted by political science researchers. Notably, the recent shift to a "pretrain and finetune" paradigm for solving NLP tasks (Devlin et al., 2019; Peters et al., 2018; discussed in Section 3) is far more powerful than conventional keyword- and topic model-based methods and, we believe, particularly well-suited to modeling KWISI concepts in text. Our second contribution is to adapt this latest generation of NLP for the task of identifying and categorizing internment rationalizations in our "noisy" dataset (Section 4). We report on its performance, relative to a conventional text classification approach (Section 5).

This new approach is more accurate than the baseline at *categorizing* a piece of text known to contain a rationalization into one of twelve substantively distinct rationalization categories. However, *identifying* a rationalization within a much larger text corpus remains a serious challenge; several remaining limitations exist for future advances to address and overcome (Section 6). These include (1) coping with noise introduced by digitization (e.g., optical character recognition errors, handwritten comments, and the physical erosion of paper documents over time), for which the new methods we use here are unprepared; and (2) using a generic vocabulary that cannot be easily modified to suit the classification task within lexically idiosyncratic text.

1 Government Rationalizations for Violating Human Rights

Following the Easter Rising uprising across the Republic of Ireland (1916), Britain annexed six counties in Ireland's Ulster Province as Northern Ireland in 1920–21. The region's Protestant majority applauded this move, while most Irish Catholics—many of whom fought against British rule in the Irish War of Independence—opposed Ulster's annexation. By the late 1960s, Northern Ireland's Catholics had experienced decades of political, social, and economic discrimination. To redress this discrimination, they initiated civil rights protests, which precipitated repression from the state. In the Troubles in Northern Ireland that followed, unrest and violence from state actors, pro-government Protestant groups, and Republican groups gripped the province. The Provisional Irish Republican Army (PIRA) perpetrated violent terrorist attacks as part of their campaign to liberate Northern Ireland from the United Kingdom.

[2]This dataset, along with the original documents, will be publicly released in accordance with British Library Guidelines.

Meanwhile, loyalist militias used violence to support Northern Ireland's Protestant-ruling government and fealty to the British Crown. In August 1971, Britain authorized internment without trial to contain the political upheaval and terrorism violence. In violation of European Convention on Human Rights and British common law, Northern Ireland interned 1,874 Catholic citizens (and, eventually, 107 loyalists) without trial.

During the twentieth and twenty-first centuries, some of the world's oldest industrialized democracies—including in Europe and the United States—responded to real or perceived national security threats by surrendering their commitments to democratic legal processes and denying rights to targeted minorities within their borders. State actors routinely rationalize these violations as necessary to maintain national security during "emergencies" that threaten a state's sovereignty, yet these actions violate the principles often heralded as a democracy's *raison d'être* (Agamben, 2005; Campbell and Connelly, 2003; Davenport, 2007; Gross and Aoláin, 2006). How do state actors discuss and rationalize their decisions to authorize human rights violations, like internment without trial? How do they negotiate between their commitments to human rights and national security, and how do these rationalizations change over the course of a national security crisis?

In the case of Northern Ireland, British officials publicly announced that internment was necessary to quell terrorist violence. However, internal documents suggest that military officials believed internment was counter-indicated. Given these discrepancies, scholars must gain access to *internal* decision-making processes—the uncurated motivations and debates that occur outside the public spotlight—in order to understand how state actors rationalize their decisions to authorize human rights violations. The recently declassified government correspondence files we analyze here provide unique evidence of how government officials internally rationalize internment policies as those policies were be crafted, authorized, and amended.

Tracking state rationalizations for rights violations makes valuable downstream contributions. Understanding how state actors leverage and oscillate between various rationalizations for rights violations provides critical information about the boundaries between military objectives, political motivations, legally authorized violence, racialized attitudes, and rights-based protections—contours which shape contemporary state responses to immigration, incarceration, crime, racially motivated violence, protests against police brutality, and many other political issues. Developing computational NLP tools to identify rationalizations could help scholars pinpoint or even anticipate government human-rights violations. We *manually* identified and categorized British officials' rationalizations for internment without trial. However, this task is a candidate for *automation*, which, if executed successfully, could amplify scholars' abilities to track or anticipate government policies that violate citizens' rights.

2 Digitized Corpus from the UK Archive

Our collection of data comprises nearly 7,000 archived pages about Northern Ireland from the British Prime Minister's security-related correspondence files during the first five years of the Troubles (1969–1974; see Figure 1 for example pages). State archives can contain extensive, textured records of governance and political institutions over considerable periods of time (Katagiri and Min, 2019). They offer scholars a unique opportunity to examine states' internal decision-making processes and the rationalizations officials leverage to support and legitimate their decisions. Analyzing these records can provide historical context for contemporary policies and elucidate political dynamics that might generalize to politics today. The archive collection we use in this analysis includes reports, memos, typed and hand-written letters, telegraphs, meeting minutes, telephone-call transcriptions, and other documentation about Britain's planning and management of internment.

To obtain this material, we manually photographed every page in every file relevant to our scope

