Supreme Court Oral Arguments Outcome Prediction

Chen Hui Fei Wang, Reza Rizky Pratama, Shradha Ganapathy, Xiomara Salazar Flores
University of Chicago
chenhui1@uchicago.edu, rezapratama@uchicago.edu,
shradag@uchicago.edu, xiomara@uchicago.edu

Abstract

In this paper, we propose a novel approach using natural language processing techniques to predict the outcomes of US Supreme Court cases based on oral argument transcripts. The aim is to understand how justices' questions and comments might reveal their decision before it is officially issued. Our dataset was the Supreme Court Oral Arguments Corpus – more specifically, a subset across George W. Bush's and Barack Obama's administrations, which allowed for a comprehensive study across different political periods. Our model focused on the dialogues from cases that spanned 16 years under the administrations of Presidents George W. Bush and Barack Obama. Features considered included text utterances, the percentage of Republican justices, the percentage of male justices, and the case development time. We found that Logistic Regression performs better in this dataset compared to other models. Despite an imbalanced dataset, our model demonstrated satisfactory performance using F1 scores as the measure, particularly highlighting the predictive power of text utterances. This study faced limitations that opened avenues for future work, such as expanding the dataset and incorporating a wider array of features and advanced NLP methodologies.

1 Introduction

1.1 Background

The Supreme Court of the US (SCOTUS) is the highest court in the United States. It holds the power of judicial review – determining whether a statute violates the Constitution. The court consists of nine Justices (a chief justice and eight associate justices). Justices can serve for life (i.e., until they die, retire, resign, get impeached, etc.), so one Justice influences decisions for many decades. Clearly, the decisions made by the SCOTUS have a great public policy impact in the US.

In general, the cases that the SCOTUS review has been given a decision by a lower court and are brought to the Court for appeal of said decision. For these cases, the Court will take briefs and conduct an oral argument with the attorneys representing the (usually 2) parties in this case. Following the oral argument, each Justice gets one vote, and the majority vote determines the case outcome. Example SCOTUS case: https://www.oyez.org/cases/2018/17-204

In this project, we aim to predict the outcome of cases (i.e., which party gets the majority vote) through oral argument transcripts. The dataset contains dialogues of English natural language text. Our project is a text classification task, where the transcripts are the input documents. The labels we want to predict are the outcomes (e.g., Y/N, the petitioner "wins").

1.2 Use Cases

Predicting SCOTUS oral argument outcomes directly benefits a variety of stakeholders across multiple domains. Below are a few groups we believe these results are helpful for:

1. **Legal professionals and legal firms**: For legal firms, predicting the likelihood of winning a case informs how they structure their cases, select clients, or construct their arguments. From a commercial/business perspective, any legal firm with access to a high-performance model can raise business profits by soliciting more clients. By predicting case outcomes with transcripts, researchers and legal scholars can gain insights into the patterns, arguments, and reasoning that tend to sway the Court.
2. **Policymakers**: Our legal expert noted that these predictions are valuable when accounting for institutional interdependence and adaptation. For example, our expert highlighted that institutions could have begun preparations (such as updating forms) for the outcome of *Obergefell v. Hodges,* which guaranteed the right to marriage for same-sex couples. With advanced notice, other institutions can gain a head start for noteworthy cases (Dicey 1885).
3. **Political activists and businesses/investors**: The former would leverage this knowledge to target lobbying efforts, strategize how to influence court decisions, or adapt public mobilization efforts (from grassroots campaigns, public awareness initiatives, etc.). The latter could use anticipated changes to comply with new regulations and preempt stock market movement.
4. **Computer Scientists, Researchers, and ML Engineers**: This group would be interested in how the complexity of this data and its predictions aids the advancement of machine learning and natural language processing (text classification, sentiment analysis, language understanding).
5. **Broader Public**: Understanding how or whether the gender of advocates on either side, the gender of justices, the degree of partisanship of the court, the alignment with the political affiliation of the government, or the inclination of the amicus curiae influences the court decision-making could also unearth potential biases within the court.

