Proposal for Automated Classification and Processing of Unstructured Financial Documents

1. Introduction

Appian Credit Union processes thousands of financial documents daily, including applications, identity documents, financial statements, and receipts. The current manual verification and organization are time-consuming and repetitive. Implementing an automated solution will enhance efficiency, reduce errors, and allow employees to focus on higher-value tasks.

2. Solution Overview

We propose an Intelligent Document Processing (IDP) system that leverages advanced AI models to automate the classification and processing of unstructured financial documents. The system will:

- **Ingest and preprocess** various document formats.
- **Extract and analyze** textual and visual information.
- **Classify documents** by type and associate them with the correct individuals.
- **Summarize key information** for quick review.
- **Store and manage** processed documents in a structured database.

3. System Architecture

The proposed architecture comprises the following components:

- **Data Ingestion Module:** Handles the upload and preprocessing of documents.
- **OCR and Text Extraction Module:** Utilizes Optical Character Recognition to digitize text from scanned documents.
- **Document Classification Module:** Employs AI models to categorize documents by type.
- **Entity Recognition and Association Module:** Identifies personal information to link documents to individuals.
- **Summarization Module:** Generates concise summaries of document content.
- **Storage and Retrieval Module:** Stores processed documents and metadata for easy access.

4. Technologies and Models

- Optical Character Recognition (OCR):
  - *Tesseract OCR:* An open-source OCR engine that supports multiple languages and is effective for digitizing text from scanned documents.
- Document Classification:
  - *LayoutLMv3:* A transformer-based model that integrates textual and layout information for document image understanding

- Named Entity Recognition (NER):
    - *SpaCy:* An open-source NLP library with pre-trained models for entity recognition, suitable for extracting personal information.
- Summarization:
    - *BERTSUM:* A variant of BERT fine-tuned for extractive summarization tasks, capable of generating concise summaries of documents.
- Storage:
    - *MongoDB:* A NoSQL database that allows for flexible storage of documents and their associated metadata.

5. Workflow

1. Data Ingestion:
    - Users upload documents through a web interface.
    - Documents are sent to the Data Ingestion Module for preprocessing.
2. OCR and Text Extraction:
    - The OCR module digitizes text from scanned documents.
    - Extracted text is passed to subsequent modules for analysis.
3. Document Classification:
    - The Classification Module uses LayoutLMv3 to categorize documents based on content and layout.
4. Entity Recognition and Association:
    - The NER module identifies personal information within the text.
    - Documents are associated with individuals using extracted identifiers.
5. Summarization:
    - The Summarization Module generates concise summaries highlighting key information.
6. Storage and Retrieval:
    - Processed documents and metadata are stored in MongoDB.
    - Users can search and retrieve documents through the web interface.

6. Implementation Plan

- **Phase 1: System Design and Planning**
    - Define requirements and success metrics.
    - Design system architecture and data flow.
- **Phase 2: Development**
    - Develop each module independently.
    - Integrate modules into a cohesive system.
- **Phase 3: Testing**
    - Conduct unit and integration testing.
    - Perform user acceptance testing with sample documents.
- **Phase 4: Deployment**
    - Deploy the system in a staging environment.
    - Monitor performance and address any issues.

  - Launch the system for production use.

7. Benefits

- **Increased Efficiency:** Automates repetitive tasks, reducing processing time.
- **Improved Accuracy:** Minimizes human errors in document classification and data entry.
- **Scalability:** Capable of handling large volumes of documents with consistent performance.
- **Enhanced Data Management:** Structured storage facilitates easy retrieval and analysis.

8. Innovation

The proposed solution introduces a novel integration of advanced AI models tailored specifically for the financial sector's document processing needs. Key innovative aspects include:

- **Multimodal Document Understanding:** By employing models like LayoutLMv3, which integrates textual and layout information, the system can comprehend complex document structures unique to financial documents.
  ArXiv
- **End-to-End Automation:** The solution automates the entire document processing workflow, from ingestion to storage, reducing manual intervention and increasing efficiency.
  Aks Desai
- **Scalability and Flexibility:** Utilizing AI-driven approaches allows the system to handle a wide variety of unstructured documents, adapting to different formats and complexities without the need for rigid templates.
  Indico Data

9. Feasibility

The implementation of this solution is feasible with current technologies:

- **AI Model Availability:** Models like Tesseract OCR, LayoutLMv3, SpaCy, and BERTSUM are readily available and have been successfully applied in similar contexts.
- **Computational Resources:** With the advent of cloud computing and specialized hardware, deploying these models at scale is practical.
- **Integration Capability:** The modular design ensures that each component can be developed and integrated incrementally, allowing for phased implementation and testing.
- **Industry Adoption:** Financial institutions are increasingly adopting AI-driven document processing solutions, demonstrating the practicality and effectiveness of such systems.
  Ailleron

10. Technical Accomplishment

The proposed system represents a significant technical achievement:

- **Advanced AI Integration:** Combining multiple AI models to handle various aspects of document processing showcases a sophisticated understanding of machine learning and natural language processing techniques.
- **Handling Unstructured Data:** Effectively processing unstructured financial documents requires overcoming challenges related to data variability and complexity, demonstrating technical prowess.
  LeewayHertz
- **Enhancing Operational Efficiency:** Implementing this system can lead to substantial improvements in processing speed and accuracy, highlighting the practical benefits of advanced AI applications in real-world scenarios.

11. Conclusion

Implementing this automated document processing system will streamline Appian Credit Union's operations, enhance productivity, and improve service quality. Leveraging advanced AI models ensures the solution is robust, scalable, and adaptable to future needs.

12. References

- Tesseract OCR
- LayoutLMv3
- SpaCy
- BERTSUM
- MongoDB