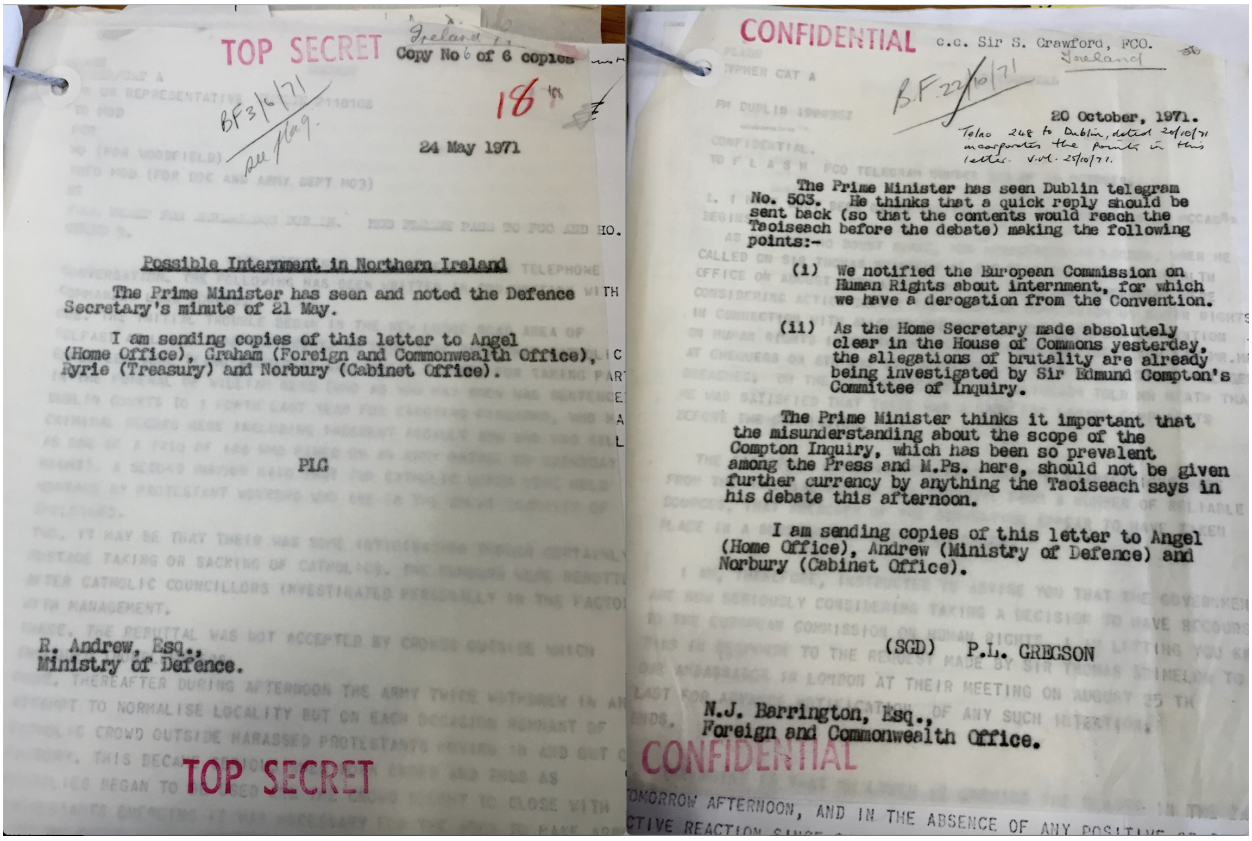


Figure 1: Digitized archive image files retrieved from the UK National Archives. **Left:** A top secret communique from the UK Ministry of Defense to the Prime Minister considering the use of internment (archive PREM 15 477, Image 5533). **Right:** Confidential communique considering whether the UK's internment policies violate European human rights law (archive PREM 15 485, Image 5222).

conditions.[3] We then adopted a multi-stage optical character recognition (OCR) process to digitize our photographs. First, we used Python's Tesseract OCR tool[4] to digitize the most legible text. However, our photographs contained considerable eccentricities, including: variation in the photography angle and technique; hand-written edits over typed text and notes in the page margins; variation in font, text size, ink darkness, and text clarity; and variation in page size, orientation, and color. These eccentricities reduced text legibility and significantly undermined the quality produced by our automated OCR process. We therefore paid an outside party to complete the OCR process and manually validate the documents (with the goal of achieving a maximum error rate of 2% at the character level). A small number of pages (approximately 200) were too illegible to digitize; in these cases, we manually transcribed any substantively relevant material. Most documents in our resulting digitized dataset contain a small amount of error, while some are considerably garbled and effectively computationally unusable.

Upon completing the OCR process, we manually identified roughly 1,400 pages directly relevant to internment. We then hand-coded and classified each sentence which contained a rationalization for internment into twelve categories, yielding roughly 2,200 rationalization occurring within the text. Each rationalization category (below) represents a distinct relationship between the state, its self-identified

[3] In addition to files from the Prime Minister's office (PREM), we included relevant files from the Ministry of Defense (DEFE) and records created or inherited by the Northern Ireland office (CJ).

[4] `https://pypi.org/project/pytesseract/`

purposes, its use of force, its accountability to existing policies or liberal human-rights standards, and its instrumental objectives. To ensure replicability in our coding scheme, we developed a deductive coding ontology, kept a detailed coding log, and iteratively amended the ontology as appropriate. All coding documentation will be made publicly available. The Appendix provides rationalization frequencies.
Rationalization categories:

- **Legal procedure**: State use of, and deferral to, existing legal policies;
- **Emergency policy**: State dismissal of "ordinary" law during exceptional times of emergency;
- **Terrorism**: State concerns about terrorist "enemies of the state";
- **Denial**: State denial that policies violate liberal legal principles or human rights protections;
- **Political**: State political or strategic military calculations that internment is advantageous;
- **Law and order**: State commitments to maintaining law and order;
- **Development and unity**: State assertions that economic development and social unity is undermined by conflict;
- **Utilitarian or deterrence**: State perceptions that internment will help respond to or reduce violence;
- **International or domestic precedent**: State reference to other precedents or examples of similar government behavior in other contexts;
- **Last resort**: State use of extreme policies as a last-case scenario;
- **Intelligence**: State suggestions that internment policies generated crucial intelligence;
- **Miscellaneous**: State use of another logic to rationalize internment.

Individual occurrences of rationalizations frequently contain overlapping or co-existing rationalization categories. Furthermore, they often use and represent words or broader ideas that overlap with surrounding text—text which does contain a rationalization. As such, these rationalization categories represent the KWISI concepts we seek to computationally identify. Although these annotations were designed to advance our substantive analysis of the universe of rationalizations for internment expressed in the archive text corpus, here they serve as "gold standard" ground truth examples of government rationalizations. We use this fully annotated corpus to train an automatic classifier and to evaluate its accuracy in dealing with KWISI political concepts.

3 Advances in Natural Language Processing

NLP methods have been applied extensively within political science. Seminal works have identified the promise of automated text-as-data methods for analyzing content and classifying text alongside the importance of retaining a role for human intuition for natural language (Grimmer and Stewart, 2013; Grimmer and King, 2011; Carlson and Montgomery, 2017; Eisenstein, Ahmed and Xing, 2011; see also: Bouchat, 2020). Researchers have used NLP and other approaches to automated text analysis to measure political conflict (Grimmer and Stewart, 2013), capture latent ideological expressions (Rheault and Cochrane, 2020), place legislatures on an ideological scale (Beauchamp, 2012), distinguish between government institutions (Nay, 2016), identify sentiments that accompany specific policy positions or

attitudes (Schoonvelde, Schumacher and Bakker, 2019), track political deliberation (Parthasarathy, Rao and Palaniswamy, 2019), understand political polarization (Bustikova et al., 2020; Peterson and Spirling, 2018), understand strategic parliamentary opposition (Dewan and Spirling, 2011), track U.S. legislative processes (Adler and Wilkerson, 2012; Collingwood and Wilkerson, 2012; Wilkerson and Casas, 2017), classify public attitudes toward the U.S. presidency (Hopkins and King, 2010), measure uncertainty in political texts (Bouchat, 2018), track media attention and frames (Boydstun, 2013; Card et al., 2015), infer international relations events and relationships (O'Connor, Stewart and Smith, 2013), analyze political text across languages (Lucas et al., 2015), identify government intent behind censorship policies (King, Pan and Roberts, 2013), analyze open-ended survey responses (Roberts et al., 2014), and quantify public political behavior on social media platforms.

