



# B.N.M. Institute of Technology

An Autonomous college under VTU

## Hybrid NLP-Based Document Classifier

Team Members:

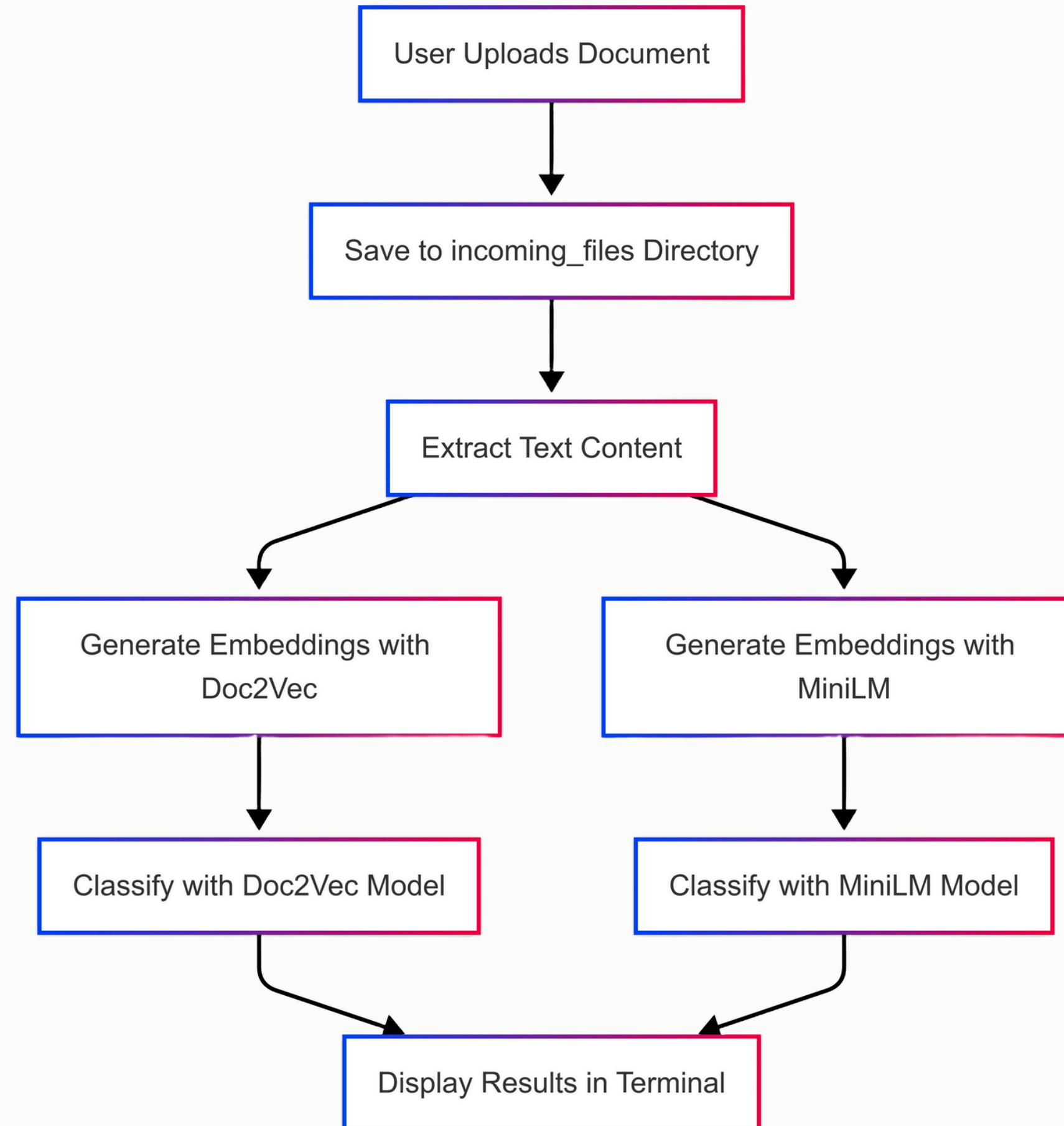
Abhay S	1BG23CS002
Abhiram H	1BG23CS004
Achintya K J	1BG23CS005

B.N.M. INSTITUTE OF TECHNOLOGY

# Objective

- To develop an intuitive and efficient Streamlit-based application that classifies uploaded PDF and DOCX documents into predefined categories.
- To compare the performance of two NLP models — Doc2Vec and all-MiniLM-L6-v2 — on real-world text classification tasks.
- To enable automated document sorting based on semantic content using word embeddings, improving organization and retrieval.
- To enable efficient document classification via a command-line tool, allowing users to upload files, trigger external model scripts, and evaluate model predictions and agreement..