1.3 Domain Research

In the field of Machine Learning, text classification models are some of the most nuanced – in recognition of the multi-faceted, ever-evolving nature of language itself. Consequently, datasets that are based on this fundamental nature of language serve as an interesting challenge for researchers. After our initial exploration through the data, we recognized that any future work would require us to have sufficient domain knowledge. To ask the right questions, we needed to have the right information. Beyond studying legal theories, our most useful source was our interviews with an expert in the field.

We reached out to Ms. Gabriela Rivera – who is an accomplished human rights lawyer – to ask about the dataset, the Supreme Court's structure, legal theories, hearings, and more. Through speaking with her, we were able to better identify model use cases, hypotheses to test, and features to explore through our model. We will be citing Ms. Rivera through this paper in acknowledgment of how her insight helped create a more meaningful model.

2 Data

2.1 Dataset

We leveraged the Supreme Court Oral Arguments Corpus from Convokit[1] to analyze data across the eight years of George W. Bush's administration and the eight years of Barack Obama's presidential administrations. These spanned from 2001 to 2017. Hence, we are studying 16 years' worth of case data. We leveraged **Senate Nomination records** to determine under which administration the Justices were appointed. Finally, we utilized a dataset from the University of California, Irvine, to predict the gender of Justices and advocates based on their first names.

2.1.1 Why did we select this subset of data?

Given that Justices on the US Supreme Court are political appointees, confirmed to their positions via the Senate, we wanted to evaluate outcomes across two different political periods. We settled on

[1] Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, Cristian Danescu-Niculescu-Mizil. 2020. "ConvoKit: A Toolkit for the Analysis of Conversations". Proceedings of SIGDIAL

choosing an administration under a Democratic President and another under a Republican President. We decided to select the Bush and Obama administrations. Both these administrations had fairly balanced courts in terms of political/legal leanings. Additionally, both Presidents Bush and President Obama appointed two justices each to the Supreme Court during their tenures, implying that the constitution of our courts remained fairly consistent with the same "number of changes." The choice of successive administrations was also important to reflect the fairly contiguous nature of the Supreme Court, where Justices serve until retirement or passing.

We decided to choose Presidents with two terms to see if there was a difference at play between how a court would vote between the first and second terms of a President and, more broadly speaking, whether the term had an impact at all. The long time span allowed us to also compare, for some justices, whether their votes differed between their times under the Bush and, subsequently, the Obama administration.

2.1.2 How did we clean the data?

To ensure a robust machine learning model, it was critical that the data we fed into our model was clean, structured, and devoid of any noise or unnecessary information. Our cleaning process for our labels and features was as follows: For feature, conversation text: the sequence of operations can be detailed as follows:

- **Combining into a single string:** case utterances were combined to ensure context preservation and represent each case as a unified narrative.
- **Stripping special characters and numbers:** the aim was to focus solely on words in the text and to reduce any potential confusion or bias induced by these non-alphabetic entities.
- **Removal of short words:** words with a length of less than three were eliminated from the text to ensure that only substantial words, likely to contribute to the overall meaning and context, were kept for subsequent processing.
- **Removing stopwords**: typically, the most common words in a language often do not contribute significantly to the semantic value of a text and thus were removed using the NLTK's stopwords list.
- **Stemming:** a text normalization technique that reduces words to their base or root form. This was done to decrease the size of the vocabulary of our text data and to amplify the signal by aggregating similar terms. We used stemming instead of lemmatization because it gave us shorter execution times while achieving almost the same result.

For the predicted label, win_side, we are interested to see the result for if petitioners win the case (1) or lose (0), with **inconclusive or deferred cases being pruned out.**

2.2 Data Analysis

We had the opportunity to consult with a legal expert whose insights have played a vital role in shaping our hypotheses and refining our research questions. This collaboration has added depth and credibility to our study on predicting SCOTUS oral argument outcomes. We focused on four core aspects to investigate the data set: justice and partisanship, gender influencing outcomes, correlations amongst voting records as well as any descriptive or additional data point that would help us understand cases.

2.2.1 Justices and partisanship

With the guidance of the legal expert, we explored how the political ideologies and affiliations of individual justices may influence their voting behavior, ultimately impacting case outcomes. By analyzing historical data, we aimed to uncover any interesting patterns or correlations that shed light on the interplay between justices and partisanship within the Supreme Court.