Though a complete review of NLP methods for political science is beyond the scope of this research letter, a brief review of the key techniques elucidates the latest developments in automated text analysis. Political scientists have generally adopted two dominant approaches: supervised text classification and unsupervised modeling.

Supervised text classification assumes that a text collection can be segmented into more-or-less independent, identically distributed instances (usually documents or sentences), each of which is given a label (typically by a human coder) that corresponds with its content. Each text is mapped deterministically to a vector "representation" or encoding, and a statistical model is estimated that maps from encodings to labels. The simplest and most common example of this approach uses (1) a "bag of words" representation with each dimension of the vector capturing absence/presence of a vocabulary word (using 1 or 0, respectively) or the frequency of a vocabulary word; and (2) a logistic regression model for classification. Even this simple method requires a number of design decisions by the researcher, including designing the label set and selecting the vocabulary (e.g., all words observed, or a small set of predefined words relevant to the concept of interest), as well as some more opaque experimental choices (e.g., the strength of the regularization hyperparameter for training the model). More sophisticated approaches replace the "bag of words" with word embedding vectors constructed from large, unannotated corpora (discussed further below) and replace the logistic regression model with a nonlinear neural network; these advances require even more decisionmaking by the researcher and more hyperparameter selection, and obtain greater accuracy at a cost of intepretability.[5]

Unsupervised modeling is distinguished from the first family of approaches by its use of unannotated text data. Typically it involves a computationally intensive optimization procedure whose byproduct is new function for encoding an arbitrary text. The canonical example is probabilistic topic modeling (Blei, 2012; Blei, Ng and Jordan, 2003), which produces a mapping of a document to proportions of automatically-inferred "topics," each associated with a distribution of words. Often the text representations discovered by these methods are interpretable to humans and useful in generating hypotheses about a text collection. As with supervised methods, there are many design decisions left to an individual researcher, many of them opaque (e.g., number of topics, the vocabulary of words considered, or the optimization procedure).[6]

These approaches, and the many hybrids created from them, are not well suited to identify KWISI concepts. We have noted that it is difficult to tie concepts like "rationalization for internment" to specific words or phrases. Unsupervised models often "fail to yield topics of their substantive interest by inadvertently creating multiple topics with similar content and combining different themes into a single

[5]Feature selection is often detached from substantive theory but can highly influence model outputs (Denny and Spirling, 2018).

[6]We note that our characterization ignores a wide literature of hybrid approaches. For example, topic representations of documents can be used within a "downstream" text classifier; topic models can be extended to capture additional variables (e.g., Roberts et al., 2013), including labels (Mcauliffe and Blei, 2008), rendering them "supervised," and a wide range of "semisupervised" learning methods have also been explored (Gururangan et al., 2019).

topic" (Eshima, Imai and Sasaki, 2020), and they often necessitate priming by human experts or interactive procedures (Hu et al., 2014). Further, when we seek to *localize* a concept within a text (rather than characterizing an entire document), these methods require fragmenting the data into smaller segments that are not truly independent from each other. In summary, these models make strong assumptions about the organization of meaning in text and are insensitive to the way context influences meaning.

One notable improvement has come in the form of "word embeddings," also called "word vectors." Word embeddings are derived from a family of unsupervised procedures that take a large corpus of text and derive a mapping from vocabulary words to vectors (typically of a few hundred dimensions). These methods can be made very computationally efficient (Pennington, Socher and Manning, 2014). A desirable, and often observed, property for word vectors is that words with similar meanings will have "nearby" vectors (in Euclidean space, for example), allowing for improved statistical sharing when the vectors are applied within supervised modeling (e.g., by aggregating the vectors for words present in a document into a single document vector). The word vectors themselves can be "pretrained" on one corpus and then made available as standalone components, or they can be trained on a specific corpus to capture the idiosyncratic properties of the words as used in that corpus. For example, Rheault and Cochrane (2020) developed word embeddings based on parliamentary corpora from Britain, Canada, and the United States to produce externally validated scaling estimates of ideological placement and to capture latent dimensions of ideology. Similarly, Nay (2016) used word embeddings based on U.S. government text corpora to identify meaningful differences between government institutions and how they talk about specific policy areas.

The most recent advance in word embeddings recognizes that the way a human interprets a word depends on the context in which it is used. At a coarse level, words have different meanings that may be unrelated (e.g., *Amazon* refers to a river and a corporation), but this phenomenon exists at arbitrarily fine-grained levels as well (e.g., the concrete meaning of *tall* is very different in *tall woman* vs. *tall skyscraper*). Peters et al. (2018) introduced an unsupervised technique for creating *contextual* word embeddings. After training on an unannotated text corpus, the method maps each word in a text sequence to a vector that is influenced by the entire sequence. The method, known as ELMo,[7] was found to offer significant advantages to accuracy measured on a wide range of NLP benchmarks.

This method and its rapidly emerging stream of successors have led to a shift of focus across NLP. Whereas previous approaches focused on specialized supervised modeling for individual language processing tasks, the new paradigm assumes that contextual word vectors are first "pretrained" on an extremely large corpus (typically of web text), and then a relatively lightweight and generic supervised model is trained on labeled input-output pairs. Importantly, this second phase also updates the parameters in the contextual embedding function through a process known as "fine-tuning."

Because contextual embedding methods have been so successful in NLP, because they shift responsibility from expensive labeled data to cheap unlabeled data, and because, in principle, they are better suited to capturing complex, KWISI concepts expressed in text, we anticipate that they can be applied as an effective text-as-data method in political science.

