Journal of Science, Computing and Engineering Research (JSCER)
Volume-9, Issue- 3, Mar-Apr 2026.
DOI: *https://doi.org/10.46379/ jscer.2026.0709*

Legal Document Simplifier Using GPT

[1]J.R.Arun Kumar,[2]Mukul Choudhary,[3]Harsh Saini,[4]Naveen Kanwat,[5]Naveen Gurjar

[1]Professor, Department of CSE, Modern Institute of Technology and Research Centre, Rajasthan, India.

[2,3,4,5]UG Student, Department of CSE, Modern Institute of Technology and Research Centre, Rajasthan, India

Article Information

Received : Mar 25 2026
Revised : Mar 25 2026
Accepted : Mar 26 2026
Published : Mar 28 2026

Corresponding Author:

Mukul Choudhary

***Abstract*—** Legal documents are often written in highly technical and complex language filled with jargon, lengthy clauses, and intricate sentence structures. While these documents are critical for everyday matters—such as contracts, agreements, property papers, and government policies—most common people without legal expertise struggle to understand them. This lack of comprehension can lead to misinterpretation, dependency on costly legal consultations, and in some cases, unintentional violation of terms. There is a need for an accessible system that can automatically simplify complex legal language into plain, everyday language without losing the original meaning or context. By using advanced Natural Language Processing (NLP) techniques and powerful models , it is possible to bridge the gap between legal professionals and the general public,

Keywords: Natural Language Processing, Legal Document Simplification, Keyword Parsing, Text Simplification ,Automated Summarization.

Citation: : J.R.Arun Kumar, Mukul Chodhary, Harsh Saini, Naveen Kanwat, Naveen Gurjar, "LEGAL DOCUMENT SIMPLIFIER", Journal of Science, Computing and Engineering Research, 9(3), MAR 2026.

I. PROBLEM STATEMENT

Legal documents such as agreements, court judgments, legislative acts, regulatory policies, and compliance documents form the backbone of governance, business operations, and public administration.

However, these documents are typically written in dense, highly technical, and formalized legal language intended for professionals with specialized expertise. For ordinary citizens, students, or business users without legal training, understanding such documents becomes a major challenge. However, this transition has created a "transparency gap" where candidates are unaware of how their professional history is being interpreted by a machine

.

The system functions by processing legal texts into small chunks used for text processing to analyze clauses, terminology, and their contextual meanings. It identifies difficult terms and breaks down legal jargon into user-friendly explanations. The platform provides concise summaries of important sections and highlights key clauses to improve comprehension.

A. The Core Mechanism

The core mechanism of the Legal Document Simplifier is based on a multi-stage Natural Language Processing (NLP) pipeline that converts complex legal text into simple and understandable language. The system first takes a legal document as input and extracts the text, which is then cleaned and preprocessed to remove noise and preserve structure.:

- **Text Extraction:** Extracting raw text from various document schemas.
- **Summarization using BART/LLM:** Standardizing text to remove noise (headers, footers, and non-standard characters).
- **Response Generation:** Top-ranked results are aggregated and the LLM synthesizes concise, accurate, simplified answers.

II. PROBLEM STATEMENT

In an age dominated by digital documentation, organizations and professionals across industries grapple with managing extensive data stored in legal document formats.

A. Complexity of Legal Language

Legal documents are written using highly technical terminology, complex sentence structures, and domain-specific jargon that are difficult for non-experts to understand**.**

- Most legal texts include long sentences, multiple clauses, and formal vocabulary designed for legal professionals.
- **Section Header Accuracy:** This makes it challenging for common users to interpret the meaning correctly, often leading to confusion or misinterpretation.

B. Lack of Accessibility for Common Users

Legal information is not easily accessible or understandable for the general public without professional assistance.

III. PROPOSED MODEL

The proposed system follows a multi-stage architecture designed to simplify and summarize complex legal documents using advanced NLP and transformer-based models.

A. Document Input and Preprocessing Module

This module handles the initial stage of the system where legal documents are uploaded and prepared for further processing.

B. Key Sub-processes in Preprocessing:

- **Accepts input in PDF, DOCX, or text format:** The system accepts multiple document formats to ensure flexibility. Users can upload contracts, agreements, or legal notices in commonly used ormats.
- **Removes unwanted symbols, headers, and formatting noise:** Unnecessary elements such as special characters, extra spaces, and formatting symbols are removed to improve processing accuracy.
- Splits large documents into smaller chunks:
- Elimination of "Digital Artifacts" such as page numbers, repeated headers/footers, and watermarks.
- Normalization of whitespace (collapsing tabs, newlines, and multiple spaces into a single space).