Key findings include:

- There are 25 justices during this time frame
- Around 75% of justices were appointed under Republican administrations
- Our top 5 Justices who voted in order from most to least were Ruth Bader Ginsbug, Clarence Thomas, Stephen Breyer, Anthony Kennedy, Antonin Scalia.

2.2.2 Gender on case outcomes

The legal expert's insights helped us develop hypotheses on how gender may affect the likelihood of achieving favorable outcomes in SCOTUS cases. Through rigorous data analysis, we sought to find significant associations between gender and case outcomes, contributing to our understanding of gender dynamics within the legal system.

Key findings that we found regarding gender on case outcomes:

- Around 80% of advocates were men
- The most prevalent names were David, John, Michael, Paul, and Thomas, but they made up less than .04% of total speakers.
- There appears to be no statistically significant relationship between the gender of advocates and case outcomes.

2.2.3 Voting records of individual justices

By studying historical voting data, we aim to understand the extent to which justices' voting patterns align or diverge. This analysis provides insights into the internal dynamics of the Court and helps us grasp the influence that individual justices may have on one another's opinions and decisions.

Key voting tendencies include that:

- Ruth Bader Ginsburg had similar voting tendencies as Sotomayor and Kagan
- Thomas had similar voting tendencies to Scalia and Alito

2.2.4 Additional descriptive and general findings

Finally, we explored several additional dimensions to gain a comprehensive understanding of the SCOTUS oral argument landscape. These included examining the number of justices in the data set, analyzing bi-grams extracted from case transcripts, studying cases by year, exploring wins per year, investigating wins per justice, and analyzing cases by month. These can be viewed in our **appendix**.

3 Methodology

3.1 Model Selection

Our baseline accuracy score was 64%. Our initial approach was to leverage a single feature – text — across 6 models in order to see which yielded the highest performance and was able to exceed our baseline accuracy. We chose the following models because:

1. **K-Nearest Neighbor.** Non-parametric method: as cases and their outcomes can be complex, this model allows us to account for the non-linearity of the relationship between text and labels (outcomes). Reason for selection: Simple, Versatile, Non-parametric
2. **Logistic Regression**. Interpretability of results: helps understand which features are more important when predicting outcomes (probabilistic). Reason for selection: Interpretable, Probabilistic, Efficient
3. **Random Forest**. Ensemble method that can handle non-linear relationships: useful as the relationship between outcomes and text is unlikely to be linear; can also help understand the importance of various features. Reason for selection: Robust, Accurate, Versatile
4. **Multinomial Naive Bayes**. Speed of algorithm relative to the size of data: this model is beneficial in text classification tasks. This model was selected as it allows for quick predictions on the large amounts of data stored in our transcripts (16 years' worth of conversations/utterances). Reason for selection: Scalable, Stable, Interpretable
5. **Linear SVC**. Robust to overfitting: this would allow us to consider a large number of features (as we were unsure at the beginning of which features were most important). Reason for selection: Effective, Versatile, Robust

6. **Perceptron**. Model simplicity: it served as our baseline – considering that we had previously cleaned/transformed our data to be linearly separable (1s and 0s). Reason for selection: Fast, Online, Simple