4 Experiments

Our experiments seek to assess the viability of pretrained-and-finetuned NLP models for KWISI concept modeling. We consider two supervised tasks:

1. Given a rationalization, classify it into our twelve-code scheme;

[7] "Embeddings from language models," where "language model" refers to the unsupervised training objective maximized using unannotated text.

2. Given a sentence, classify it as containing a rationalization or not.

For both tasks, the dataset's text is lowercased and divided into a training set (80%) of the data, a development set for selecting hyperparameters (10%), and a test set for evaluation (10%). Multiple sentences on a given archive page are assigned to the same portion of the data to avoid exposing the model to any page-level idiosyncrasies. For each task, we compare two models: a baseline and an exemplar of the pretrain-and-finetune approach.

Baseline. We use a bag-of-words logistic regression model. The vocabulary for this model was selected using only the training data, splitting on whitespace and punctuation. We represent each sentence as a vector the size of the vocabulary, with each element of the vector being the count of how many times its corresponding vocabulary word appears in that (lowercased) sentence. An important hyperparameter in using logistic regression is the regularizer, an auxiliary objective used during parameter estimation to penalize extreme weights, which are associated with overfitting. We apply ℓ_2 regularization and select a hyperparameter strength from values $\{10^{-4}, 10^{-3}, \ldots, 10^{3}\}$, using the F_1 score on the development set as the criterion for selecting this strength.

Pretrain-and-finetune. We select RoBERTa (Liu et al., 2019) as an examplar of the new family of pretrain-and-finetune approaches. RoBERTa is derived from BERT (Devlin et al., 2019), a successor to ELMo (Peters et al., 2018). It was pretrained by researchers at Facebook on 160GB of English text from books, Wikipedia entries, and news and other web text (gathered via the Internet Archive's Common Crawl). The pretrained model is publicly available.[8] Importantly, when using a pretrained model, we inherit the vocabulary it was built with, which in this case consists of "wordpieces" representing common subword-to-word-length units that appeared in RoBERTa's training data. To build this vocabulary, small subword units were iteratively merged and added to an original vocabulary of characters until doing so stopped showing gains on language modeling using those subword units as a vocabulary. During tokenization using these wordpieces, the text's tokens are determined by following this same iterative merging process, with common text (from the perspective of RoBERTa's training data) being represented by fairly large subword units.

Without any further training on RoBERTa's original masked token prediction objective, we finetune RoBERTa on our training data to maximize the log-probability of the correct label, applying instance weighting to correct for class imbalances in the training data.[9] Our training approach uses the Adam optimizer (Kingma and Ba, 2014), stopping after either ten passes over the training set or until the F_1 score on the development set stops increasing (whichever comes first), in order to avoid overfitting. We select two hyperparameters by comparing F_1 score on the development set: the learning rate (from $\{0.00001, 0.00002, 0.00003\}$) and the batch size (16 or 32).[10]

[8] We use the base version of RoBERTa with 12 layers from `https://huggingface.co/transformers/pretrained_models.html`

[9] For example, on the second task, there are roughly 1,500 positive instances and more than 14,000 negative instances in the training data. We therefore weight the loss contributed by each positive instance roughly nine times as much as the loss contributed by each negative instance, to discourage the model from learning to take advantage of this class imbalance to classify everything as negative.

[10] The learning rate controls how much the model's parameters move in response to a calculated gradient. In the learning process, models split the data into smaller "batches" of data instances (here, sentences). For each batch, we perform a single update to the model's parameters. These different hyperparameter options are empirically chosen to maximize performance but do not correspond to any substantive judgments about the underlying task.

	Logistic regression			Fine-tuned RoBERTa		
	Precision	Recall	F_1	Precision	Recall	F_1
Task 1 "Terrorism"	0.489	0.564	0.524	0.610	0.641	0.625
Task 1 "Intl-Domestic Precedent"	1.000	1.000	1.000	1.000	1.000	1.000
Task 1 "Denial"	0.480	0.480	0.480	0.500	0.560	0.528
Task 1 "Political-Strategic"	0.583	0.778	0.667	1.000	0.778	0.875
Task 1 "Development-Unity"	0.600	0.300	0.400	0.500	0.600	0.545
Task 1 "Legal Procedure"	0.561	0.605	0.582	0.676	0.658	0.667
Task 1 "Emergency-Policy"	0.414	0.324	0.364	0.520	0.351	0.419
Task 1 "Law-and-order"	0.545	0.667	0.600	0.500	0.778	0.609
Task 1 "Utilitarian-Deterrence"	1.000	0.400	0.571	0.000	0.000	Und.*
Task 1 "Last-resort"	1.000	0.500	0.667	1.000	0.500	0.667
Task 1 "Intelligence"	0.200	0.333	0.250	0.500	0.667	0.571
Task 1 "Misc"	0.441	0.455	0.448	0.485	0.485	0.485
Task 1 macro-average	0.512	0.502	0.498	0.574	0.559	**0.560**
Task 2	0.326	0.577	0.416	0.622	0.289	0.394

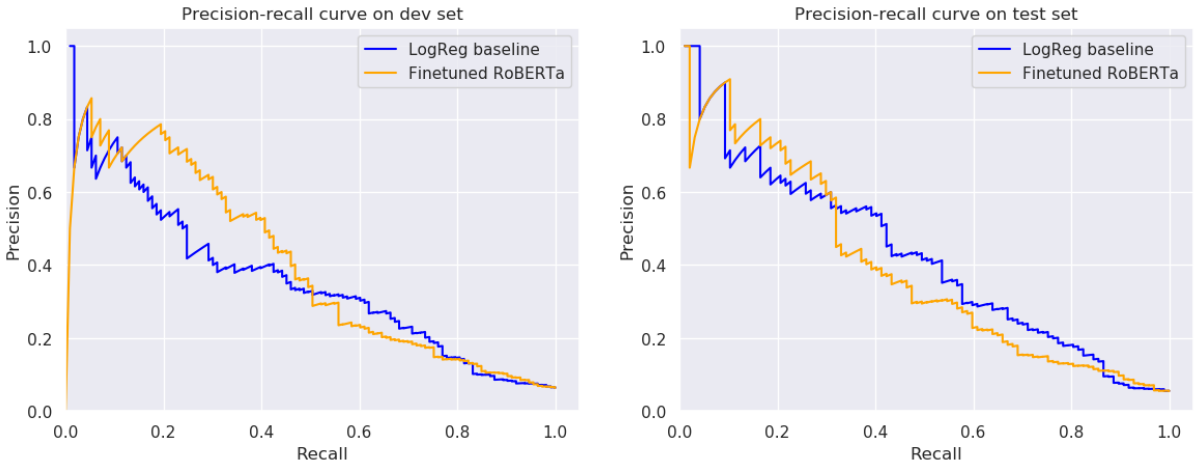
Table 1: Task performance computed on the test set. The macro-average for each statistic was calculated by performing a weighted average of each label's statistic, weighted by the true number of that label appearing in the test set. Values marked as "Und." had a denominator of 0 and were counted as 0 for the purposes of determining the corresponding macro-average; the fine-tuned RoBERTa model labeled 8 sentences as Utilitarian-Deterrence, all incorrectly. The overall accuracies of the models were: 50.2% (logistic regression) and 55.9% (RoBERTa) on task 1 and 91.1% (logistic regression) 95.1% (RoBERTa) on task 2. For task 1, hyperparameters selected on the development set were as follows: logistic regression regularization strength = 0.01; RoBERTa batch size = 16; RoBERTa learning rate = 0.00001. For task 2, hyperparameters selected on the development set were as follows: logistic regression regularization strength = 10; RoBERTa batch size = 32; RoBERTa learning rate = 0.00003.

