Partisan Bias in the US Federal Court System

Annabelle Huether
Center for Data Science
New York University
amh9750@nyu.edu

Mary Nwangwu
Center for Data Science
New York University
mcn8851@nyu.edu

Allison Redfern
Center for Data Science
New York University
amr10211@nyu.edu

Abstract

This research project addresses the current gap in understanding the partisan dynamics within the US legal system by focusing on lower courts, which constitute the majority of the judicial landscape. BiLSTM and BERT based models were leveraged to predict the partisan ideology and topic area of lower-court decisions in the United States based on their formal delivered opinions. By aggregating these predictions at the judge level, we analyze the extent to which judges align with their political party, allowing us to identify trends in the increasing or decreasing polarization of judges over time and unveil the specific topics contributing to this trend within the judiciary.

1 Introduction

How partisan is the United States court system? There is increasing public acknowledgement that the judiciary, especially the US Supreme Court, is susceptible to partisan influence, and that such influence can affect court outcomes. However, data for lower federal courts, which make up the vast majority of the US legal system, is harder to study. This project leverages Natural Language Processing (NLP) techniques in an attempt to predict the partisan direction (liberal or conservative) and the general topic area of lower federal court decisions based on the text of their formal delivered opinions. To address the question of lower court partisan influence, these predictions are aggregated to the judge level to examine how the partisanship of individual judges and topics changes over time. We ultimately find that model predictions imply a shift over time toward more conservative decision making in the lower courts.

2 Related Work

The legal field has seen a growing use of NLP applications, driven by the potential of handling unstructured data in documents. Previous studies (Wan et al. 2019; Noguti et al. 2020; Song et al. 2022) addressed NLP model limitations, progressed from text length challenges to domain-specific embeddings, and tackled multi-label classification of legal categories. This study focuses on binary classification of partisanship and multi-classification of topic area for lower federal court decisions, offering a comprehensive analysis compared to prior research. To address input text length issues, we employ text summarization instead of document chunking. Our methodology also involves using LEGAL-BERT, pre-trained on legal documents, to enhance precision. Leveraging Supreme Court decisions for pre-training enables predictions of partisanship and topics in lower court decisions, providing valuable insights into evolving dynamics within the entire US court system and enriching discourse on partisan influences in the legal landscape.

3 Problem Definition and Algorithm

3.1 Task

Supreme Court opinion data is very well-studied and labeled, and has been frequently analyzed for trends in terms of partisanship and topic area; however, lower federal court opinion data has never been labeled or analyzed quantitatively at a large scale. The main task of this project is to create a model that can manufacture this metadata for lower court opinions, as well as analyze the results to understand trends in judge partisanship over time. Due to their high-quality and availability, the Supreme Court opinion texts and labels serve as the training data for partisan direction and topic area classification models. Then, using the best models for partisan and topic area classification, inference is performed on the lower court opinion texts to create partisan direction and topic area labels for these opinions. Finally, these new labels are analyzed using judge metadata to view trends across lower court decisions.

This NLP-based method for assigning partisan and topic labels to lower court decisions has its advantages and drawbacks. While simple labels facilitate model training, assuming clear labels for every text may oversimplify the degree of partisanship. Using only decision text as a feature aids generalization by avoiding author bias, but it assumes a comparable distribution between Supreme Court and lower court decision text. Despite potential downsides, this approach is a valuable initial effort to extract sought-after lower court metadata.

3.2 Algorithm

This project explores three different NLP models for partisan direction classification and topic area classification. Each model is described below, including a more detailed algorithmic explanation of the best performing model family, BERT.

3.2.2 BiLSTM. Short for Bi-Directional Long Short Term Memory, this model is a member of the Recurrent Neural Network (RNN) family, and leverages the LSTM neural network architecture to prevent gradient vanishing. This model learns bidirectional representations from text with an added reversed LSTM layer. In this project, a BiLSTM was trained on Supreme Court decision texts that were tokenized using the Natural Language Toolkit (NLTK) and vectorized using Word2Vec embeddings.

