Analyzing U.S. Case Law with NLP and Data Visualization - Team 21

Paul Hutchison, Emre Duman, Bipin Koirala, Hung-Yu Shih, Mahdi Ghanei

Georgia Tech

1. Introduction

- Legal research is a critical part of the justice system. Lawyers and judges rely on past cases to make decisions — but reading through millions of legal documents is slow, manual, and inefficient.
- With over 6.7 million U.S. court cases available, traditional keyword searches fall short. They miss deeper patterns, relationships, and trends hidden in complex legal texts.
- Our goal: To make legal research smarter, faster, and more interactive using data analytics and visualization. We developed a system that helps users explore case law with natural language processing (NLP), network graphs, and predictive modeling — turning text into insights.

2. Dataset

We used **Caselaw Access Project** dataset, available directly at case.law or via HuggingFace at https://huggingface.co/datasets/free-law/Caselaw_Access_Project

It contains 6.7 million U.S. court cases from 1658-2019. Total dataset size is approximately 300 GB and is processed sequentially due to memory limits. The data is temporal in that each case includes a year field. Overall, the data contains 23 different fields for each case such as opinion text, judge name, etc.

We did different analyses on different sections of our dataset to make the computational costs manageable. This includes numeric metrics for all the states, text analysis for all Georgia cases, and winning party analysis for all the Alaska cases.

3. Our Approach

Our key contributions include a combination of machine learning and interactive visualizations:

Similarity Search

- We compute semantic similarity between legal cases using **sentence embeddings** from **Legal-BERT** and **BGE**. To stay efficient, we embed only the first **512 tokens** per case; enough to capture key context.
- We apply K-Nearest Neighbors (KNN) with Euclidean distance to find related cases. To handle large-scale data, we use **FAISS**, enabling fast and scalable similarity search.

Case Categorization & Complexity

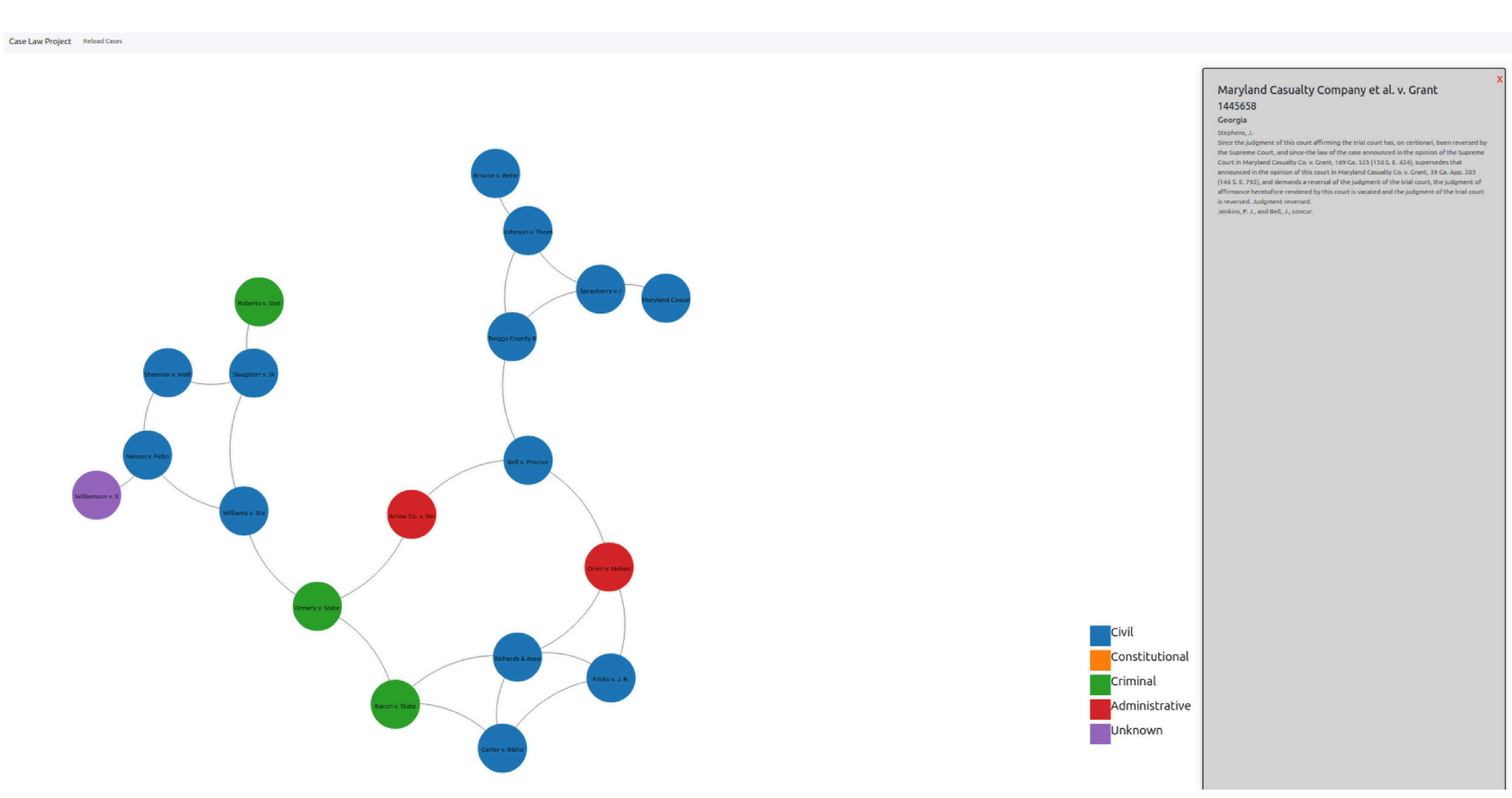
- We use Google-hosted LLM (Gemini) to classify each case as **criminal, civil, administrative,** or **constitutional** based on the case opinion text in the dataset.
- The same LLM estimates case complexity on a scale of 1 (simple) to 10 (complex). This helps us analyze patterns across case types and complexity levels.

