Issue Area Discovery from Legal Opinion Summaries using Neural Text Processing

Avi Bleiweiss
BShalem Research, Sunnyvale, U.S.A.

Keywords: Legal Domain, Issue Area Prediction, Transformers, Language Model, Deep Learning.

Abstract: Applying existed methods of language technology for classifying judicial opinions into their respective issue areas, often requires annotation voting made by human experts. A tedious task nonetheless, further exacerbated by legal descriptions consisting of long text sequences that not necessarily conform to plain English linguistics or grammar patterns. In this paper, we propose instead a succinct representation of an opinion summary joined by case-centered meta-data to form a docket entry. We assembled over a thousand entries from court cases to render our low-resourced target legal domain, and avoided optimistic performance estimates by applying adversarial data split that ensures the most dissimilar train and test sets. Surprisingly, our experimental results show that fine-tuning a pretrained model on standard English recovers issue area prediction by 9 and 8 F1 percentage points over a pretrained model on the legal domain, for macro and weighted average scores, respectively.

1 INTRODUCTION

Evidenced by the proliferation of legal-tech companies, applying Artificial Intelligence (AI) to law is slowly transforming the profession. Notably the application of natural language processing (NLP) to legal text attracted great interest in the community. Chen (2019) proposes that machine-learned predictive analytics of case outcomes can be used to measure judge bias or fairness. Whereas Volokh (2019) takes a leap forward and envisions that deep learning-based text generation models will produce well enough judicial opinions and soon allow for replacement of human judges by machines.

The reasoning behind particular legal opinions and rulings are often cited in other legal cases. Binding precedent helps ensure court rulings remain consistent among similar cases. In practice, precedent entails the classification of case texts and decisions, a highly time consuming and labor intensive task, when curated by humans. Lame (2005), Evans et al. (2007), and Ashley and Brüninghaus (2009) were of the earliest to propose automating the analysis of legal opinions using language technology. Their work facilitated an essential tool to law practitioners for streamlining precedent.

Modern approaches to various NLP tasks utilize `BERT`-derivative pretrained language models (Devlin et al., 2019) that are based on the transformer network (Vaswani et al., 2017). Fine-tuning `BERT` models on domain-specific data displayed marked performance gains (Xia et al., 2020). Our work explores the feasibility of a distilled version of standard `BERT` to deliver similar results on the legal domain, by applying transfer learning from structured text to non-standard legal text.

In this paper, we adopted sustainable NLP for the legal domain. Rather than using long spans of text, our model is fed with judicial opinion summaries that are considerably more concise. Our system contrasts inference performance of `DistilBERT` (Sanh et al., 2019) with `LegalBERT` (Chalkidis et al., 2020), when fine-tuned for multi-class text classification on the train split of our opinion summary dataset. Our contribution is twofold: (1) a high-quality and sustainable opinion summary dataset scraped from `FindLaw`, [1] and paired with case-centered meta-data extracted from the modern US Supreme Court Database (`SCDB`; Spaeth et al., 2020), and (2) through qualitative evaluation of issue area prediction, we show that using a synoptic outline of opinion content renders an effective adaptation of the legal domain to standard text, and aids in perceiving law as a generalized NLP problem. To the extent of our knowledge, we are the first to use `DistilBERT` on the legal domain. We made

[1] https://reference.findlaw.com/

Bleiweiss, A.
Issue Area Discovery from Legal Opinion Summaries using Neural Text Processing.
DOI: 10.5220/0010974300003116
In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART 2022) - Volume 3*, pages 938-944
ISBN: 978-989-758-547-0; ISSN: 2184-433X

Table 1: Our legal-case object representation.