3.2.2 BERT and LEGAL-BERT. Short for Bi-Directional Encoder Representations from Transformers, BERT is a pre-trained large language model (LLM) using a masked language modeling objective to learn bidirectional representations of sequences. In this project, a BERT base model (uncased) with 110M parameters was fine-tuned for classification. BERT is pre-trained on BooksCorpus and English Wikipedia. LEGAL-BERT is a family of BERT models pre-trained specifically on data from the legal domain ($\sim$450K documents). In this project, LEGAL-BERT-BASE (uncased) with 110M parameters was also fine-tuned for classification.

BERT and LEGAL-BERT share the same algorithm for processing examples. BERT utilizes a transformer-based architecture with self-attention mechanisms, allowing it to understand word and phrase context in sequences for improved text representation. BERT (base) has a 12-layer, 12-head encoder. For a judge's opinion text, for example, BERT's tokenizer transforms the text into tokens using WordPiece Tokenizer, adds a special token for classification, and assigns each token an index. These embeddings, combined with positional encodings, are processed in parallel through 12 self-attention layers, refining contextual representations. The final pooled representation is fed into a softmax classification head, generating probability distributions over classes. The class with the highest probability is chosen. During fine-tuning, BERT is optimized by minimizing cross-entropy loss through the comparison of predicted and real labels.

These models face limitations due to computational resources and configurations. As sequence lengths increase, BiLSTM requires more memory and resources than available on NYU's High Performance Computing (HPC) system for deeper training. BERT and LEGAL-BERT are constrained to token counts below 512, leading to implementation of text summarization of the decision texts for training. The assumption is that the first 512 tokens of input sequence summaries sufficiently capture information for these models to learn meaningful representations.

4 Experiment Evaluation

4.1 Data

The datasets utilized for model training and inference were obtained from CourtListener and the Supreme Court Database (SCDB). CourtListener is an open-source research website which consolidates all US federal court data. All opinion text from the Supreme Court ($\sim$28K records) and lower courts ($\sim$9M records) were sourced from CourtListener as well as judge and docket metadata. SCDB provides more comprehensive metadata on US Supreme Court cases from 1791 to 2021 including the Supreme Court case decision direction and topic area. The decision direction, or partisan direction, codes the partisan stance of Supreme Court decisions as liberal, conservative, or unspecified. The topic area categorizes the case into broader topics, such as Criminal Procedure, Civil Rights, First Amendment, and more. The dataset was filtered to begin at the year 1930 in order to capture the onset of modern party alignments under Franklin D. Roosevelt. For non-baseline models, preprocessing involved removing Supreme Court Justice names and titles ("Mr." or "Ms.") to enhance generalization. Opinion text, extracted from HTML and XML formats, was standardized for uniformity, ensuring data consistency and readiness for classification models. Mentioned in the previous section, token classification models often face a 512-token limit, requiring truncation of longer sequences. To preserve essential information in lengthy court opinions, we employed text summarization with LongT5, optimized for handling long input sequences. A base-sized LongT5 model, fine-tuned on the BOOKSUM dataset, summarized Supreme Court decision texts efficiently. Summarization, executed on the NYU HPC GPU A100, processed batches of 5 documents. A secondary round with a batch size of 1 was implemented for initial batches exceeding CUDA memory limits.