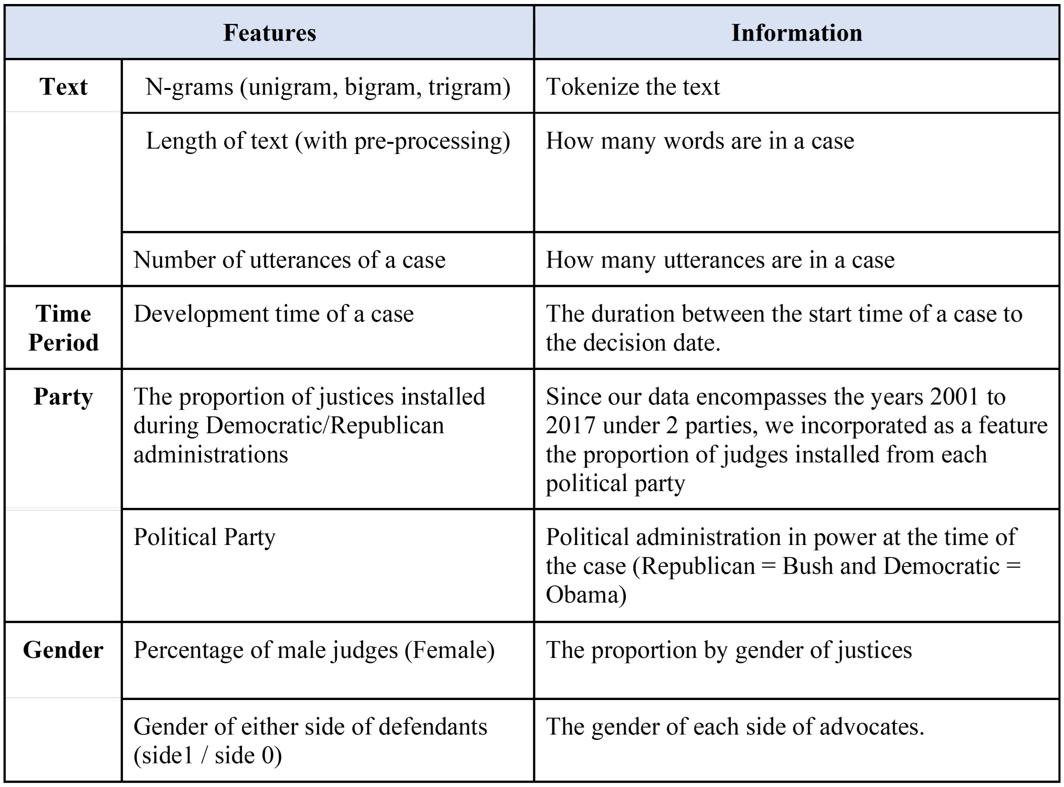
Besides that, we also attempted classification models utilizing BERT (Bidirectional Encoder Representations from Transformers). By considering a range of models, we gained a comprehensive understanding of their strengths and weaknesses, enhancing the robustness of our conclusions.

3.2 Features and Feature Engineering

For each of these models, we approached **text vectorization** by leveraging **CountVectorizer, TF-IDF, and DictVectorizer**. The goal here was to determine – with a singular, important feature (text) – which combination of text vectorization methods and hyperparameters would give us the best performance for the same subset of data. Note that for each of these models, the data was uniformly cleaned (in the same manner).

Unsurprisingly, we found that text vectorized using **TF-IDF with bigrams, on average, yielded higher performance (both accuracy and f1 scores)**. This is likely because our data was longer / more extensive and contained a lot of ‘unimportant’ words (words that did not contribute materially to our predictions).

While n-grams are a common feature when performing text classification, we also developed other features to explore all possibilities. Each of our predictions is based on **bigrams** of utterances. To this, we incorporated features like Text, Time Period, Party, and Gender. The details about our features are shown in Figure 1.



Features		**Information**
Text	N-grams (unigram, bigram, trigram)	Tokenize the text
	Length of text (with pre-processing)	How many words are in a case
	Number of utterances of a case	How many utterances are in a case
Time Period	Development time of a case	The duration between the start time of a case to the decision date.
Party	The proportion of justices installed during Democratic/Republican administrations	Since our data encompasses the years 2001 to 2017 under 2 parties, we incorporated as a feature the proportion of judges installed from each political party
	Political Party	Political administration in power at the time of the case (Republican = Bush and Democratic = Obama)
Gender	Percentage of male judges (Female)	The proportion by gender of justices
	Gender of either side of defendants (side1 / side 0)	The gender of each side of advocates.

Figure 1: Feature Selection

4 Results and Analysis

4.1 Model Performance

How did we select our performance metric (accuracy v precision v recall v F1)?

Our project uses F1 scores as our measure of performance. Our sample was imbalanced, with "yes" votes outweighing "no" votes by 75%. In other words, one of our classes was more prevalent than the other. Additionally, when evaluating features such as gender, we found that our dataset's advocates were 80% male (v. female).

