Predicting Supreme Court Case Outcomes

Baraa Zekeria
DSC 161: Text as Data

March 11, 2024

Research Question

- **Research Question:** *Can we predict the outcome of Supreme Court cases using the opinions, concurrence, and dissent in the court's decisions?*
- Goal $\Rightarrow$ Understanding if the **language** used in court opinions, concurrence, and dissent can predict case outcomes **has significant implications for legal practice and judicial decision-making**.

$$\text{win\_side} = f(\text{justia\_section})$$

Data Source

- Data sourced from ”*Super-SCOTUS: A multi-sourced dataset for the Supreme Court of the US*” by Biaoyan Fang et al.
  - Aims to address complexity of US Supreme Court judiciary by integrating various procedural phases and resources.
  - Provides a comprehensive dataset connecting language documents with extensive metadata.

Data Description

- Supreme Court cases from 2010 to 2015
  - *Why?* Justices are the same
- 

Figure: justia_section Word Cloud

- The feature of interest is the `justia_sections` variable, encompassing opinions, concurrences, and dissents.

Data Cleaning and Text Preprocessing

- Parsed dictionary-like strings into dictionaries for relevant columns.
- Cleaned and preprocessed text data by removing stopwords, lemmatizing, and removing numbers.

Modeling: TF-IDF Vectorization

- Split the dataset into training and testing sets using an 80-20 split.
- Applied TF-IDF vectorization to the textual content in `justia_sections` (44K+ features)

TF-IDF Formula:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

where:

- t represents a term (word) in the document
- d represents a document
- D represents the set of all documents
- $\text{TF}(t, d)$ is the term frequency of term t in document d
- $\text{IDF}(t, D)$ is the inverse document frequency of term t across all documents in D

Modeling: TF-IDF Vectorization

Figure: justia_section Word Cloud

Modeling: Classification

- Used **Naive Bayes** and **Logistic Regression** models for predicting the case outcomes based on the features extracted from `justia_sections`.

Results

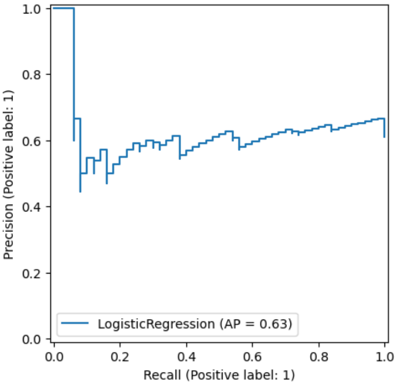
- Utilized cross-validation to choose the best model, Logistic Regression
  - Limitation of Naive Bayes: *naively* assumes each word is independent
- Achieved an **accuracy of 62%** on the test set with the best model.
- 

Figure: Logistic Regression Percision-Recall Curve

Limitations

- **Limited Feature Set**: reliance solely on textual content from `justia_sections` may overlook valuable metadata
- **Data Imbalance**: imbalance between classes ("affirmed" vs. "reversed") impacted model performance
- **Simplified Models**: oversimplify complex relationships
  - **Naive Bayes**: assumption of independence between words, leading to potential loss of context and meaning.
  - **Logistic Regression**: linear decision boundaries may struggle to capture nonlinear relationships present in legal texts, limiting the model's ability to discern subtle patterns
- **Bias and Fairness**: Risks of bias exist in the dataset and models, requiring careful mitigation for fair predictions.
- **Generalizability**: Model applicability beyond the dataset and timeframe needs validation for real-world use.

Next Steps

- Possibly include additional features related to case metadata such as oral arguments, amicus briefs, petitioner, respondent, etc.
- Consider incorporating more advanced natural language processing techniques