4.2 Methodology

In line with project goals, partisan direction evaluation focused on accuracy, calibration, and area under the receiver operating characteristic curve (AUC). Topic area classification was assessed using accuracy and accuracy by class. Accuracy measures overall correctness, while calibration ensures predicted probabilities align with true label likelihood. AUC evaluates the model's ability to distinguish liberal and conservative directions, indicating discriminatory power. These metrics collectively validate the reliability and performance of the classification models in this study. Before applying NLP techniques, a baseline Gradient Boosting model using Word2Vec embeddings averaged across documents was employed. Subsequently, three NLP models—BiLSTM, BERT, and LEGAL-BERT—were implemented and compared (section **3.2**). BiLSTM was chosen for its efficacy in multi-classification tasks with short-text legal documents. BERT, known for capturing complex contextual information, suited nuanced language in legal documents. LEGAL-BERT, a BERT variation pre-trained on legal data, aimed at leveraging domain-specific pre-training. These diverse model choices were made to explore different NLP techniques, each selected for specific strengths and potential to enhance partisan direction or topic area classification performance. Prior to model training or fine-tuning, the Supreme Court opinion text was randomly split into training (21,107 documents), validation (2,700 documents), and test (2,618 documents) sets. During the split it was ensured that two opinions on the same case, such as dissent and a concurrence, would appear in the same dataset to prevent leakage. For each of the three models and two objectives, multiple hyperparameters and architectures were trained or fine-tuned using the Supreme Court training data and assessed using the validation data. The best model of each type was chosen using validation accuracy, and had a final assessment on the test data to compare against the other models.

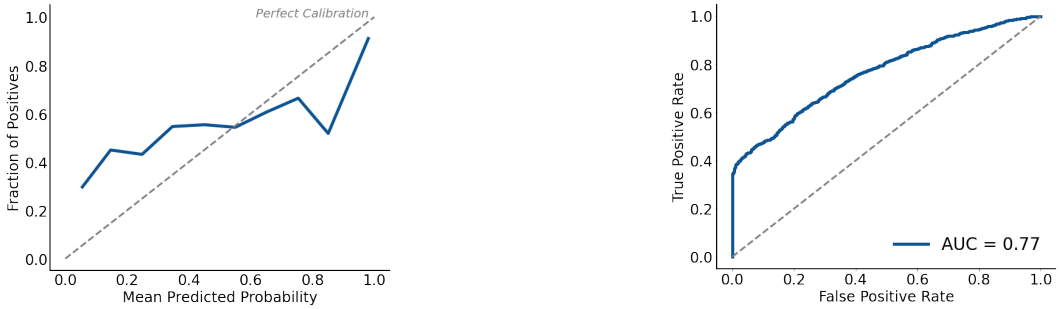
4.3 Results

4.3.1 Partisan Direction. The best hyperparameters and architectures of each partisan direction classification model are shown in Table 1. The calibration and AUC of each model was also found for partisan direction classification. Figure 1a and Figure 1b display the calibration and AUC for the best model, LEGAL-BERT. LEGAL-BERT achieves the highest accuracy on the test set, in addition to being fairly calibrated. LEGAL-BERT also has the largest AUC at 0.77, outperforming the other three models in ability to distinguish between liberal and conservative opinion texts.

4.3.2 Topic Area. The best hyperparameters and architectures for the topic area classification models are displayed in Table 2. The model with the highest accuracy for topic area classification is also LEGAL-BERT. The accuracy by topic area for LEGAL-BERT is shown in Table 3, revealing that it

Table 1: Partisan direction classification model architecture, hyperparameters, & accuracy comparison

Model	Training Data	Architecture & Hyperparameters	Test Accuracy
Gradient Boosted Trees (baseline)	Supreme Court Full Text	9 node maximum depth, 200 estimators	0.6264
BiLSTM	Supreme Court Summarized Text	2 100-node BiLSTM layers, 0.05 dropout, Dense layer (sigmoid activation), 15 epochs, binary cross-entropy loss, 0.001 step size, 250 batch size	0.6318
BERT	Supreme Court Summarized Text	3 epochs, binary cross-entropy loss, softmax head, 2.00×10^{-5} learning rate, 0.1 weight decay, 32 batch size	0.6742
LEGAL-BERT	Supreme Court Summarized Text	5 epochs, binary cross-entropy loss, softmax head, 2.00×10^{-5} learning rate, 0.1 weight decay, 32 batch size	0.6788



(a) Calibration curve.