Using accuracy scores would be misleading as the model would be able to achieve high accuracy merely by predicting the majority class – which would leave us unable to correctly identify potential losses / the minority class. Cases that are won and those that are lost are equally important to legal professionals, policymakers, activists, and industries. Hence, we would not be able to select solely precision or recall either.

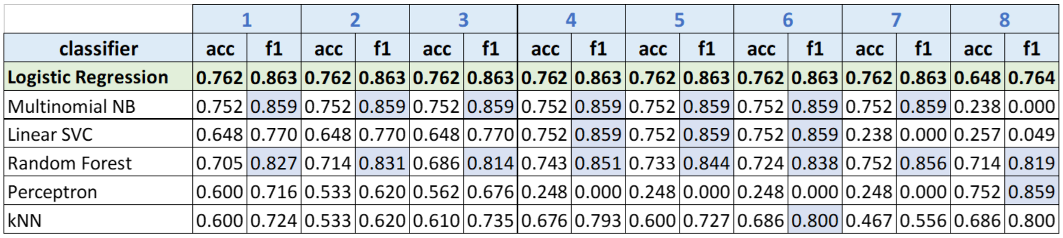
The F1 score allowed us to combine/balance precision and recall in a single value; we could, thus, identify our best-performing model despite class imbalances. It takes into account both false positives and false negatives – identifying both wins and losses – lending to a comprehensive evaluation of model performance.

Testing the Model

We tested different variations of the models. Each column below corresponds to different features included in the model:

- 1 = utterances (TFIDF Bigram)
- 2 = utterances + rep_pct (%that is republican)
- 3 = utterances + male_jpc (% of justices that are men)
- 4 = utterances + develop_time (timeframe of the case)
- 5 = utterances + rep_pct + male_jpc
- 6 = utterances + male_jpc + develop_time
- 7 = utterances + rep_pct + male_jpc + develop_time
- 8 = utterances + rep_pct + male_jpc + develop_time with oversampling

The result of this variation of the models is as shown in Figure 2 and Figure 3. Using the Logistic Regression model, the coefficient for rep_pct, male_jpc and develop_time are .012, -.082, -.0004. This shows that the other variables besides the utterance are weak predictors for this analysis.



	1		2		3		4		5		6		7		8	
classifier	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**	**acc**	**f1**
Logistic Regression	**0.762**	**0.863**	**0.762**	**0.863**	**0.762**	**0.863**	**0.762**	**0.863**	**0.762**	**0.863**	**0.762**	**0.863**	**0.762**	**0.863**	**0.648**	**0.764**
Multinomial NB	0.752	0.859	0.752	0.859	0.752	0.859	0.752	0.859	0.752	0.859	0.752	0.859	0.752	0.859	0.238	0.000
Linear SVC	0.648	0.770	0.648	0.770	0.648	0.770	0.752	0.859	0.752	0.859	0.752	0.859	0.238	0.000	0.257	0.049
Random Forest	0.705	0.827	0.714	0.831	0.686	0.814	0.743	0.851	0.733	0.844	0.724	0.838	0.752	0.856	0.714	0.819
Perceptron	0.600	0.716	0.533	0.620	0.562	0.676	0.248	0.000	0.248	0.000	0.248	0.000	0.248	0.000	0.752	0.859
kNN	0.600	0.724	0.533	0.620	0.610	0.735	0.676	0.793	0.600	0.727	0.686	0.800	0.467	0.556	0.686	0.800

Figure 2: results across various features

In an attempt to get better results, we utilized the BERT model for sequence classification. This was executed utilizing both 'distilbert-base-uncased' and 'bert-base-uncased'. To fine-tune these models, we employed the AdamW optimizer with a weight decay of 0.01, set the model to save after every 1000 steps, and conducted it over 10 epochs. However, even after this rigorous testing and fine-tuning, the performance of the BERT models was found to be inferior to that of the logistic regression model. This might be because the BERT models might require more extensive tuning, including adjustments to learning rates, the optimizer, or the weight decay.