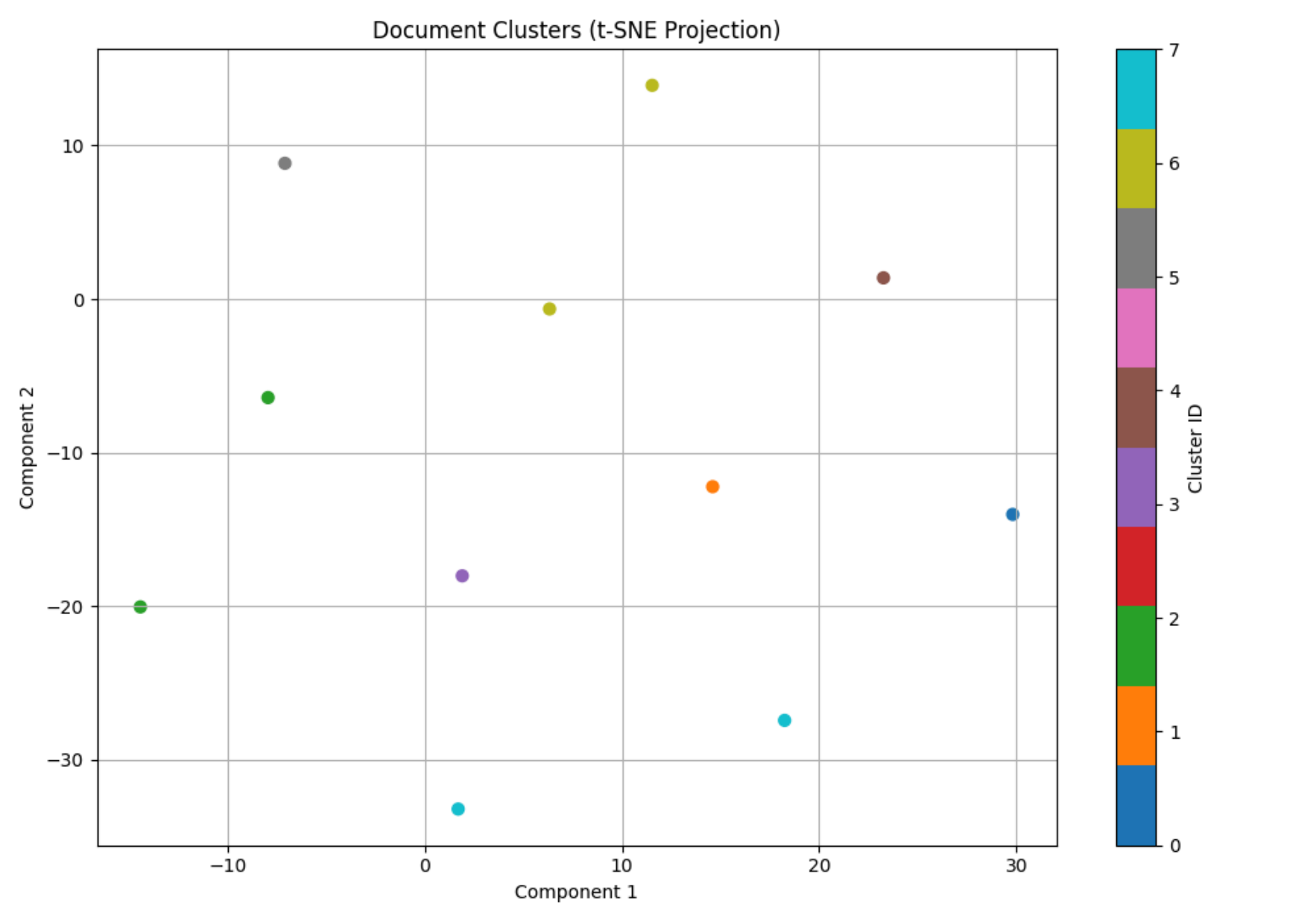
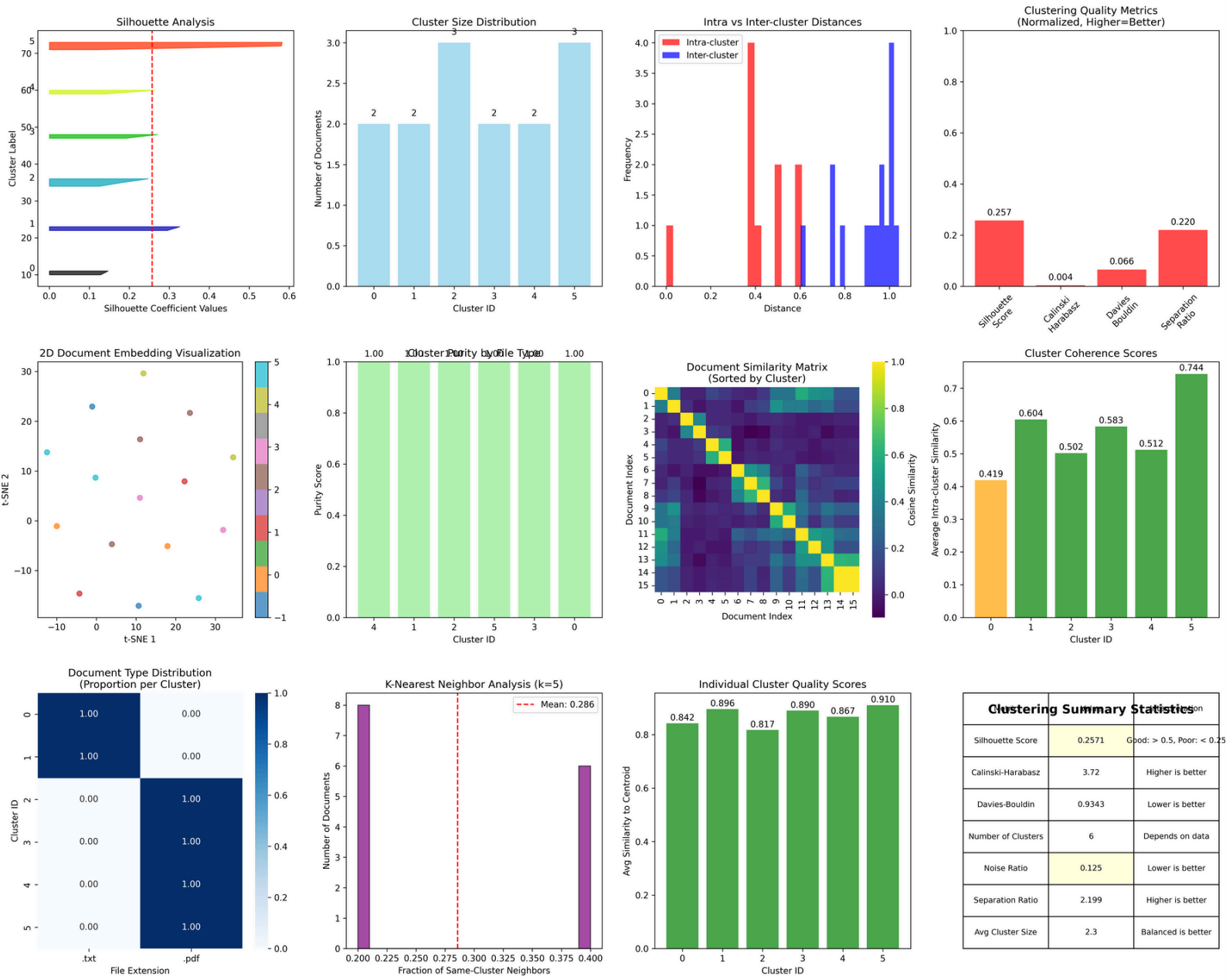
# Methodology and Workflow