(b) ROC curve.

Figure 1: Partisan direction classification LEGAL-BERT test set results.

has an accuracy greater than 75% for half of the classes, and that this high accuracy is not limited to topic areas with the highest sample size in the test set. Topic areas with a smaller sample size in Supreme Court data, such as "Miscellaneous" and "Private Action", need more data to be accurately classified.

4.3.3 Inference on Lower Courts. The best partisan direction and topic area classification model, LEGAL-BERT, was then used to perform inference on all authored lower court opinion texts. As a first attempt to check the accuracy of these models on the lower court data, 116 opinion texts were hand-labeled by Aaron Kaufman, Assistant Professor of Political Science at NYU Abu Dhabi and

Table 2: Topic area classification model architecture, hyperparameters, & accuracy comparison

Model	Training Data	Architecture & Hyperparameters	Test Accuracy
Gradient Boosted Trees (baseline)	Supreme Court Full Text	9 node maximum depth, 200 estimators	0.7108
BiLSTM RNN	Supreme Court Summarized Text	4 50-node BiLSTM layers, 0.05 dropout, Dense layer (softmax activation), 10 epochs, categorical cross-entropy loss, 0.001 step size, 250 batch size	0.6165
BERT	Supreme Court Summarized Text	3 epochs, categorical cross-entropy loss, softmax head, 5.00×10^{-5} learning rate, 0.01 weight decay, 32 batch size	0.7552
LEGAL-BERT	Supreme Court Summarized Text	3 epochs, categorical cross-entropy loss, softmax head, 3.00×10^{-5} learning rate, 0.01 weight decay, 32 batch size	0.7872

Table 3: Topic area classification LEGAL-BERT accuracy by topic & sample size in test set

Topic Area	Sample Size	LEGAL-BERT Test Accuracy
Unions	105	88.6%
Federal Taxation	147	88.4%
Criminal Procedure	604	88.2%
Economic Activity	512	82.4%
First Amendment	244	82.0%
Civil Rights	416	80.8%
Privacy	34	76.5%
Attorneys	33	69.7%
Judicial Power	295	64.7%
Due Process	112	50.0%
Federalism	111	45.9%
Miscellaneous	4	0.0%
Private Action	1	0.0%

the mentor for this project. For partisan direction, the results were 0.5070 accuracy for all model prediction probabilities and 0.6316 accuracy for predictions with greater than 0.9 probability. For topic area, the sample accuracy was 0.4615 for all labeled topics and 0.65625 when excluding the "Private Action" topic area, since this type of case had a very low frequency in the Supreme Court data but is a common topic area in lower court decisions. The results of this hand-labeling exercise are to be validated further with additional input from experts in the legal domain, and therefore should only be considered preliminary.

The lower court inference results were then analyzed across various criteria. Generally, the model predicts that lower courts overall are becoming more conservative. Figure 2a shows the average likelihood of judges having opinions with partisanship that opposes their appointing presidents' parties over time. Predictions illustrate that judges appointed by conservative presidents are becoming less likely to oppose their ideology, while judges appointed by liberal presidents are increasingly opposing their party. Therefore, both types of judges are shifting toward conservative ideology over time. This trend can also be seen in individual judges' time since their first case. Figure 2b depicts the average liberal decision ratio of judges during their term, showing that both judges appointed by liberal and conservative presidents become more conservative over the course of their term.