Docket Number	00-949
Term	2000
Name	GEORGE W. BUSH v. ALBERT GORE, JR.
Issue Area	Civil Rights
Direction	Conservative
Decision	Per Curiam Argued
Case ID	70
Opinion Summary	the absence of specific uniform standards for manual vote recounts, especially where the evidence shows that vote counting standards varied or changed within counties, violates the equal protection clause.

our opinion summary dataset available from a public repository. [2]

2 RELATED WORK

Although text classification is a widely researched area motivated by great practical importance, NLP methods applied to domain-specific problems have been relatively understudied. In an early work, Nallapati and Manning (2008) have explored binary classification in the domain of legal docket entries. They showed that state-of-the-art (SotA) Machine Learning (ML) classifiers perform poorly to capture the semantics of the text. Katz et al. (2017) built an ML model based on the random forest method for predicting decisions of the Supreme Court of the United States (SCOTUS). Relying mainly on SCDB (Spaeth et al., 2020) meta-data as inputs, and less on legal opinion text, they achieved prediction accuracy of 70.2% at the case outcome and 71.9% at justice vote level. Primarily supporting law professionals to efficiently perform an exhaustive search of case related documents, Merchant and Pande (2018) proposed an automated text summarization system that captures concepts from lengthy judgments by applying latent semantic analysis (LSA). Their model achieved a moderate unigram ROUGE-1 score of 0.58 on average.

Recently, Wan et al. (2019), Chalkidis et al. (2019), and Soh et al. (2019) reviewed the classification of lengthy legal documents with the main objective to ameliorate input constraint imposed by BERT (Devlin et al., 2019) on token-length exceeding 512 terms. They chose an elemental BiLSTM or BiGRU neural architectures with attention that were fed by Doc2Vec or GloVe embeddings (Le and Mikolov, 2014). Wan et al. (2019) used a data splitting approach that improved performance by 1 F1 percentage point, although at a prohibitively expensive storage-complexity of the dataset they used. Surprisingly, Soh et al. (2019) showed traditional ML baselines to outperform pretrained language models.

On the other hand, Chalkidis et al. (2020) applied BERT models to downstream legal tasks and studied the performance impact when trading off domain pretraining and fine-tuning. In our work, we used their LegalBERT model, fine-tuned on our opinion summary dataset for the task of issue area classification.

3 DATA

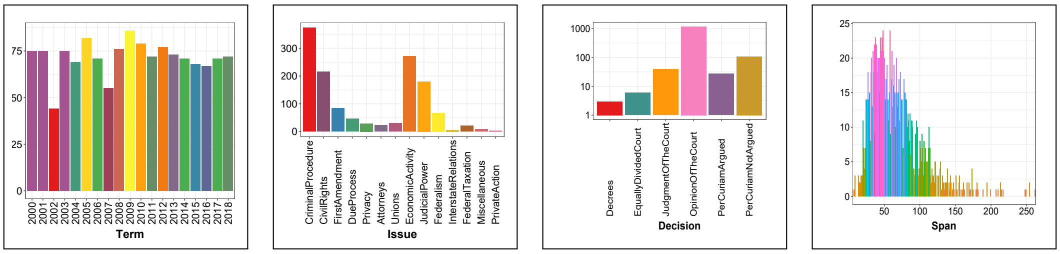
We constructed a new dataset for our experiments that consists of opinion summaries we scraped from the Opinion Summaries Repository made available by FindLaw. [3] FindLaw provides public access to summaries of published opinions that span about two decades from 2000 till 2018, and pertain to U.S. and selected state supreme and appeals courts. Our study centers around the U.S. Supreme Court that has consistently issued between 70–90 opinions per term over the past twenty years (Figure 1a). We successfully merged 1,358 opinion summaries with case centered meta-data from SCDB (Spaeth et al., 2020), [4] using a matching docket number. Our representation of a case object includes the substantive issue area and decision direction attributes, and the decision-type outcome variable (Table 1).

