MPBA G519 – NLP FOR BUSINESS
PROJECT REPORT

MBA BUSINESS ANALYTICS
SECOND SEMESTER 2024-25

TOPIC - CONTRACT REVIEW AND LEGAL DOCUMENT ANALYSIS USING NLP

SUPERVISOR: PROF. SATANIK MITRA

DATE: 21st APRIL 2025

SUBMITTED BY-

1. KARTIK VIJAY BADKAS - 2024H1540809P
2. SATHEESH M K - 2024H1540810P
3. SARVESH KULKARNI - 2024H1540820P
4. GUZZU ADITYA - 2024H1540827P
5. SAKET PITALE - 2024H1540836P

Table of Content:

Abstract:

Legal contract review is a time-consuming and error-prone process that demands deep legal knowledge and meticulous attention to detail. LegalDocNLP is a Natural Language Processing (NLP) based system developed to automate and simplify the review process of legal contracts. The application extracts standard contractual clauses, classifies them with Legal-BERT, flags risky clauses, and generates concise summaries using T5. Built using Python and Flask, the system also supports keyword context analysis, enabling faster, more accurate, and scalable contract analysis.

1. Introduction:

Contracts form the foundation of most business agreements. Manual review of legal documents is labor-intensive, subjective, and prone to oversight. The advent of NLP provides an opportunity to leverage machine learning models trained on legal text to automate much of this process. This project implements a contract analyzer that reads a PDF file, extracts key clauses, classifies them using Legal-BERT, summarizes them using LegalT5, and displays the results in a structured, user-friendly web interface.

2. Problem Statement:

Manual contract review involves:

- Identifying risky clauses
- Understanding obligations
- Locating key parties and dates
- Analyzing terminations, indemnities, liabilities, etc.

All of the above are resource-heavy and inconsistent. The challenge is to build a system that:

- Automates clause classification
- Flags risky content
- Summarizes clauses
- Allows legal professionals to work faster and with fewer errors

3. Objective:

- Automate legal clause identification using Legal-BERT
- Summarize clauses using LegalT5
- Classify clauses by type (e.g., Termination, Indemnity, Confidentiality)
- Display extracted content via a user-friendly Flask interface
- Allow keyword-based clause context extraction

4. Literature Review / Base Paper Summary:

The project is built upon foundations laid in the CUAD (Contract Understanding Atticus Dataset) and research papers involving Legal-BERT and LegalT5. CUAD is a curated dataset of legal contracts annotated for clause types. Legal-BERT, built upon the BERT architecture, is pre-trained on legal documents and is ideal for classification tasks in the legal domain. LegalT5 is an adaptation of the T5 transformer model, trained to generate clause summaries.

5. Methodology:

The methodology includes the following components:

- Step 1: PDF Extraction (using PyMuPDF/fitz)
- Step 2: Clause Segmentation (sentence tokenization)
- Step 3: Clause Classification using Legal-BERT
- Step 4: Labelling using keyword mapping
- Step 5: Summarization using LegalT5
- Step 6: Keyword Context Analysis
- Step 7: Render through Flask web app

6. Technologies Used:

- Programming Language: Python 3.10+
- Libraries:
  - transformers (HuggingFace)
  - nltk
  - Flask
  - PyMuPDF (fitz)
  - json / os / re / torch

7. Implementation:

This section outlines the implementation.

a) Clause Classification-
   Using Legal-BERT, each sentence is passed through a pre-trained model:

```
# Load models and tokenizers
summarizer_tokenizer = T5Tokenizer.from_pretrained( pretrained_model_name_or_path: "SEBIS/legal_t5_small_summ_en", use_fast=False)
summarizer_model = T5ForConditionalGeneration.from_pretrained("SEBIS/legal_t5_small_summ_en")

bert_tokenizer = BertTokenizer.from_pretrained("nlpaueb/legal-bert-base-uncased")
bert_model = BertForSequenceClassification.from_pretrained( pretrained_model_name_or_path: "nlpaueb/legal-bert-base-uncased", num_labels=2)
```

b) Clause Labeling-

Keyword-based tagging is applied for labels like “Indemnity”, “Confidentiality”, etc.

```
# Define your clause labels
LABELS = [
    "Confidentiality",
    "Indemnity",
    "Termination",
    "Payment Terms",
    "Governing Law",
    "Intellectual Property",
    "Arbitration",
    "Exclusivity",
    "Non-compete",
    "Other"
]

1 usage
def classify_clause(clause_text):
    inputs = tokenizer(clause_text, return_tensors="pt", truncation=True, padding=True)
    with torch.no_grad():
        outputs = model(**inputs)
    logits = outputs.logits
    predicted_class_id = logits.argmax().item()
    return LABELS[predicted_class_id]
```

c) Clause Summarization-

Using T5 model trained on legal text:

```
from transformers import T5Tokenizer, T5ForConditionalGeneration
from transformers import BertTokenizer, BertForSequenceClassification
import torch
import nltk
import fitz  # PyMuPDF

nltk.download('punkt')

# Load models and tokenizers
summarizer_tokenizer = T5Tokenizer.from_pretrained( pretrained_model_name_or_path: "SEBIS/legal_t5_small_summ_en", use_fast=False)
summarizer_model = T5ForConditionalGeneration.from_pretrained("SEBIS/legal_t5_small_summ_en")
```

d) Flask Integration-

A front-end is provided to:

