Legal Text Classification - Project Summary

Problem Statement:
The goal of this project is to automatically classify U.S. Supreme Court legal opinions into predefined categories using Natural Language Processing (NLP) and machine learning techniques. Manual classification is time-consuming and inconsistent; this project provides an automated, scalable solution.

Approach:
1. Data Collection:
   - Supreme Court Database (SCDB 2025 Release 01) containing case opinions and metadata.
2. Data Preprocessing:
   - Clean text: remove punctuation, lowercase, remove stopwords.
   - Tokenize and prepare text for modeling.
3. Feature Extraction & Topic Modeling:
   - Use TF-IDF vectorization to represent text numerically.
   - Apply Latent Dirichlet Allocation (LDA) to extract main topics.
4. Classification:
   - Train a Logistic Regression (LR) model to classify cases into categories.
   - Optionally, compare with BERT embeddings for improved accuracy.
5. Evaluation:
   - Metrics: Accuracy, Precision, Recall, F1-Score
   - Visualizations: Confusion matrix, performance plots

Implementation Overview:
- Jupyter notebooks implement each step:
  1. Data preprocessing
  2. Topic modeling
  3. Classification
  4. Evaluation
- Models and vectorizers are saved for reproducibility.
- Results show effective classification and highlight potential improvements.

Conclusions / Challenges:
- Logistic Regression provides a simple baseline with decent performance.
- BERT embeddings improve semantic understanding but require more resources.
- Main challenge: cleaning and preprocessing large text data efficiently.

Future Work:
- Explore additional deep learning models (e.g., RoBERTa, LegalBERT).
- Implement multi-label classification for cases spanning multiple categories.