In Figure 1b, we show the distribution of the fourteen SCDB-defined issue areas across our entire dataset objects. Overall, the spread depicted is fairly uneven, as criminal procedure, economic activity, civil rights, and judicial power dominate with about 83 percentage points of the total cases. The decision direction allocation is nearly uniform with a majority of 698

[2] https://github.com/bshalem/jos

[3] https://caselaw.findlaw.com/summary.html

[4] Case centered data on: http://scdb.wustl.edu/data.php



(a) (b) (c) (d)

Figure 1: Distributions across our entire opinion summaries dataset of (a) term, (b) issue area (unbalanced), (c) decision (in logarithmic scale), and span length in words of (d) opinion summaries.

and 644 conservative and liberal descriptions, respectively, along with only 16 unspecifiable cases. Using a logarithmic scale, we review in Figure 1c the distribution of the six court decision types showing a significant bias with 1,175 samples, or 86%, classified as Opinion of the Court, and distant second 135 Per Curiam cases.

Algorithm 1: Data Split Dissimilarity.

```
Input  : $S$  a list of opinion summaries
Output: $w$  Wasserstein distance
emd($x,y,p=1$) return $\|x-y\|^p$ ;
$w \leftarrow 0$;
// X/Y for Train/Test sets
$X, Y \leftarrow$ ResampleRandomSplit($S$);
$X_e, Y_e \leftarrow$ ExtractEmbeddings($X, Y$);
for $y \in Y_e$ do
    for $x \in X_e$ do
        $w \leftarrow w +$ emd($x, y$);
    end for
end for
```

Distribution of span extent in words for the opinion summaries are shown in Figure 1d. Opinion summary lengths range from 6 to 262 tokens with an average sequence length of 67 words. In contrast, the entire opinion text typically extends from 10,000 to 50,000 words (Wan et al., 2019)— an expansion of about two orders of magnitude. For example, the published report of the BUSH v. GORE case reaches across 61 pages of a PDF file with a total of 24,804 words, [5] and is considerably reduced to a compact 28-token summary paragraph as shown in Table 1. Similarly, Soh et al. (2019) report 6,968 tokens on average for Singapore Supreme Court judgments, and Koreeda and Manning (2021) has 2,254 for contractual data.

4 SETUP

In this section, we provide details of dataset preprocessing and fine-tuning methodology.

Corpus Preprocessing. Our scraped opinion summaries underwent several cleanup steps to fit our task. First, to perform plausible predictions of an issue area, we required to alleviate the uneven presence of issue labels (Figure 1b) by remapping the fourteen SCDB categories into five classes and produce a balanced dataset, as shown in the distribution of Table 2. However, Superior Court decisions were severely skewed in the dataset toward a signed opinion (Figure 1c), such that predicting an outcome deemed impractical and thus excluded from the scope of this paper. In the course of matching SCDB meta-data with FindLaw opinion summaries, we found about a dozen of cases with repeated docket numbers that we removed. Lastly, we pulled out numerical entities from many opinion summaries that contain section references, or else they may impact issue area prediction adversely.

We apportioned our target legal data into train and test sets with an 80-20 split that amounts to 1,085 and 273 case-summary pairs, respectively. To avoid overestimating real performance (Søgaard et al., 2021), we used adversarial data splits by maximizing the Wasserstein distance w between the train and test partitions that were generated across randomly resampling our data for ten times ($\text{argmax}_{1 \leq i \leq 10} w_i$). We further outline our method in Algorithm 1, noting that it is performed for each resampling iteration. After reshuffling the data and generating splits that each retain issue area balance, we extracted a single embedding vector for every opinion summary. We then pair test and train embeddings, calculate an Earth Mover's Distance (EMD; Rubner et al., 1998), and return the sum of all distances. The time complexity of the algorithm is $O(mn)$, where m and n are the size of the test and train sets, respectively.

[5] https://supremecourt.gov/opinions/USReports.aspx

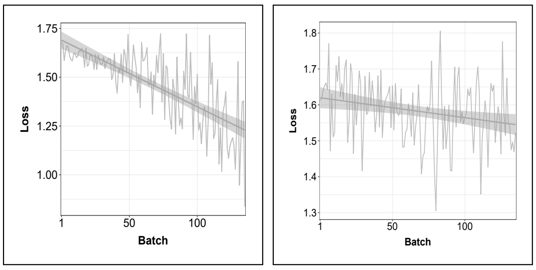
Table 2: Balanced data set after remapping issue area into five classes. Judicial Power examples merged with First Amendment, while the Other case category includes samples from Due Process, Privacy, Attorneys, Unions, Federalism, Interstate Relations, Federal Taxation, Private Action, and Miscellaneous.