# Key Assumptions

- Users will upload documents that are semantically rich and contain sufficient textual content for accurate classification.
- The pre-trained embeddings (Doc2Vec and MiniLM) are assumed to generalize well to the types of documents being uploaded.
- The uploaded files are free from significant noise, errors, or language inconsistencies that might degrade embedding quality.
- Both models (Doc2Vec and MiniLM) assume language is English and embeddings are aligned to English text semantics.

# Model Evaluation and Analysis





# Project Summary and Outcomes

## Project Summary

- Built a Streamlit-based document classifier for .pdf and .docx files using Doc2Vec and all-MiniLM-L6-v2 embeddings.
- Enabled document classification via CLI with support for external model execution and side-by-side prediction comparison from independent scripts.

## Project Outcomes

- Successfully deployed a CLI-based prototype for real-time document classification with necessary metrics.
- Validated the semantic effectiveness of embedding-based models (Doc2Vec and MiniLM) for categorizing varied document types.
- Established a foundation for future multi-model NLP systems, with extensibility for broader classification tasks.

# Future Improvements and Extensions

- Add support for multi-label classification for documents spanning multiple categories.
- Add a user-friendly frontend using streamlit for drag and drop upload.
- Integrate OCR to handle scanned or image-based PDFs.
- Incorporate Google's Universal Sentence Encoder or BERT variants for deeper contextual understanding.
- Enable model training from uploaded data to adapt to domain-specific documents.
- Develop a feedback loop where user corrections improve model accuracy over time.
- Add analytics dashboard for tracking document types, prediction trends, and model performance.

# Reflections and Learning Outcomes

- Gained hands-on experience in end-to-end NLP pipeline design using pre-trained embedding models and training a model.
- Understood trade-offs between different word embedding techniques like Doc2Vec and MiniLM.
- Identified challenges in text extraction, formatting inconsistencies, and model generalization.
- Recognized the importance of clear evaluation metrics and cross-model comparisons for trust.