2024

JP Morgan classification for legal documents project

USING CRISP-DM

YEMPARA BHUVANAIKA

|

J P Morgan classification for legal documents

1.Bussiness Understanding

❖ Business Objective:

The goal is to automate the classification of various legal documents using an AI-driven approach which is equivalent to JP Morgan's COIN. The main objective is that the system should classify documents with a high level of accuracy, improving efficiency, reducing the manual classification and also the system should minimize the errors in processing of the document.

❖ Accessing the situation:

This will help us to understand the starting point and defining the objectives and goals and understanding the scope of the project clearly. We need to determine the types of legal documents involved, the challenges faced such as volume, legal implications, complexity.

❖ Determining data mining goals:

Build a model using data mining techniques that can accurately classify legal documents into predefined categories (e.g., contracts, agreements, disclosures, etc.) with high accuracy. Develop a model that maintain scalability, high standards of accuracy and compliance.

❖ Project Plan:

Perform tasks such as data collection, data preparation, model development, and model evaluation. Set timelines, identification of contributors (e.g., data scientists, legal advisors.)

2.Data Understanding

❖ Data Collection:

Collect the data and explore the data to understand the hidden patterns and features present in legal documents

Gather the initial data from databases, file systems, websites or other sources (e.g., commercial loan agreements, credit-default swaps, and custody agreements.)

❖ Data Description:

Describe the data by understanding the format (e.g., PDFs, Word documents), Data attributes (Features) (e.g., Metadata, Content Features, Structural Features). Data quality considerations (e.g., Completeness, Consistency, Accuracy, Redundancy.) Data volume and storage.

❖ Data Exploration:

- **Overview of the Dataset**

  Conduct exploratory data analysis to identify key features that differentiate document types (e.g., contracts, legal briefs.) and document length to examine the terms if word count and page count.

- **Content Analysis and Structural Analysis**

  Textual data exploration (e.g., Word Cloud, Keyword Frequency.). Formatting Patterns (e.g., use of bullet points, tables, headers.). Use clustering techniques to visualize patterns in the data.

❖ Data Quality Verification:

Check for data quality issues such as incomplete documents, inconsistent labeling, accuracy verification, Duplication checking, Data integrity verification.

3.Data Preparation

❖ Data Selecting:

Prepare the dataset for modeling by cleaning, transforming, and structuring the data. Select the appropriate text features, metadata and structural elements for classification.

❖ Data Cleaning:

Data preprocessing steps should be performed. Clean the data by handling missing values, examine the data for inconsistencies, duplicates, and errors. Remove noise from the data (e.g., headers, footers and advertisements.). Normalize the text (e.g., convert text to lowercase, remove punctuation.). Perform tokenization, stop words removal and stemming.

❖ Data Integration:

Integrate preprocessed data into a single file or single dataset, combining all the text features, document metadata and any other relevant features.

❖ Data Formatting:

Reduce the dimensionality of the dataset. Apply techniques like PCA or LDA if the feature space is too large, to simplify the model and improving performance. Perform Label encoding which is converting categorical labels into numerical values in the dataset.

4.Modeling

❖ Model selection:

Choose suitable classification algorithms, such as Support Vector Machines, Naïve Bayes, Logistic Regression, Random Forests, or deep learning models like RNNs, CNNs, BERT, GPT, considering the nature of the text data in the dataset.

❖ Test Design:

- **Splitting data:**

  Splitting the data as training and testing data performing some methods like cross-validation techniques to ensure that the model is well trained and generalizes effectively and accurately to unseen data.

❖ Model Building:

Train the selected model with the preprocessed dataset, tune with hyperparameters to get optimal performance and high accuracy of the model.

❖ Model Assessment:

Evaluate the model using metrics such as accuracy, precision, recall, confusion matrix, F1-score will help model is performing in classifying legal documents and help to identify areas for improvement.

5.Evaluation

- **Evaluating results:**

  Evaluate model’s performance to ensure it meets business goals. Compare the model’s predictions with the actual classifications to check whether the model is performing similar to business objective. Conduct validation for the accuracy of the classification and also make sure to ensure the minimization of time and errors.

- Reviewing the Process:

  Once completion of the evaluation result for the classification validate the model with legal advisors or data scientists to ensure that it correctly classify or predict clauses according to legal standards.

- Determine Next Steps:

  If the model we are selected do not perform much well and does not meet expectations, then gather and collect more amount of data, perform good feature extraction process or try using different algorithm techniques in order to select the best algorithm for the implementation to meet the business objective. If everything is running well and meets the business objective, then it’s time for deployment of the model.

6.Deployment

❖ Deployment Planning:

Implement the model to classify legal documents in real-time application. Integrate the model with existing J.P. Morgan’s existing COIN system. Choose a deployment environment (e.g., AWS, Google, Azure.). Implement techniques for importing and processing legal documents which include integrating with document management systems. Develop a dashboard for stakeholders to interact with the classification system, acquiring results and managing the documents. Ensure that the legal documents are protected through encryption and access controls.

❖ Monitoring and Maintenance

Setting up a system that will continuously monitor the model's performance. By Monitoring we can make adjustment as needed.

- ### Reviewing the Project

Review the project to identify the areas for the improvement in order to meet the business objective and it will keep track for successful implementation of the model.

- ### Finalizing the Project

Ensure that the entire process is implemented or not from initializing the data to deployment is documented and find if any challenges are occurred and the successful implementation of the entire project and everything is organized in a correct manner.