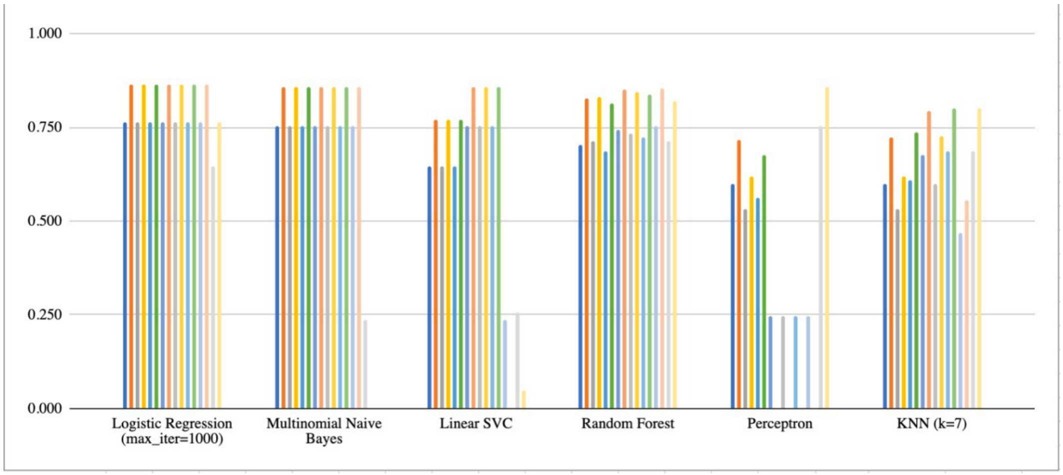


Figure 3: chart comparing results across various features

We ultimately settled on Logistic Regression for our model due to higher accuracy and F1 scores and interpretability of results, which is important when considering the legal, political, social, and economic consequences of our predictions/outcomes.

We considered text utterances, the percentage of Republican appointees, the percentage of male judges, and development time. Surprisingly, we obtained identical results when using only a single feature as compared to when using all four features combined. This unexpected outcome suggests that the predictive power of text utterances is so strong that it overshadows the influence of the other features.

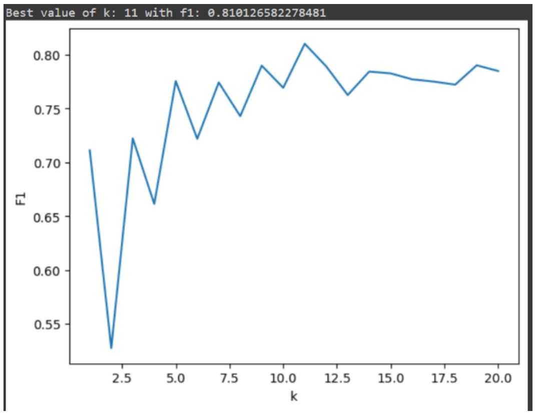
4.2 Treating Imbalance and Hyperparameter Tuning

In our study, we experimented with both undersampling using RandomUnderSampler and oversampling using SMOTE techniques to address the class imbalance in the dataset. However, contrary to our expectations, we observed that neither undersampling nor oversampling led to improved results compared to not using any sampling techniques. We would need to investigate these results in future work further. This could be due to the unique characteristics of the dataset involving language – particularly its complexity (especially when considering legal arguments). Additionally, it may be that RandomUnderSampler and SMOTE techniques may not have been suitable for this NLP task.

In order to find optimal hyperparameters, we performed tuning for our kNN and Logistic Regression models. We fine-tuned 'k' (the number of nearest neighbors the model considers when making a prediction) in our K-Nearest Neighbors model to arrive at the optimal value. Using F1 scores to measure model performance, it was found that a 'k' value of 11 provided the highest F1 score, suggesting it was the optimal choice. The best K value is as shown in Figure 4

For our logistic regression model, we ran grid search to explore multiple permutations of penalty parameters ('l1', 'l2', and 'None'), solvers ('liblinear' and 'lbfgs'), and regularization strengths ('C' values of 0.1, 1, and 10). We evaluated the performance of various combinations using the F1 score across 10-fold cross-validation. These results are visualized in Figure 5. Remarkably, we found that three combinations yielded the same F1 score: (1) an 'L2' penalty with a 'C' value of 0.1 and 'lbfgs' solver, (2) an 'L2' penalty with a 'C' value of 1 and 'lbfgs' solver, and (3) an 'L1' penalty with a 'C' value of 1 and 'liblinear' solver. These combinations offer the most effective compromise between model complexity and predictive performance in our Logistic Regression model.