C. Abstractive Summarization Module (BART Model):

- **Transformer Encoder-Decoder :**BART uses an encoder to understand the input text and a decoder to generate summaries, enabling high-quality text generation.
- **Context Understanding:** The model captures relationships between different parts of the text, which is important for legal documents with interdependent clauses.
- Preservation of Key Information: Important legal points such as obligations, conditions, andclauses are retained to avoid loss of meaning.

D. Output Generation and User Interface Module

After processing, the system organizes and displays the results in a user-friendly format for easy access and interaction.

- **Summary Integration:** All processed chunks are combined into a single coherent summary to maintain logical flow**.**

A. Component Diagram:

The Component Diagram illustrates the structural relationship among the software components.

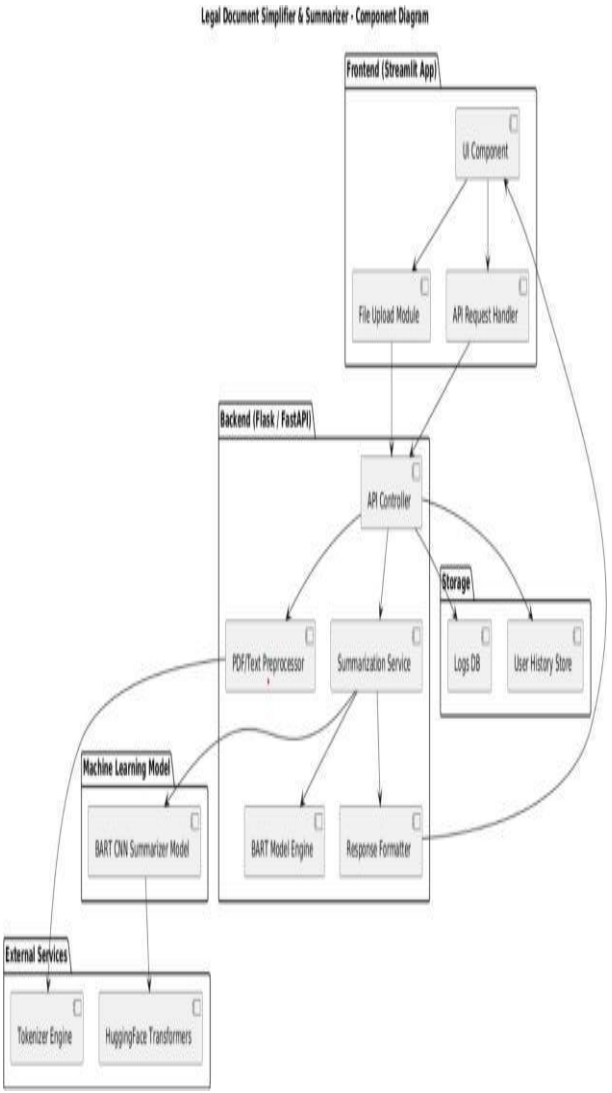

Fig. 1. Component Diagram

IV. TECH STACK

The proposed Legal Document Simplifier system is built using a combination of frontend technologies, backend frameworks, NLP models, and supporting tools to ensure efficient processing, scalability, and user-friendly interaction.

A. Frontend Technologies (Streamlit)

The frontend of the system is developed using Streamlit, which provides a simple and efficient way to build interactive web applications using Python.

It allows users to upload legal documents, view simplified summaries, and interact with the system in real time. Streamlit eliminates the need for complex frontend development and ensures a clean, user-friendly interface that is accessible even to non-technical users.

B. Backend Technologies

The backend is built using Python and the Flask framework, which handles the core logic of the system. Flask manages communication between the frontend and the processing modules through APIs.

It is an core of its mian function. It processes user requests, handles document uploads, and coordinates tasks such as preprocessing, summarization, and simplification.

C. NLP and Machine Learning Models

The system uses advanced NLP models such as BART-Large CNN and GPT to perform summarization and simplification.

BART is responsible for generating concise summaries of legal documents, while GPT rewrites the content into simple and understandable language.

D. Text Processing Libraries

Libraries such as PyPDF2 and python-docx are used to extract text from PDF and Word documents.