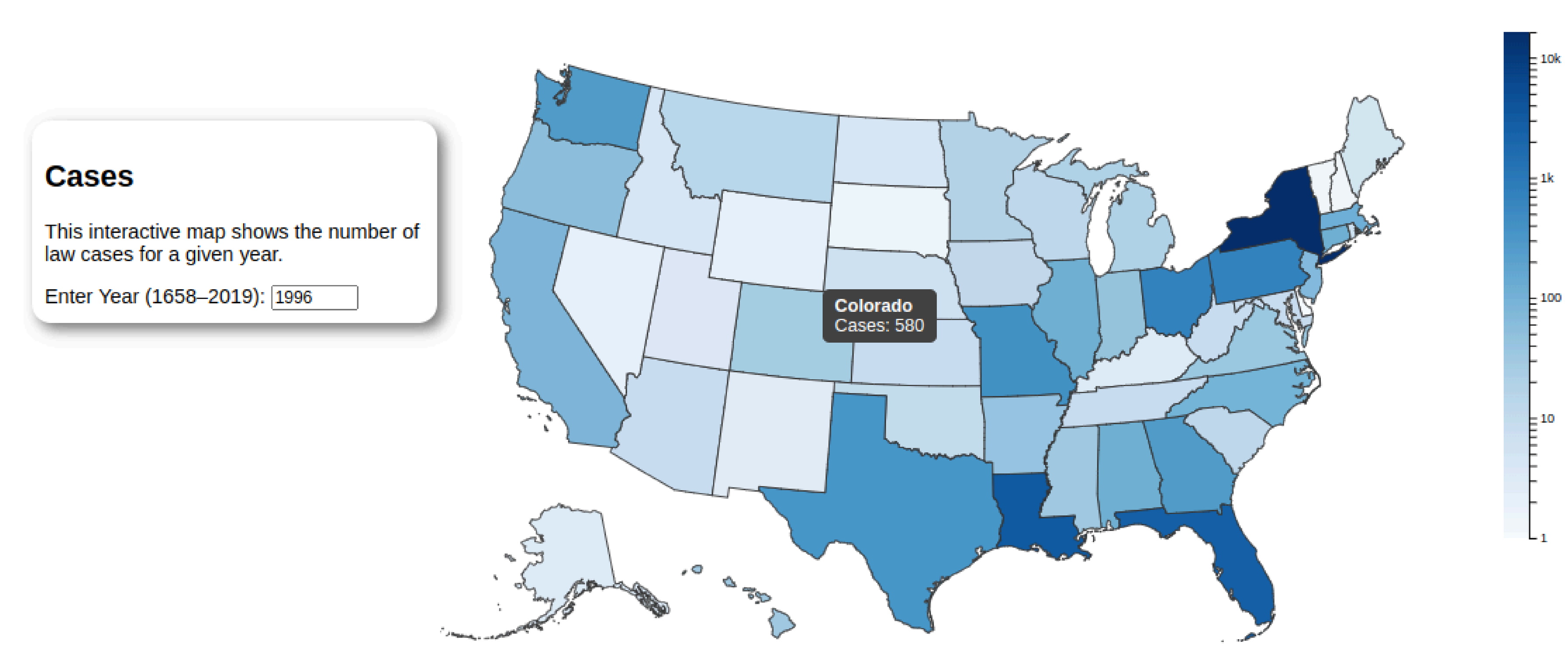
Winning Party Prediction

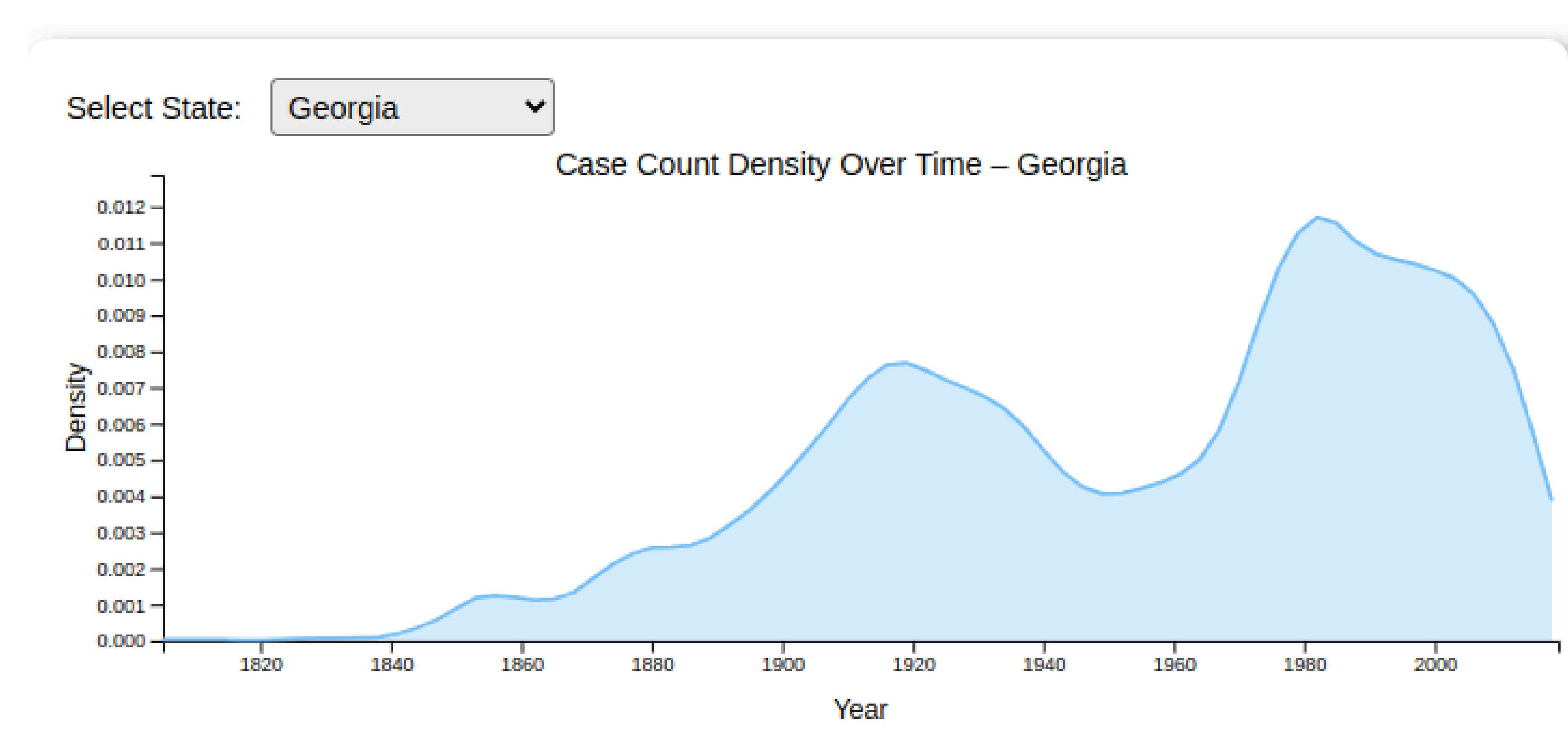
- We used **889 Alaska cases** to extract the "FACTS" and "CONCLUSION" sections to label each case as "**appellant**" or "**appellee**" using a pre-trained Legal-BERT model.
- After labeling, we removed these sections and trained:
  - GRU classifier with DistilBERT embeddings
  - Transformer encoder for comparison
- We used Weighted-Random-Sampler to handle class imbalance and reported metrics (Accuracy, F1) for both models.