Criminal Procedure	Civil Rights	Economic Activity	Judicial Power	Other
374	215	271	264	234

Table 3: F1 scores of multi-class opinion classification comparing `distilbert-base` with `legal-bert-base` models. The support column indicates the number of ground-truth cases for an individual issue area.

	`distilbert-base`			`legal-bert-base`			
Issue Area	Precision	Recall	F1	Precision	Recall	F1	Support
Criminal Procedure	0.65	0.97	0.78	0.67	0.99	**0.8**	75
Civil Rights	0.71	0.12	**0.20**	0.8	0.09	0.17	43
Economic Activity	0.61	0.80	**0.69**	0.42	0.93	0.58	55
Judicial Power	0.42	0.51	**0.46**	0.49	0.34	0.4	53
Other	0.47	0.17	**0.25**	0	0	0	47
Accuracy			**0.58**			0.54	273
Macro Average	0.57	0.51	**0.48**	0.48	0.47	0.39	273
Weighted Average	0.58	0.58	**0.52**	0.49	0.54	0.44	273

Fine-Tuning. Our experimental framework consists of `DistilBERT` (Sanh et al., 2019), [6] a generic and simple language-model that leverages knowledge distillation during pretraining, and `LegalBERT` (Chalkidis et al., 2020), [7] a model pretrained on 12GB of diverse English legal text from the fields of legislation, court cases, and contracts. We used the uncased version of both models, after converting the text of opinion summaries to lowercase. To perform multi-class classification, we added a linear layer that reduces the network output to our five issue areas and followed with a softmax activation. We used the Adam optimizer (Kingma and Ba, 2014) with a cross-entropy loss function, a learning rate of $1e^{-5}$, and applied a fixed 0.3 dropout.



(a) `DistilBERT`. (b) `LegalBERT`.

Figure 2: Loss progression during fine-tuning on our train-set in one out of four and ten epochs for (a) `DistilBERT` and (b) `LegalBERT`, respectively.

In Figure 2, we show loss behavior in a single epoch during fine-tuning `DistilBERT` (Figure 2a) and `LegalBERT` (Figure 2b) on our opinion summary train-split. The loss pattern in both models has a spiky appearance, but nonetheless consistently descending as the batch count increases. Overall, fine-tuning on `DistilBERT` took four epochs and more than twice long for ten epochs on `LegalBERT`. We used a batch size of 8 and 32 opinion summaries in fine-tuning and inference, respectively.

5 EXPERIMENTS

We ran inference on our opinion summary test set using a checkpoint of both the fine-tuned language model and the vocabulary augmented with tokens contributed by our train set.

In Table 3, we provide classification F1 scores comparing the base models of `DistilBERT` with `LegalBERT`, each fine-tuned on our opinion summary train-set. Performance results are presented for both an individual issue category and cumulative macro and weighted averages. F1 scores for the areas of criminal procedure and civil rights came out relatively close with a slight edge exchanging between the language models. `DistilBERT` had a moderate 6% F1 advantage for judicial power, on the other hand, `DistilBERT` dominated with convincing F1 margins of 11 and 25 percentage points for predicting economic activity and the other issue group, respectively.

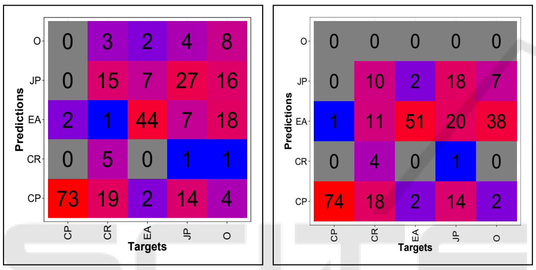
[6] https://huggingface.co/distilbert-base-uncased

[7] https://huggingface.co/nlpaueb/legal-bert-base-uncased

Table 4: Comparing F1 scores against external baselines.

