**Document Classification and Information Extraction**

**1. Introduction**

The objective of this project was to develop a solution for information extraction and document classification using a combination of machine learning techniques. The task involved extracting the title and authors from PDF files and classifying each document into predefined categories such as Tables, Classification, Key Information Extraction, Optical Character Recognition, Datasets, Document Layout Understanding, and Others. We utilized two distinct models primarily for comparison to understand their strengths and weaknesses in handling this task.

**2. Methodology**

**Information Extraction**

For the information extraction task, we utilized a hybrid approach combining PDF parsing tools and Natural Language Processing (NLP) techniques:

* **PDF Parsing**: We employed PyMuPDF to extract text data directly from the first page of each PDF document. This tool allowed us to efficiently handle text extraction, maintaining the structural integrity of the document.
* **Named Entity Recognition (NER)**: To identify titles and authors, we used a pre-trained NLP model for NER. The NER model focused on recognizing entities classified as PERSON for author names and TITLE or similar labels for document titles. This step was crucial in accurately pinpointing relevant metadata.

**Document Classification**

We approached document classification using two different methodologies to enhance accuracy and gain insights into the document's content:

1. **BERT-based Approach with LDA**:
   * We used BERT (Bidirectional Encoder Representations from Transformers) from the Hugging Face Transformers library to generate document embeddings, capturing their contextual meaning.
   * To enhance topic detection within the text, we employed Latent Dirichlet Allocation (LDA), which helped identify the underlying topics and further refine the categorization.
   * The combined BERT and LDA approach enabled a robust classification system, leveraging the semantic power of BERT with the topic modeling capability of LDA to improve accuracy in document classification.
2. **Graph-Based Approach with BM25**:
   * We constructed a graph-based model using the Node2Vec algorithm, where words and documents were treated as nodes, and edges represented their relationships.
   * The importance of words in this model was calculated using the BM25 weighting scheme, a ranking function that assesses term importance based on its frequency and relevance within the documents.
   * Node2Vec generated embeddings that effectively captured the graph structure, and K-means clustering was used to group similar documents, followed by category matching based on keyword similarity.

**3. Challenges**

During the project, we encountered several challenges:

* **Text Extraction**: PDF formatting inconsistencies occasionally led to difficulties in accurately extracting titles and author names. We addressed this by refining the text parsing logic and using heuristic rules.
* **Model Comparison**: Comparing the BERT-LDA approach with the graph-based method required careful analysis to understand their respective strengths. While BERT-LDA excelled in capturing semantic nuances, the graph approach offered valuable insights into document connectivity.

**4. Results**

* **Information Extraction**: The hybrid approach successfully extracted titles and authors with a high level of accuracy. The use of NER models significantly reduced errors in identifying author names.
* **Document Classification**:
  + **BERT-LDA Approach**: This model demonstrated strong performance by effectively categorizing most documents into their respective categories, utilizing both semantic analysis and topic modeling.
  + **Graph-Based Approach with BM25**: This method excelled in capturing document relationships based on word importance and frequency, which led to accurate clustering and category mapping, complementing the BERT-LDA approach.

The final JSON output was structured according to the required schema, categorizing documents into predefined groups with the correct metadata.

**5. Conclusion**

This project demonstrated the benefits of using two distinct models for comparison—BERT-LDA for semantic understanding and topic detection, and a graph-based approach for structural analysis using BM25. Each model offered unique advantages that enriched the document classification process. Future improvements could include:

* **Enhanced PDF Parsing**: Integrating more sophisticated parsing techniques to handle various PDF formats and layouts.
* **Fine-tuning Classification Models**: Further fine-tuning the classification models with more labeled data to enhance performance.
* **Graph Embeddings**: Exploring other graph embedding techniques to improve the clustering and classification accuracy.

**6. References**

* **Libraries and Tools**: PyMuPDF, Hugging Face Transformers, Node2Vec, BM25, LDA, K-means, Matplotlib.
* Research on Transformer models, Graph-based algorithms, and topic modeling techniques for Natural Language Processing.