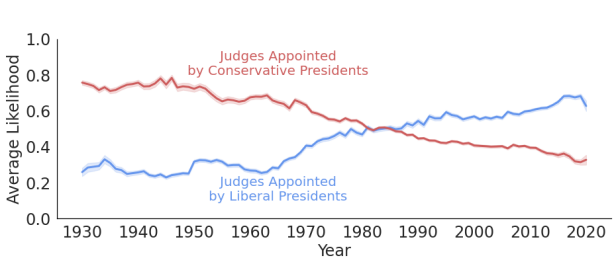
When considering this trend toward conservative opinions in the lower courts across topic, there is consistency across predicted topic areas as well. Looking at just the top 5 predicted topics, which makes up over 90% of all predictions, Figure 2c illustrates the average liberal decision ratio over time for each topic area. Decisions regarding these main topics are becoming more conservative over time. This trend is consistent when stratifying across appointing presidents, as both parties are becoming more conservative on each topic over time.

4.4 Discussion

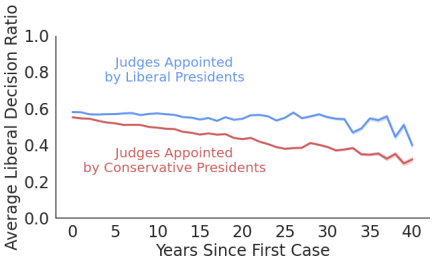
The results of the study illustrate that the task of utilizing Supreme Court text data to predict the partisan labels and topic area labels of lower court opinions is inherently challenging. Sub-optimal accuracy might suggest that these models do not have the capacity nor size to learn the necessary representations of the opinion texts, or rather, that the labels "liberal" or "conservative" are too broad to express degree and granularity of partisanship. For example, this problem arises for decision texts that take both a pro-liberal and a pro-conservative stance in the same document, or fall more into a moderate category rather than in a liberal or conservative one. As such, the results from inference on lower courts should be considered preliminary, and require further validation and analysis.

5 Conclusions

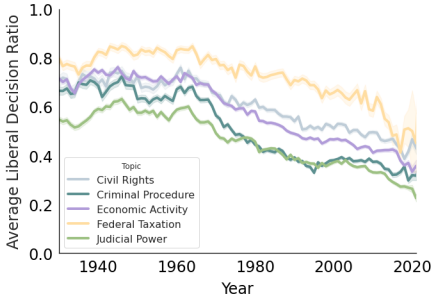
Out of the models implemented in this project, LEGAL-BERT achieved the highest accuracy on the Supreme Court test set for partisan classification and topic area classification. In addition, LEGAL-BERT was also fairly calibrated. Over all model types, accuracy on topic area classification was much higher than accuracy on partisan classification. After obtaining predictions from the



(a) Predicted average likelihood for a judge's decision to oppose appointing president's ideology over time.



(b) Predicted average liberal decision ratio over a judge's term.



(c) Predicted average liberal decision ratio for top 5 topics over time.

Figure 2: LEGAL-BERT lower court inference results.

best LEGAL-BERT model, preliminary results revealed a shift over time toward more conservative decision making in the lower courts.

The results provide insights into the challenges of the attempted classification task and shortcomings of the applied methodology. Legal texts pose difficulty due to their length and complex language, suggesting a need for models with exceptional capabilities in capturing intricate semantic relationships over longer sequences. To improve accuracy, future extensions could explore larger or different types of language models. Another challenge lies in discerning partisan direction from opinion text, which is a more abstract concept. To enhance future work, defining partisanship along a spectrum and incorporating Supreme Court judges' voting directions and unanimity in the training data could provide a more nuanced representation of partisan degree. Additionally, incomplete data for lower courts and assumptions about similarity between lower and Supreme Court data distributions hinder inference interpretation. Future approaches could address this by augmenting Supreme Court data with human-labeled samples from lower court data to align distributions more closely.

Overall, this project is a valuable first attempt at extracting coveted lower court metadata to discern trends in judicial partisanship and topics over time, and serves as a strong foundation for future work. These results lend to a better understanding of whether the lower courts are heavily influenced by partisanship, which has implications for realms beyond political science. Fair and impartial justice is a fundamental aspect of democracy and must be upheld to preserve the protection of civil liberties. Therefore, understanding the influences of partisan bias in the judicial system is crucial, and encourages democratic action to maintain the legitimacy and fairness of the United States government.