In general, the 'L2' penalty seems more suitable across regularization strength. 'L2' regularization tends to distribute error among all the terms (more stable solution), which might be the reason why it performs better. Better performance on higher regularization strength suggested that our model might be prone to overfitting without this regularization.



```
Best value of k: 11 with f1: 0.810126582278481
```

Figure 4: Best parameters for kNN

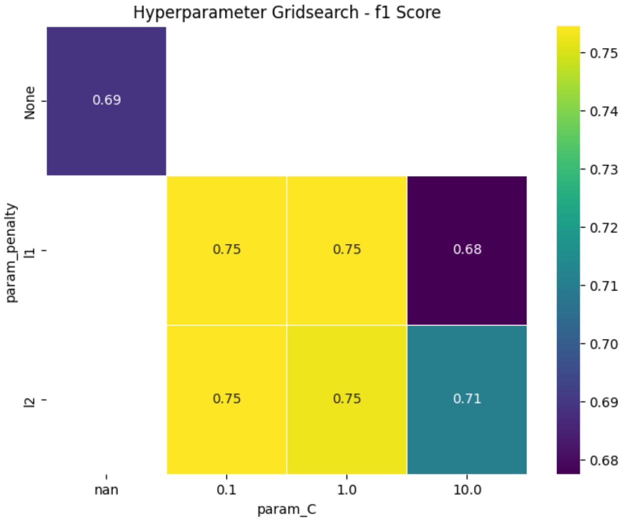


Figure 5: Best parameters Logistic Regression

5 Limitations and Future Work

Working on a complex NLP project such as ours did mean that we encountered uphill challenges throughout. These are detailed below and should provide context as to why and how our model functions as it does:

1. **Imbalanced Dataset:** As mentioned above – despite our intentions while subsetting it – we, unfortunately, found that our dataset was imbalanced. The majority class far outweighed the minority, resulting in the model's accuracy being higher merely by predicting 'Yes' rather than because of the model's construction itself. This also meant that we may have missed potentially relevant data points and suffered from concerns about generalizability. Our restricted, unbalanced dataset may be constrained when applied to contexts outside our timeframe. This resulted in us having to adjust what performance metric we were using, having to oversample and swap 1s and 0s in order actually to test the model's performance. **Future Work:** due to the paucity of time and resources, we had to restrict our dataset and limit the features we were incorporating into our model. We want our future research to work across an expanded dataset (be it a longer time scope/with additional data sources) and consider features such as the influence of amicus curiae briefs on outcomes, changing outcomes during the second terms of presidents v. their first terms, etc.

2. **Changing Dynamics:** This also returns to the generalizability concern. Legal interpretations and precedents evolve over time, influenced by changing societal, political, and cultural factors. By confining our analysis to a specific timeframe, we may not fully capture the temporal dynamics and shifts in legal ideologies and decision-making patterns. This may limit the extent to which our findings can be applied or extrapolated to different time periods. **Future Work:** In the future, we would like our model to identify and quantify biases in case outcomes, legal arguments, or legal language to evaluate institutional fairness and structural inequalities to include more political and cultural factors.
3. **Limited NLP Toolkit:** as we hadn't covered NLP topics in class, we constantly had to resolve blockers arising from limited NLP/text classification knowledge and leverage a variety of resources. With a project as complex as this and with as limited a time frame as a quarter, this meant that our model may not have performed as well as we would have liked. **Future Work:** Building a tool to offer predictions and construct visualizations that can be leveraged by media, legal professionals, policymakers, activists, and businesses.

5.1 Learnings

From our limitations came our learnings. One of the biggest learnings we encountered through this project reflected what Dr. Anna Zink mentioned in her presentation on the 16th of May, 2023. That **domain knowledge is an asset** in building meaningful models that address on-the-ground models. Having access to a legal professional helped us better understand the data and finetune our areas of interest. It helped us ask insightful questions during our data exploration and identify features of importance when fine-tuning our model. In all, it contributed vastly to creating a more generalizable model.