5 Findings

Task 1: Classifying Rationalizations

Given a short excerpt of text known to express a rationalization (typically one or two sentences), this task requires mapping it to one of our twelve codes. The two models' performance by category and in aggregate on the test set of 213 sentences is reported in Table 1. Given that some classes are thematically similar, we also train models on the same data with an edited labeling scheme, which condenses the set of twelve labels down to six meta categories; we report these results in Table 3 in the Appendix, but do not discuss them further here.

The most-frequent-class baseline in the twelve-way version of this task would lead to an accuracy of roughly 18.3%. (Information about how each class was represented in each split of the data is located in Table 2.) Our baseline and fine-tuned RoBERTa model achieve accuracies of 50.2% and 55.9%, respectively. Computing the F_1 score for each class and averaging those weighted by the number of true examples of each class in the test set, our RoBERTa model achieves a macro-average F_1 of over 0.06 over the baseline. We also note that our RoBERTa model's classwise F_1 score meets or exceeds the classwise F_1 score of our baseline for every class except one (Utilitarian-Deterrence), which comprises five of the test set's examples. These results demonstrate the strength of RoBERTa on a task with relatively curated data.



(a) Precision-recall curves for development set

(b) Precision-recall curves for test set

Figure 2: Precision-recall curves on both development and test sets on task 2 of finding rationalizations. The two held-out sets exhibit similar trends, but their differences also underscore the variance between different documents of the same corpus, which we used to split our data.

Task 2: Detecting Rationalizations

This task is a binary sentence classification problem: Given a sentence,[11] does it express a rationalization for internment? Any sentence that overlapped with any hand-coded rationalization annotation was given a positive label. Any sentence that did not overlap with a hand-coded rationalization received a negative label. The negative-label set of data therefore included many chunks of text that represent OCR transcription failures.

For this task, the vast majority of sentences were negative; in each split of the data, negative instances outnumbered positive instances by a ratio of at least nine to one. On the test set, for which only 97 out of 1,767 sentences were positive, a baseline accuracy of labeling everything as negative would achieve an accuracy of 94.5%. Our logistic regression model reached an accuracy of 91.1%, while our RoBERTa model reached an accuracy of 95.1%. Performance is reported in Table 1, with precision-recall curves given in Figure 2. We see in Table 1 that, at the default threshold, our logistic regression baseline performs better with respect to recall of positive instances, while RoBERTa exhibits higher precision, but neither model performs particularly well.

These results demonstrate that, in a task exemplifying such challenges as severe label imbalance, identifying KWISI concepts, and transcription difficulties, both older and current NLP models still struggle. We discuss this further in Section 6.

6 Discussion

We first examine the performance of our two models for task 1. We select examples from the test data that highlight the relative strengths of these different models, and which help to explain discrepancies in performance. One point to highlight is that our RoBERTa-based model is capable of correctly classifying instances in the absence of a strong keyword signal from any one of its tokens. For example, in

[11] We used simple computational heuristics (e.g., separating according to sentence-ending punctuation that does not appear to be part of an acronym, section number, title, etc.) to determine the sentence boundaries.

response to the sentence, "You will know that it is my intention to seek the agreement of my colleagues to new procedures which in the event or internment being brought at an end will make It possible to keep in custody those members of the IRA whose continued liberty is a threat to security," our baseline model incorrectly predicted the label "Terrorism", while our RoBERTa-based model correctly predicted "Emergency-Policy" as the label.[12] It also becomes evident from examining misclassified instances that more context than just the sentence level might be required in some cases: both models misclassified a number of short, ambiguous sentences , but in these cases, the ambiguous sentence alone, removed from its context, provides insufficient information for classification. For this reason, one potential avenue for improvement on this kind of task could be to integrate document-level context into models.

We now examine task 2 through more examples that illustrate when both models perform well, and when they encounter difficulties. The following sentence is typical of a positive sentence that both models correctly flag: "In the opinion of the Attorney General it would be possible to provide by legislation for the establishment of a procedure whereby persons would be subject to preventive detention on the exercise of the discretion of a tribunal which was satisfied that the persons brought before it had been concerned with terrorism." This is a setting in which words have been cleanly transcribed, and the sentence is long and relatively self-contained. This is in contrast to a sentence such as the following, which both models failed to identify as positive: "Now there is no sign of co-operation from Jack Lynch's government at the moment." In order to understand the implied meaning of this sentence, more information about the surrounding document is required.[13]

One additional difference between these tasks is that almost all of the severe digitization errors were only included in task 2. Given that *both* positive and negative sentences contain transcription errors, though, this may have had an outsized impact on our fine-tuned version of RoBERTa. Our baseline's vocabulary is extracted from our training data for this task; RoBERTa's vocabulary is fixed once pre-training has started, well before it is exposed to our fine-tuning data. In contrast to task 1, this task of *finding* rationalizations in a broader section of the corpus exposes the model to severe OCR errors which have rendered parts of the data unintelligible, and which were overwhelmingly part of examples classed as negative. (Any hand-coded rationalizations that were extracted from the corpus for task 1 were hand-corrected during that process.) These unintelligible sentences would have been analyzed into mostly very short (subword) tokens, of which there are a small number, from RoBERTa's wordpiece tokenization. These tokens are not inherently meaningful and are expected to provide more or less random signal; since they are infrequent, this will lead to spurious correlations with labels and poor generalization. There may be other factors in play as well, but these transcription errors are certainly a problem for current models.