6 Lessons Learned

The main challenges that were encountered in applying our data science skills to a real-world problem were related to HPC resources and data quality. Limited knowledge of HPC necessitated understanding best practices, setting up environments, and optimizing configurations for efficient job execution. Learning NYU's HPC system enhanced our ability to leverage specific processors for a variety of tasks. Additionally, managing disparate and incomplete data from multiple sources required extensive preprocessing and filtering. Handling such "messy" data has equipped us to tackle similar data challenges in future projects.

7 Bibliography

bert-base-uncased · Hugging Face. https://huggingface.co/bert-base-uncased.

nlpaueb/legal-bert-base-uncased · Hugging Face. https://huggingface.co/nlpaueb/legal-bert-base-uncased.

pszemraj/long-t5-tglobal-base-16384-book-summary · Hugging Face. https://huggingface.co/pszemraj/long-t5-tglobal-base-16384-book-summary.

Text classification · Hugging Face. https://huggingface.co/docs/transformers/tasks/sequence_classification.

Albrecht, J., S. Ramachandran, and C. Winkler (2020). *Blueprints for text analysis using Python: machine learning-based solutions for common real world (NLP) applications* (First edition ed.). Sebastopol, CA: O'Reilly Media, Inc. OCLC: on1156041908.

Bonica, A. and M. Sen (2021, February). Estimating Judicial Ideology. *Journal of Economic Perspectives 35*(1), 97–118.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019, May). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs].

Free Law Project. CourtListener. https://www.courtlistener.com/help/api/bulk-data/.

Guo, M., J. Ainslie, D. Uthus, S. Ontanon, J. Ni, Y.-H. Sung, and Y. Yang (2022, May). LongT5: Efficient Text-To-Text Transformer for Long Sequences. arXiv:2112.07916 [cs].

He, H., L. Jiang, W. Yuan, X. Pan, Y. Kuang, and D. Rothermel (2023). NYU DS-GA 1011 Calendar and Course Content. https://nyu-cs2590.github.io/fall2023/calendar/.

MEHMET TEKMAN. LSTM Text Classification - Pytorch. https://kaggle.com/code/mehmetlaudatekman/lstm-text-classification-pytorch.

Noguti, M. Y., E. Vellasques, and L. S. Oliveira (2020, July). Legal Document Classification: An Application to Law Area Prediction of Petitions to Public Prosecution Service. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. ISSN: 2161-4407.

Song, D., A. Vold, K. Madan, and F. Schilder (2022, May). Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. *Information Systems 106*, 101718.

Spaeth, H. J., L. Epstein, A. D. Martin, J. A. Segal, T. J. Ruger, and S. C. Benesh. Supreme Court Database. http://Supremecourtdatabase.org.

Tunstall, L., L. v. Werra, and T. Wolf (2022). *Natural language processing with transformers: building language applications with Hugging Face* (First edition ed.). Sebastopol, CA: O'Reilly Media. OCLC: on1266359932.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin (2023, August). Attention Is All You Need. arXiv:1706.03762 [cs].

Wan, L., G. Papageorgiou, M. Seddon, and M. Bernardoni (2019, December). Long-length Legal Document Classification. arXiv:1912.06905 [cs].

Zheng, L., N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho (2021, June). When does pretraining help?: assessing self-supervised learning for law and the CaseHOLD dataset of 53,000+ legal holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, São Paulo Brazil, pp. 159–168. ACM.

8 Student Contributions

Annabelle Huether, Mary Nwangwu, and Allison Redfern worked on Exploratory Data Analysis, Data Pre-Processing, Data Merging, Baseline Model, BiLSTM Model Building, Analysis of Results, Poster, Report
Annabelle Huether implemented BERT/LEGAL-BERT
Mary Nwangwu implemented LongT5 Text Summarization
Allison Redfern implemented BiLSTM, Visualizations