System	Documents	Tokens	Labels	Model	F1
Soh et al. (2019)	623	6,968	31	`bert-base`	0.43
				`bert-large`	0.45
ContractNLI	607	2,254	3	`bert-base`	0.53
				`legal-bert-base`	0.51
Sarkar et al. (2021)	230	768	2	`sentence-bert-base-zs`	0.59
				`sentence-bert-base-fs`	**0.67**
Ours	1,085	67	5	`distilbert-base`	0.48
				`legal-bert-base`	0.39

Notably `LegalBERT` had no definite prediction for cases that were part of the other collection. We contend that the corpora used to pretrain `LegalBERT` might be scope limited and thus challenged by uncovered issue areas of the other group that possesses relatively weak inter-area similarity.



(a) `DistilBERT`. (b) `LegalBERT`.

Figure 3: Confusion matrix representation for multi issue-area classification on (a) `DistilBERT` and (b) `LegalBERT`.

Overall, `DistilBERT` surpassed `LegalBERT` prediction quality consistently for both macro and weighted averages with an F1 score of 48% and 52%, thus gaining 9 and 8 F1 percentage points, respectively. Similarly, the balanced accuracy measure for `DistilBERT` was 0.58 vs 0.54.

In Figure 3, we review the confusion matrix representation for our task of multi issue-area classification. While true positive counts along matrix diagonals are fairly consistent across our models, error rate per class varies. For example, the Economic Activity issue area has 28 false positives and 11 false negatives for a total of 39 misclassifications on `DistilBERT` (Figure 3a), whereas on `LegalBERT` (Figure 3b) there is a much higher rate of 74 incorrect predictions.

Next, we contrast our performance against external baselines that use derivatives of the `BERT` model.

Soh et al. (2019) used a fairly imbalanced dataset of 51 labels in total that was limited to the 30 most frequent issue areas, as the rest of the labels were mapped to the others label. They trained each model type using three classifiers on different training subsets to mitigate remaining label imbalance. However, their work avoids performing fine-tuning and for our analysis we used their F1 scores for a ten percentage point holdout of their train set.

ContractNLI (Koreeda and Manning, 2021) is a document level natural language inference (NLI) tool for reviewing contracts. A hypothesis might thus contradict, entail, or be neutral to a contract. The three-label classification task showed that performance of contradicting labels are impacted far more adversely compared to entailment labels, due to imbalanced label distribution in the dataset. Similar to our results, fine-tuning a model pretrained on legal corpora proved mixed results and did benefit NLI marginally.

Exploring a predictive coding system for regulatory compliance, Sarkar et al. (2021) proposed Few-shot (FS) learning to classify financial-domain data using `SentenceBERT` (Reimers and Gurevych, 2019), and compared it against a Zero-shot (ZS) approach using a pretrained NLI `BART` model (Lewis et al., 2020). Their manually labeled dataset is extremely small and composes sentence-level promissory and non-promissory examples, of which a sentence tagged promissory is considered the hypothesis. `SentenceBERT` encodes each sentence into a fixed-sized embedding vector.

In Table 4, we use macro-average F1 scores to rate performance of multi-class classification, with the intention to refrain from overly optimistic results. F1 scores for ContractNLI are reported as the average of contradiction and entailment labels for each model. Although the systems we considered are fairly diverse in both their goals and data— train or fine-tuning set size and average tokens per document or sentence— scores are nonetheless comparatively concurring. Zero-shot and Few-shot binary text classification were expected to be the highest scoring with F1 of 59% and 67%, respectively. ContractNLI is second with a slight edge of 3 F1 percentage points over our system when pretrained on legal corpora, mainly due to a relatively simpler classification task.

6 CONCLUSION

Our study motivates the use of opinion synopses rather than long-length descriptions to predict issue areas at scale. We analyzed qualitatively whether legal is a domain-specific problem from an NLP tools perspective, or a domain that could be generalized based on representation and the task of interest. Finetuned on our balanced dataset with the most dissimilar splits, we showed that a sustainable generalized language-model is more train-efficient and outperformed a model pretrained on a specialized legal domain.