These tools convert document content into a format suitable for processing.

E. Ollama

Ollama is an open-source tool used to run large language models (LLMs)—such as Llama 3, Mistral, and LLaVA—locally on your own computer, rather than in the cloud.

It simplifies setting up AI models for private, offline use, making it popular for developing

.

V. RESULT SCREENSHOTS

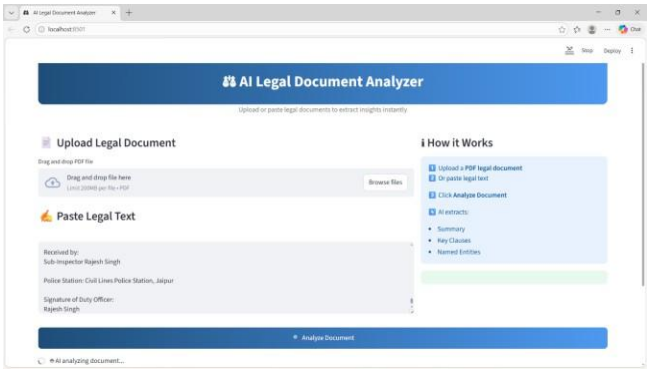

Fig. 2. Home Page.

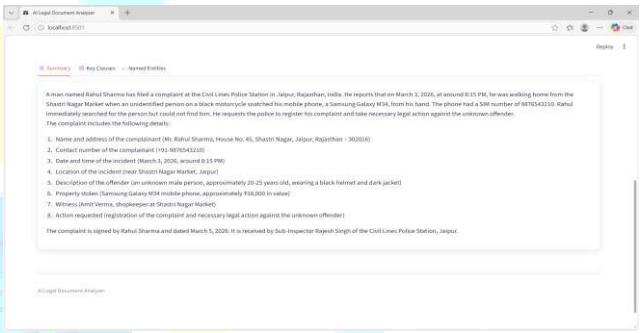

Fig. 3. Output Page

VI. CONCLUSION

A. Project Synthesis

The "**Simplify Law**" project successfully integrates advanced Natural Language Processing (NLP) techniques and transformer-based models to address the challenge of understanding complex legal documents. The system combines multiple stages, including document extraction, preprocessing, chunking, abstractive summarization using BART, and simplification using GPT-based models.

This pipeline ensures that lengthy and jargon-heavy legal texts are transformed into concise and user-friendly summaries while preserving overall architecture demonstrates an effective synergy between data processing, machine learning, and user interface design, resulting in a scalable and efficient solution.

- **Integration of Multiple Technologies:**

The project successfully combines Natural Language Processing, transformer-based models, and web technologies into a unified system.

B. Core Achievements

The project has delivered on all its foundational promises:

- **User Interface Layer:** This layer provides the inter connection point between the user and he system it is built using streamlit and allows user to upload legal document and view simplified output.
- **Input Acquation Model:** The Module handles the invoke of document in multiple format such as PDF, DOCX or plain text
- **Simplication module :** The module converts legal

  Content into simple and easy language it uses gpt

  Based module to write complex sentences.
- **Storage and Data Management Layer:** This layer manages storage of uploaded documents, processed summaries, and related metadata using databases such as PostgreSQL or MongoDB.

C. Concluding Remarks

In conclusion, the Legal Document Simplifier represents a meaningful advancement toward making legal information more accessible and understandable for the general public. By reducing dependency on legal professionals for basic interpretation, the system promotes transparency, efficiency, and informed decision-making. The project lays a strong foundation for future enhancements such as multilingual support, domain-specific model training, and integration with legal databases. Overall, it demonstrates how AI can bridge the gap between complex legal systems and everyday users, contributing to a more informed and empowered society.

REFERENCES

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017)..

[2] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners

[3] Ollama (2026). *Ollama Documentation (Version 3.2.x)*. Retrieved from https://docs.ollama.com/api/usage/3.0.x/

[4] Streamlit (2026). *Streamlit Documentation (Fitz)*. Retrieved from https://docs.streamlit.io/develop/api-reference

[5] Python Software Foundation. (2026). *python-docx: Create and Update Microsoft Word .docx files*. Retrieved from https://python-docx.readthedocs.io/en/latest/ legalese. (2025).

[8] legalese. (2025). Even lawyers do not like legalese : A Guide for Employers and Candidates. national library of medicine.