7 Conclusions

Political scientists have applied a variety of NLP methods—including supervised text classification and unsupervised text modeling—to great effect. However, these approaches are not particularly well suited to deal with KWISI concepts—those that are specific, complex, and cognitively identifiable to human coders but not neatly tied to specific words or phrases.

In this research letter, we demonstrated how new advances in automation for text-related tasks, originating in the computing field of natural language processing (NLP), offer potential improvements

[12] Because not all members of the IRA engaged in terrorism, we only coded references to IRA members that were explicitly engaged in violence as part of the "Terrorism" rationalization.

[13] In its document context, this sentence contained an implicit rationalization for internment that hinged on the behavior of other governments. British officials hoped that Mr. Lynch (the leader of Ireland) would also adopt internment, and that doing so would rationalize and legitimize Northern Ireland's internment policies.

for scaling qualitative analysis and dealing with KWISI concepts in "noisy" data. Here, we use a novel data source—recently declasified archives of the UK Prime Minister's correspondence during the "Troubles in Northern Ireland"—and assess the effectiveness of emerging techniques from NLP in identifying and categorizing rationalizations for human rights violations in that collection. In doing so, this research makes two specific contributions. First, it models a pipeline for retrieving, digitizing, and using NLP methods to analyze government archive documents. Second, it demonstrates the promise that these advances in NLP offer toward *classifying* KWISI concepts into pre-defined categories, and also highlights existing challenges that remain to accomplishing tasks that involve less human oversight (e.g., identifying KWISI concepts within broader swaths of political text).

A variety of additional research steps may further improve the model's performance but require careful consideration. First, addressing (and potentially eliminating) the corpus's unintelligible transcriptions of text could reduce an overload onto the relatively few short elements of the model's vocabulary. This is defensible insofar as the unintelligible text is randomly dispersed across the text corpus we examine here (and we believe that it is randomly dispersed). More importantly, much remains unknown about recent modeling advances in NLP. As is the case for a variety of NLP and text-as-data approaches (Denny and Spirling, 2018), fine-tuning a pretrained model involves making a variety of design decisions for which there is not necessarily a clear, appropriate choice. These discretionary decisions may be more consequential for a model's performance than we are currently aware. Scholars who adopt recent NLP models should remain cognizant of these existing uncertainties. With these caveats in mind, recent advances in natural language processing offer considerable promise for scaling scholars' analyses of KWISI political concepts.

References

Acree, Brice DL, Justin H Gross, Noah A Smith, Yanchuan Sim and Amber E Boydstun. 2020. "Etch-a-Sketching: Evaluating the post-primary rhetorical moderation hypothesis." *American Politics Research* 48(1):99–131.

Adler, E. Scott and John Wilkerson. 2012. "Congressional Bills Project: 1947–2008.".

Agamben, Giorgio. 2005. *State of Exception*. Chicago, IL: University of Chicago Press.

Beauchamp, Nick. 2012. "Using text to scale legislatures with uninformative voting.".

Blei, David M. 2012. "Probabilistic topic models." *Communications of the ACM* 55(4):77–84.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of machine Learning research* 3(Jan):993–1022.

Bouchat, Sarah. 2020. Text Classification and Clustering. In *Sage Handbook for Research Methods in Political Science and International Relations*, ed. Robert J. Franzese and Luigi Carini. Sage Publications.

Bouchat, Sarah B. 2018. "InDetermination: Measuring Uncertainty in Social Science Texts.".

Boydstun, Amber E. 2013. *Making the news: Politics, the media, and agenda setting*. University of Chicago Press.

Bustikova, Lenka, David S Siroky, Saud Alashri and Sultan Alzahrani. 2020. "Predicting Partisan Responsiveness: A Probabilistic Text Mining Time-Series Approach." *Political Analysis* 28(1):47–64.

Campbell, Colm and Ita Connelly. 2003. "A model for the 'war against terrorism'? Military intervention in Northern Ireland and the 1970 Falls Curfew." *Journal of Law and Society* 30(3):341–375.

Card, Dallas, Amber Boydstun, Justin H Gross, Philip Resnik and Noah A Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. pp. 438–444.

Carlson, David and Jacob M Montgomery. 2017. "A pairwise comparison framework for fast, flexible, and reliable human coding of political texts." *American Political Science Review* 111(4):835–843.

Collingwood, Loren and John Wilkerson. 2012. "Tradeoffs in accuracy and efficiency in supervised learning methods." *Journal of Information Technology & Politics* 9(3):298–318.

Davenport, Christian. 2007. *State Repression and the Domestic Democratic Peace*. New York, NY: Cambridge University Press.

Denny, Matthew J and Arthur Spirling. 2018. "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It." *Political Analysis* 26(2):168–189.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. "BERT: Pre-training of deep bidirectional transformers for language understanding." *arXiv:1810.04805 [cs.CL] for Computational Linguistics (NAACL)* .

Dewan, Torun and Arthur Spirling. 2011. "Strategic opposition and government cohesion in Westminster democracies." *American Political Science Review* pp. 337–358.

Eisenstein, Jacob, Amr Ahmed and Eric P Xing. 2011. "Sparse additive generative models of text." pp. 1041–48.

Eshima, Shusei, Kosuke Imai and Tomoya Sasaki. 2020. "Keyword Assisted Topic Models." *arXiv preprint arXiv:2004.05964* .

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political analysis* 21(3):267–297.

Grimmer, Justin and Gary King. 2011. "General purpose computer-assisted clustering and conceptualization." *Proceedings of the National Academy of Sciences* 108(7):2643–2650.

Gross, Oren and Fionnuala Ní Aoláin. 2006. *Law in Times of Crisis: Emergency Powers in Theory and Practice*. Vol. 46 Cambridge, MA: Cambridge University Press.

Gururangan, Suchin, Tam Dang, Dallas Card and Noah A Smith. 2019. "Variational pretraining for semi-supervised text classification." *arXiv preprint arXiv:1906.02242* .