Interactive Visualizations

- **Network Graphs:** We visualize semantic links between cases using similarity search. Each node is a case; edges connect k-nearest neighbors based on Legal-BERT/BGE embeddings. Users can:
  - Select a "seminal" case to view its neighbors
  - Filter by **case type** or **winning party**
  - Click nodes to view **case ID, title, or opinion**
- **Choropleth Map:** We mapped case volumes by U.S. state and year. The final choropleth allows users to visually compare case density across jurisdictions and time.
- **KDE Plot:** To show how case counts evolve over time, we create Kernel Density Estimation (KDE) plots for each state.
  - We used *Epanechnikov* kernel with fixed bandwidth *h=10*
  - Users choose a state from a dropdown to view its trend

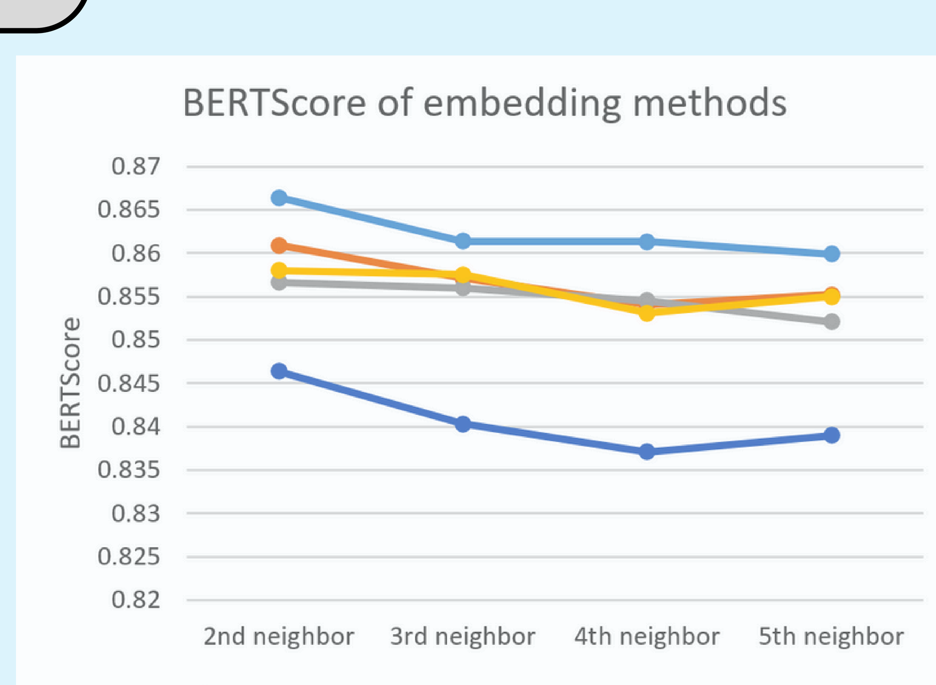






4. Experiments & Results

- Similarity search is evaluated using the average BERTScore over a fixed query set
- The Legal-BERT and BGE embedding outperformed the traditional baselines (BOW, TFIDF, Doc2Vec)
- 58% of cases are civil, 33% are criminal, 6% are administrative, 2% constitutional.
- The GRU–Bidirectional model achieved an accuracy of 81% with a macro F1 score of 47%. Transformer Encoder model achieved a slightly lower accuracy of 74% while obtaining a higher macro F1 score of 51%



Method	1st neighbor	2nd neighbor	3rd neighbor	4th neighbor	5th neighbor	Avg
Legal-BERT	1	0.8664	0.8614	0.8613	0.8599	0.8898
BGE	1	0.8609	0.8571	0.8541	0.8552	0.88546
BOW	1	0.8566	0.856	0.8545	0.8521	0.88384
TF-IDF	1	0.858	0.8575	0.8531	0.855	0.88472
Doc2Vec	0.9934	0.8464	0.8403	0.8371	0.839	0.87124

Model	Sentence Embedding	Accuracy	Macro F1 Score
GRU – Bidirectional	DistilBERT	81%	47%
Transformer Encoder	DistilBERT	74%	51%

5. Conclusion

- Successfully integrated NLP, ML, and visualization to analyze U.S. case law, significantly enhancing traditional legal research methods.
- Legal-BERT and BGE embeddings outperformed conventional approaches (BOW, TF-IDF, Doc2Vec), highlighting the advantage of domain-specific NLP models.
- Developed interactive network visualization, enabling intuitive exploration of semantic relationships between cases.
- Choropleth maps and KDE plots effectively visualized geographic and temporal trends in legal cases, providing accessible insights.
- Predictive models (GRU-Bidirectional and Transformer Encoder) showed good accuracy for winning party predictions, though imbalanced classes presented ongoing challenges.
- Future improvements include refining predictive models for better class sensitivity, expanding data coverage, and further optimizing visualization techniques.