ClauseWise – Project Documentation

1. Project Overview

ClauseWise is an AI-powered legal document analyzer designed to simplify, decode, and classify complex legal documents. It empowers lawyers, businesses, and individuals by automating the review and understanding of legal language using advanced natural language processing (NLP) and large language models (LLMs).

2. Objectives

The primary goals of ClauseWise are to:

- Automatically simplify complex legal clauses into plain, understandable language.

- Identify and extract key legal entities such as parties, dates, monetary values, and obligations.

- Break lengthy documents into individual, analyzable clauses.

- Classify documents into categories such as NDAs, lease agreements, employment contracts, etc.

- Support various input formats like PDF, DOCX, and TXT.

- Provide a simple, interactive, and user-friendly interface for legal analysis.

3. Technologies Used

Layer	Tools & Technologies
Frontend	Streamlit (UI), Gradio (optional)
NLP & AI	Python, HuggingFace Transformers, IBM Watson NLU, SpaCy
Document Parsing	PyMuPDF (PDF), python-docx (DOCX)
Classification	Scikit-learn (TF-IDF + ML), BERT (HuggingFace)

Infrastructure | Local machine or cloud-hosted (optional)

4. System Architecture & Workflow

1. Document Upload: Users upload PDF, DOCX, or TXT documents via Streamlit UI.

2. Text Extraction: Text is extracted using PyMuPDF (for PDFs), python-docx (for DOCX), or standard I/O (for TXT).

3. Clause Segmentation: Extracted text is passed through SpaCy to break it down into individual clauses.

4. Clause Simplification: Each clause is passed through a HuggingFace T5 model (or similar) to generate simplified output.

5. Named Entity Recognition: Entities are extracted using SpaCy and/or IBM Watson NLU.

6. Document Classification: The full document is classified using either a traditional ML model or a BERT-based classifier.

7. Output Display: The simplified clauses, extracted entities, and classification results are displayed in an intuitive format via Streamlit.

5. Modules Implemented

- File Handler: Handles uploading and text extraction for PDF, DOCX, and TXT files.

- Clause Segmenter: Uses SpaCy to segment documents into logical legal clauses.

- Simplification Engine: Utilizes pretrained transformer models to simplify complex legal text.

- NER Processor: Extracts structured information such as entities and obligations.

- Classifier: Predicts the category of the document (e.g., NDA, lease).

- Frontend Interface: Built in Streamlit, allowing users to upload files, view output, and interact with results.

6. Challenges Faced

- Clause Boundary Detection: Legal text formatting varies widely, making clause segmentation non-trivial.

- Simplification Quality: Ensuring that simplified clauses retain legal accuracy and context.

- Entity Extraction Accuracy: Differentiating between legal parties, dates, and unrelated named entities.

- Multi-format File Parsing: Handling formatting inconsistencies across different document types.

7. Future Enhancements

- Incorporate feedback loops for user corrections to improve clause quality.

- Fine-tune classification and simplification models on legal-specific datasets.

- Add support for multilingual legal documents.

- Provide downloadable or shareable outputs in formats like PDF or CSV.

- Integrate with document management systems and legal CRMs.

8. Business & Social Impact

- Time Efficiency: Reduces legal review time from hours to minutes.

- Cost Saving: Cuts down on external legal consultation for basic document understanding.

- Accessibility: Makes legal content understandable to non-lawyers, freelancers, and small businesses.

- Transparency & Compliance: Improves contract visibility and reduces risk of missed obligations.

9. Conclusion

ClauseWise presents a practical and intelligent solution to a common challenge in legal operations: understanding and analyzing complex legal documents. By combining natural language understanding, document parsing, and AI-powered simplification in a single tool, ClauseWise enhances legal efficiency, accessibility, and transparency.