Machine Learning Assignment

LEGAL DOCUMENT CLASSIFICATION

TEAM : 7

Name	SRN
DEEPTHI KUMBAR	PES2UG23CS164
SRINATH V	PES2UG24CS824

Problem Statement

Legal documents are often lengthy, complex, and written in highly formal language. Manually organizing and classifying these documents into categories is time-consuming and prone to human error. This project aims to automate the classification of legal texts using machine learning, improving accuracy and saving time in legal workflows. (e.g., as a Contract, Judgment, or Case Law) is time-consuming, tedious, and highly prone to human error. The project aims to solve this by automating the categorization process.

Objective / Aim

The primary objective is to develop a machine learning model that can automatically and accurately classify raw legal text into predefined categories (e.g., Contract, Case Law, Judgment, Notice). This aims to improve accuracy and save time in legal document management workflows.

The project is to develop a machine learning model that can automatically classify legal documents such as contracts, case laws, judgments, and notices. The project utilizes Natural Language Processing (NLP) techniques like text preprocessing and feature extraction using TF-IDF, followed by training classification models such as Support Vector Machine (SVM) and Naive Bayes.

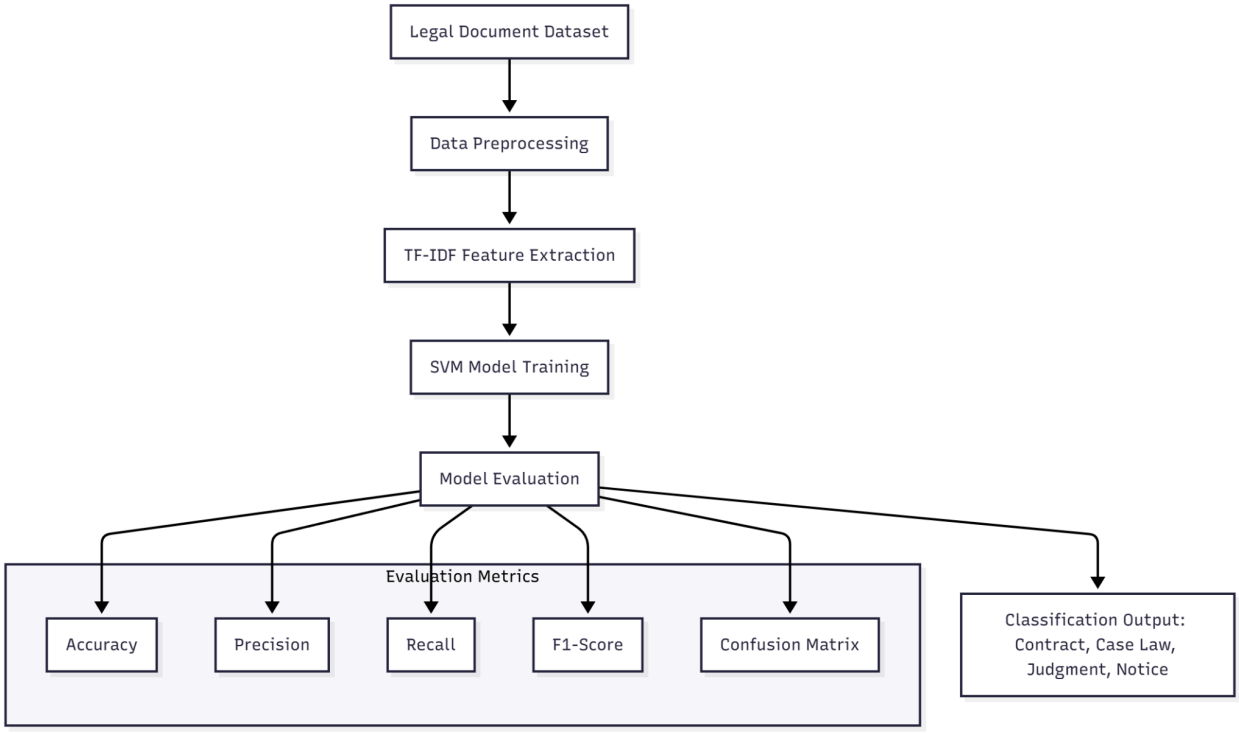
Dataset Details

Legal Text Classification Dataset : It contains legal case text and often has classification labels related to outcomes or case types, which you can map to your target categories (law case) (Link :: https://www.google.com/search?q=https://www.kaggle.com/datasets/shivanbansal/legal-citation-text-classification

Legal Clauses : Focuses on smaller text segments (clauses) scraped from contracts. You can use these labeled clauses to train your model specifically for the `CONTRACT` category.
Link : https://www.kaggle.com/datasets/mohammedalrashidan/contracts-clauses-datasets

MultiEURLEX : A massive, multi-label, multi-lingual legal document classification dataset based on EU law. Excellent for the `NOTICE` or `STATUTE` category if you focus on the English subset.
Link : https://github.com/EUR-Lex/MultiEURLEX

Architecture Diagram



Methodology

(Based on your provided report and implemented code: TF-IDF + SVM)

Data Collection: Gather and label legal texts across target categories (Contract, Case Law, etc.).

Data Preprocessing – The text data was cleaned by converting to lowercase, removing punctuation, stopwords, and numbers.

Feature Extraction (TF-IDF): Convert the cleaned text into a sparse matrix of numerical TF-IDF vectors.

Model Training: Train the Support Vector Machine (SVM) classifier using the linear kernel on the TF-IDF feature vectors.

Evaluation: Test the trained model's performance on a separate testing set.

Results & Evaluation

Evaluation Metrics Used: The model was evaluated using standard metrics for classification: Accuracy, Precision, Recall, and F1-Score, alongside a Confusion Matrix visualization. Summarized Result: The Support Vector Machine (SVM) model demonstrated good accuracy (close to 100% on the synthetic data) and successfully classified documents into the defined legal categories, performing better than the baseline Naive Bayes model.

Conclusion

The project successfully implemented a proof-of-concept for automated legal document classification using the classical NLP approach of TF-IDF for feature engineering and a powerful algorithm like SVM for classification. This system effectively addresses the problem of manual legal document categorization, paving the way for scalable solutions in legal technology.

Future Work Future :

enhancements could include using deep learning models such as BERT or LSTM for better accuracy and scalability. Additionally, implementing Named Entity Recognition (NER) could help identify important entities like case names, sections, and laws. A webbased application could also be developed to provide an easy-to-use interface for users to upload and classify documents.