Hopkins, Daniel J and Gary King. 2010. "A method of automated nonparametric content analysis for social science." *American Journal of Political Science* 54(1):229–247.

Hu, Yuening, Jordan Boyd-Graber, Brianna Satinoff and Alison Smith. 2014. "Interactive topic modeling." *Machine learning* 95(3):423–469.

Katagiri, Azusa and Eric Min. 2019. "The Credibility of Public and Private Signals: A Document-Based Approach." *American Political Science Review* 113(1):156–172.

King, Gary, Jennifer Pan and Margaret E Roberts. 2013. "How censorship in China allows government criticism but silences collective expression." *American Political Science Review* pp. 326–343.

Kingma, Diederik P and Jimmy Ba. 2014. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* .

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:1907.11692 [cs.CL]* .

Lucas, Christopher, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer and Dustin Tingley. 2015. "Computer-assisted text analysis for comparative politics." *Political Analysis* 23(2):254–277.

Mcauliffe, Jon D. and David M. Blei. 2008. "Supervised topic models." *Advances in neural information processing systems* pp. 121–128.

Nay, John J. 2016. "Gov2vec: Learning distributed representations of institutions and their legal text." *arXiv preprint arXiv:1609.06616* .

O'Connor, Brendan, Brandon M. Stewart and Noah A. Smith. 2013. "Learning to extract international relations from political context." pp. 1094–1104.

Parthasarathy, Ramya, Vijayendra Rao and Nethra Palaniswamy. 2019. "Deliberative Democracy in an Unequal World: A Text-As-Data Study of South India's Village Assemblies." *American Political Science Review* 113(3):623–640.

Pennington, Jeffrey, Richard Socher and Christopher D Manning. 2014. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* .

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer. 2018. "Deep Contextualized Word Representations." *arXiv:1802.05365 [cs.CL]* .

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.

Rheault, Ludovic and Christopher Cochrane. 2020. "Word embeddings for the analysis of ideological placement in parliamentary corpora." *Political Analysis* 28(1):112–133.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. "Structural topic models for open-ended survey responses." *American Journal of Political Science* 58(4):1064–1082.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi et al. 2013. The structural topic model and applied social science. In *Advances in neural information processing systems workshop on topic models: computation, application, and evaluation*. Vol. 4 Harrahs and Harveys, Lake Tahoe.

Schoonvelde, Martijn, Gijs Schumacher and Bert N Bakker. 2019. "Friends With Text as Data Benefits: Assessing and Extending the Use of Automated Text Analysis in Political Science and Political Psychology." *Journal of Social and Political Psychology* 7(1):124–143.

Spirling, Arthur. 2012. "US treaty making with American Indians: Institutional change and relative power, 1784–1911." *American Journal of Political Science* 56(1):84–97.

Wilkerson, John and Andreu Casas. 2017. "Large-scale computerized text analysis in political science: Opportunities and challenges." *Annual Review of Political Science* 20:529–544.

A Appendix

Rationalization Frequencies

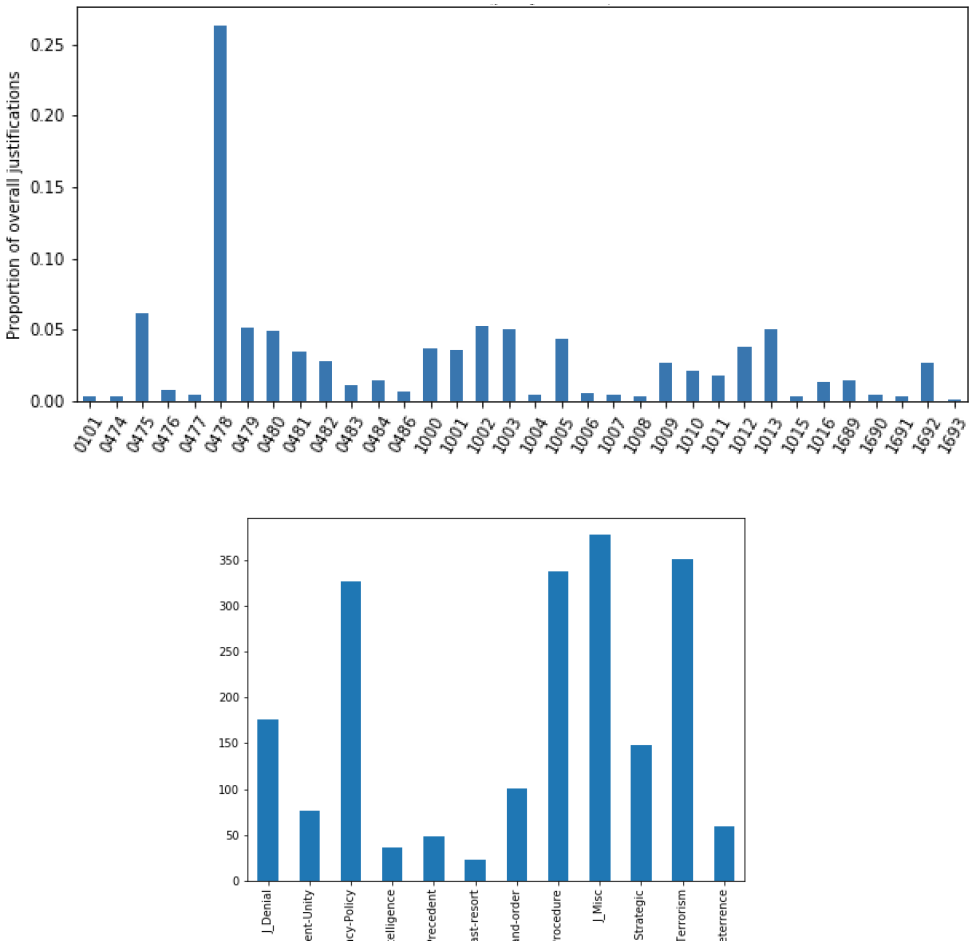


Figure 3: *Top:* Rationalization frequency per PREM 15 file (proxy for time). Internment introduced during PREM 15 0478. *Bottom:* Rationalization frequency (count) per category.