Our results carve several avenues of future research such as improve performance by removing name entities from summaries, apply text simplification to auto-generate opinion abstractions from long documents, and expand our work to a broad class of prediction tasks in legal studies. As it becomes increasingly important to develop simple, efficient, and reproducible domain-agnostic systems for neural text processing, we hope our approach will help the NLP community to further expand prediction analysis to other humanity disciplines.

ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their insightful suggestions and feedback.

REFERENCES

Ashley, K. D. and Brüninghaus, S. (2009). Automatically classifying case texts and predicting outcomes. *Artificial Intelligence and Law*, 17(2):125–165.

Chalkidis, I., Fergadiotis, E., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2019). Extreme multi-label legal text classification: A case study in EU legislation. In *Natural Legal Language Processing Workshop*, pages 78–87, Minneapolis, Minnesota.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., and Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: (EMNLP)*, pages 2898–2904, Online.

Chen, D. (2019). Judicial analytics and the great transformation of American law. *Artificial Intelligence and the Law*, 27(1):15–42.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 4171–4186, Minneapolis, Minnesota.

Evans, M., McIntosh, W., Lin, J., and Cates, C. (2007). Recounting the courts? applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies*, 4(4).

Katz, D. M., Bommarito, M. J., and Blackman, J. (2017). A general approach for predicting the behavior of the supreme court of the united states. *PLOS ONE*, 12(4).

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980. http://arxiv.org/abs/1412.6980.

Koreeda, Y. and Manning, C. D. (2021). Contractnli: A dataset for document-level natural language inference for contracts. Available at https://export.arxiv.org/abs/2110.01799v1.

Lame, G. (2005). *Using NLP Techniques to Identify Legal Ontology Components: Concepts and Relations*, page 169–184. Springer-Verlag, Berlin, Heidelberg.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International Conference on Machine Learning (ICML)*, pages 1188–1196, Bejing, China.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880, Online.

Merchant, K. and Pande, Y. (2018). NLP based latent semantic analysis for legal text summarization. In *Advances in Computing, Communications and Informatics (ICACCI)*, pages 1803–1807, Bangalore, India.

Nallapati, R. and Manning, C. D. (2008). Legal docket classification: Where machine learning stumbles. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 438–446, Honolulu, Hawaii.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China.

Rubner, Y., Tomasi, C., and Guibas, L. (1998). A metric for distributions with applications to image databases. In *IEEE Conference on Computer Vision*, pages 59–66.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108. http://arxiv.org/abs/1910.01108.

Sarkar, R., Ojha, A. K., Megaro, J., Mariano, J., Herard, V., and McCrae, J. P. (2021). Few-shot and zero-shot approaches to legal text classification: A case study in the financial sector. In *Natural Legal Language Processing (NLLP)*, pages 102–106, Online.

Søgaard, A., Ebert, S., Bastings, J., and Filippova, K. (2021). We need to talk about random splits. In *European Chapter of the Association for Computational Linguistics (EACL)*, pages 1823–1832, Online.

Soh, J., Lim, H. K., and Chai, I. E. (2019). Legal area classification: A comparative study of text classifiers on Singapore Supreme Court judgments. In *Natural Legal Language Processing Workshop*, pages 67–77, Minneapolis, Minnesota.

Spaeth, H. J., Epstein, L., Martin, A. D., Segal, J. A., Ruger, T. J., and Benesh, S. C. (2020). Supreme court database (SCDB), version 2020 release 01. Available at http://supremecourtdatabase.org.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008. Curran Associates, Inc., Red Hook, NY.

Volokh, E. (2019). Chief justice robots. *Duke Law Journal*, 68(2):1135–1192.

Wan, L., Papageorgiou, G., Seddon, M., and Bernardoni, M. (2019). Long-length legal document classification. *CoRR*, abs/1912.06905. http://arxiv.org/abs/1912.06905.

Xia, P., Wu, S., and Van Durme, B. (2020). Which *BERT? A survey organizing contextualized encoders. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 7516–7533, Online.