Additionally, our limited toolkit meant that we were learning much of the technical skillset needed for this project on the go. As it was our first time working on NLP tasks, we found it quite challenging but did grow a lot. Some lessons we took away included performing **more in-depth analysis when identifying our subsets to avoid imbalanced data**. We also learned of the importance of leveraging confusion matrices to identify where specifically our model was failing; previously, we just reviewed high-level statistics, but in projects such as this, granularity should also be considered. Another interesting technique learned through this project was to **swap 1s and 0s** if our dataset favored 1s – this would let us confirm if our model's performance was achieved merely by predicting the majority class or if it was due to the abilities of the model itself.

In conclusion, this project provided valuable learnings and insights through the identification and understanding of its limitations. We recognized the immense value of domain knowledge and the impact it has on building meaningful and applicable models. Additionally, the project highlighted the importance of continuous learning and acquiring technical skills specific to the task at hand as we navigated the challenges of working with NLP tasks for the first time. Overall, this project provided a valuable learning experience and laid the foundation for future advancements in this field.

6 Conclusion

Our project endeavored to predict the outcomes of U.S. Supreme Court cases based on oral argument transcripts. We exploited text classification methods, specifically utilizing logistic regression, and applied techniques to tackle the issue of class imbalance within our dataset. Our exploration spanned 16 years' worth of case data from two different presidential administrations, and our model considered multiple features such as text utterances, percentage of Republican justices, percentage of male justices, and case development time.

Interestingly, our Logistic Regression model with 'lbfgs' solver, 'L2' penalty and higher regularization strengths outperformed others, demonstrating a high F1 score. It became apparent that utterances carried such significant predictive power that they overshadowed the influence of the other features. Despite our initial assumptions, neither undersampling nor oversampling improved our results, suggesting the need for future investigation into suitable techniques for balancing class distributions in such NLP tasks.

However, our study also revealed several limitations. The imbalanced dataset and the focus on a specific timeframe constrain our model's generalizability. The evolving nature of legal interpretations, as influenced by societal, political, and cultural changes, is another factor that could limit our model's future applicability. Finally, we faced challenges associated with our limited exposure to NLP techniques.

These insights, however, have set the stage for future research. Expansion of the dataset and consideration of additional influencing factors such as amicus curiae briefs, political and cultural dynamics, and structural inequalities could enhance the model's performance and generalizability. Our study also highlighted the value of domain knowledge in building accurate and meaningful models, an aspect that will guide our future work.

This project served as an exciting exploration of the intersection of law and machine learning. Through our research we learned that with continued refinement and development, machine learning models have the potential to offer significant contributions to legal analytics. Despite the limitations and complexities encountered, our model's performance brings optimism for the predictive potential in legal cases. By further expanding our research in this area and refining our methodologies, we are looking forward to contributing more to the rich landscape of legal prediction using machine learning tools.

Acknowledgments

We would like to express our appreciation to Karen Zhou, our Lead TA, for her guidance and support throughout this project. Her feedback was integral to our progress and the successful completion of this work. We would also like to thank our professor, Chenhao Tan, for his valuable instruction and insights. Their assistance was key to the success of our project, and we are grateful for their contributions. Furthermore, we want to extend our gratitude to Gabriela Rivera whose legal expertise was necessary to even kick-start this process and to ensure that we were asking the right questions. We are grateful for her patience and for extending us her time.

Project Code

The project code is at https://github.com/rezarzky/supreme-prediction

References

The project code is at https://codi-gen.github.io/

[1] A.V. Dicey, The Law of the Constitution (1885) (Oxford University Press edition, ed. J.W.F. Allison, 2013, p. 69).

[2] Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, Cristian Danescu-Niculescu-Mizil. 2020. "ConvoKit: A Toolkit for the Analysis of Conversations." Proceedings of SIGDIAL.

[3] Rules of the Supreme Court of the United States (2019). RULE 28: ORAL ARGUMENT