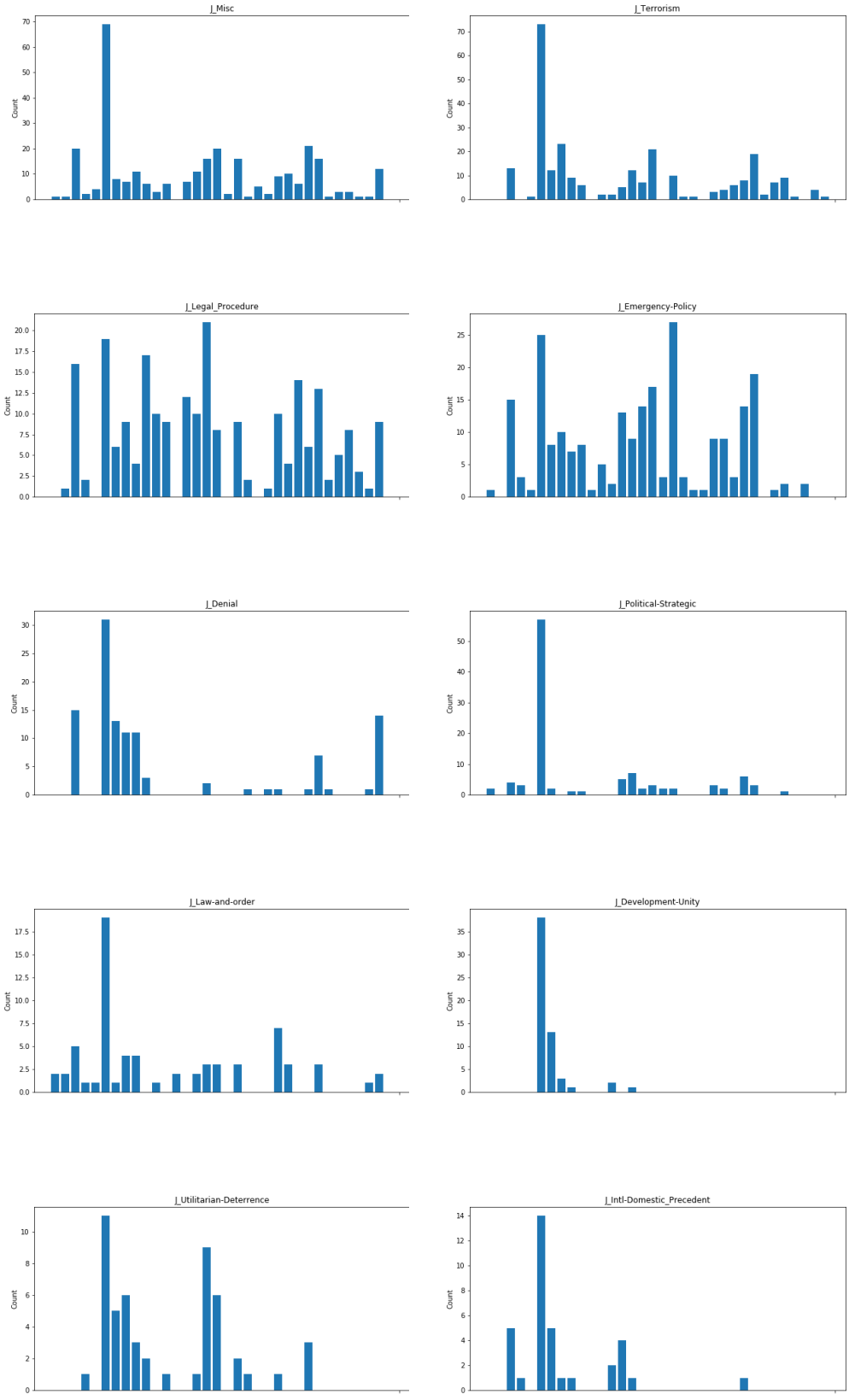


Figure 4: Rationalization frequency with PREM folders over time (raw counts). Rationalizations occurring very infrequently (Intelligence and Last Resort) were omitted from this figure.

	Training sents	Dev. sents	Test sents	Total docs
Terrorism	265	48	39	209
Intl-Domestic Precedent	39	6	3	42
Denial	130	21	25	123
Political-Strategic	123	16	9	100
Development-Unity	56	11	10	59
Legal Procedure	279	21	38	239
Emergency Policy	261	29	37	237
Law-and-order	78	14	9	84
Utilitarian-Deterrence	48	7	5	47
Last-resort	18	3	2	22
Intelligence	31	2	3	28
Misc	312	32	33	249

Table 2: For task 1, the number of sentences of each label in each split of the data, along with the total number of documents containing all of a label's sentences. We created our train/dev/test splits by randomly assigning all sentences in a single document together.

Model Outputs

	Logistic regression			Fine-tuned RoBERTa		
	Precision	Recall	F_1	Precision	Recall	F_1
Task 1 "Political"	0.636	0.778	0.700	0.667	0.889	0.762
Task 1 "Security"	0.521	0.446	0.481	0.595	0.446	0.510
Task 1 "Rights not violated"	0.517	0.536	0.526	0.567	0.607	0.586
Task 1 "Terrorism"	0.458	0.564	0.506	0.566	0.769	0.652
Task 1 "Legal"	0.543	0.658	0.595	0.667	0.684	0.675
Task 1 "Misc"	0.581	0.419	0.486	0.595	0.512	0.550
Task 1 macro-average	0.530	0.526	0.522	0.602	0.601	**0.594**
Task 2	0.326	0.577	0.416	0.622	0.289	0.394

Table 3: Task performance for all 6 classes on the test set. The macro-average for each statistic was calculated by performing a weighted average of each label's statistic, weighted by the true number of that label appearing in the test set. The overall accuracies of the models were: 52.6% (logistic regression) and 60.1% (RoBERTa) on task 1 and 91.1% (logistic regression) 95.1% (RoBERTa) on task 2. For task 1, hyperparameters selected on the development set were as follows: logistic regression regularization strength = 0.0001; RoBERTa batch size = 32; RoBERTa learning rate = 0.00002. For task 2, hyperparameters selected on the development set were as follows: logistic regression regularization strength = 10; RoBERTa batch size = 32; RoBERTa learning rate = 0.00003.

	Training sents	Dev. sents	Test sents	Total docs
Political	123	16	9	100
Security	436	55	56	327
Rights not violated	169	27	28	154
Terrorism	265	48	39	209
Legal	279	21	38	239
Misc	368	43	43	274

Table 4: For task 1, the number of sentences of each label in each split of the data, along with the total number of documents containing all of a label's sentences. We created our train/dev/test splits by randomly assigning all sentences in a single document together.