- Upload contracts
- View clause-wise labels and summaries

```
<!doctype html>
<html>
<head>
    <title>Upload Contract</title>
</head>
<body>
    <h2>Upload Your Contract PDF</h2>
    <form method="POST" enctype="multipart/form-data">
        <input type="file" name="pdf_file" required>
        <input type="submit" value="Analyze">
    </form>
</body>
</html>
```

e) Keyword Context Extraction-
Important keywords are highlighted with context from surrounding clauses.

```
2 usages
def extract_keywords_contexts(pdf_path, top_n=5, window=40):
    doc = fitz.open(pdf_path)
    keyword_counts = Counter()
    keyword_contexts = defaultdict(list)

    for page_num, page in enumerate(doc, start=1):
        text = page.get_text()
        words = re.findall( pattern: r'\b\w+\b', text.lower())
        keyword_counts.update(words)

    stopwords = set([
        'the', 'and', 'to', 'of', 'a', 'in', 'that', 'is', 'for', 'with', 'as',
        'on', 'at', 'by', 'an', 'be', 'this', 'are', 'from', 'or', 'it', 'was',
        'which', 'we', 'not', 'can', 'has', 'have', 'will', 'may', 'shall'
    ])

    for word in list(keyword_counts):
        if word in stopwords or len(word) < 3:
            del keyword_counts[word]

    top_keywords = keyword_counts.most_common(top_n)

    for keyword, _ in top_keywords:
        for page_num, page in enumerate(doc, start=1):
            text = page.get_text().lower()
            matches = [m.start() for m in re.finditer(r'\b{}\b'.format(re.escape(keyword)), text)]
            for match in matches:
                start = max(0, match - window)
                end = min(len(text), match + window)
                context = text[start:end]
```

8. Results:

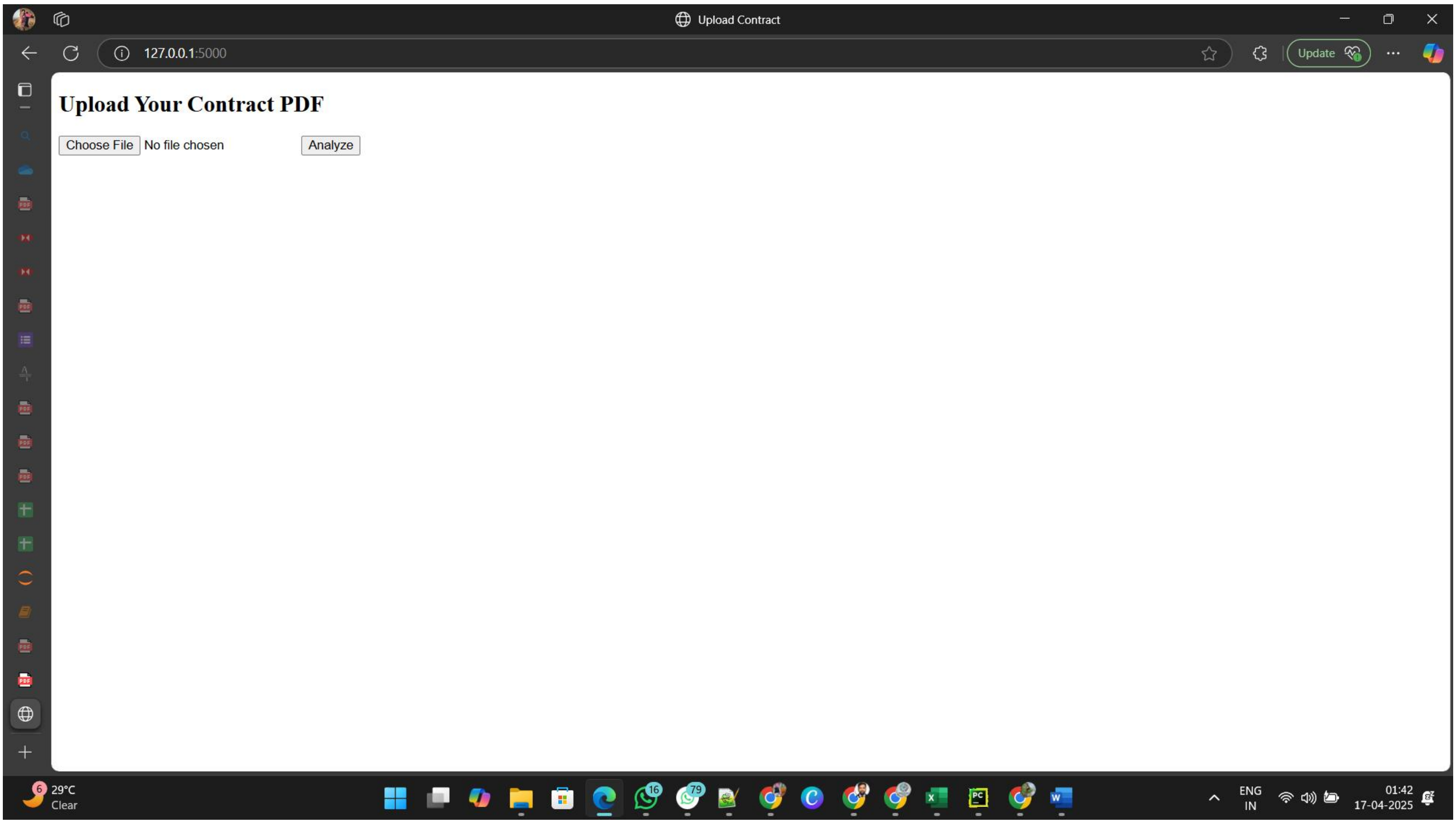
Results are presented in a structured HTML interface. Each clause is displayed with:

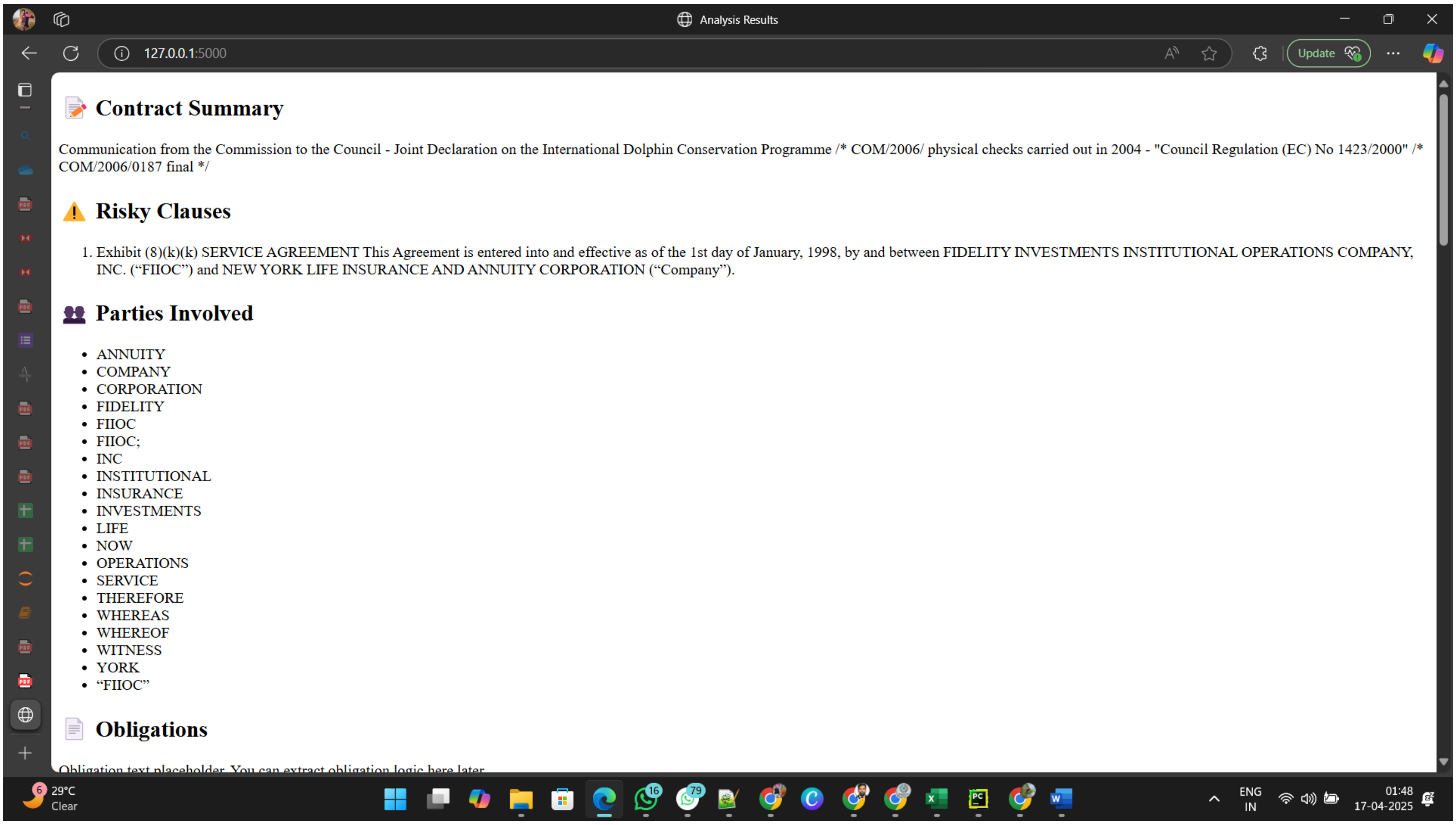
- Clause Label (e.g., Termination)
- Full Text
- Abstractive Summary

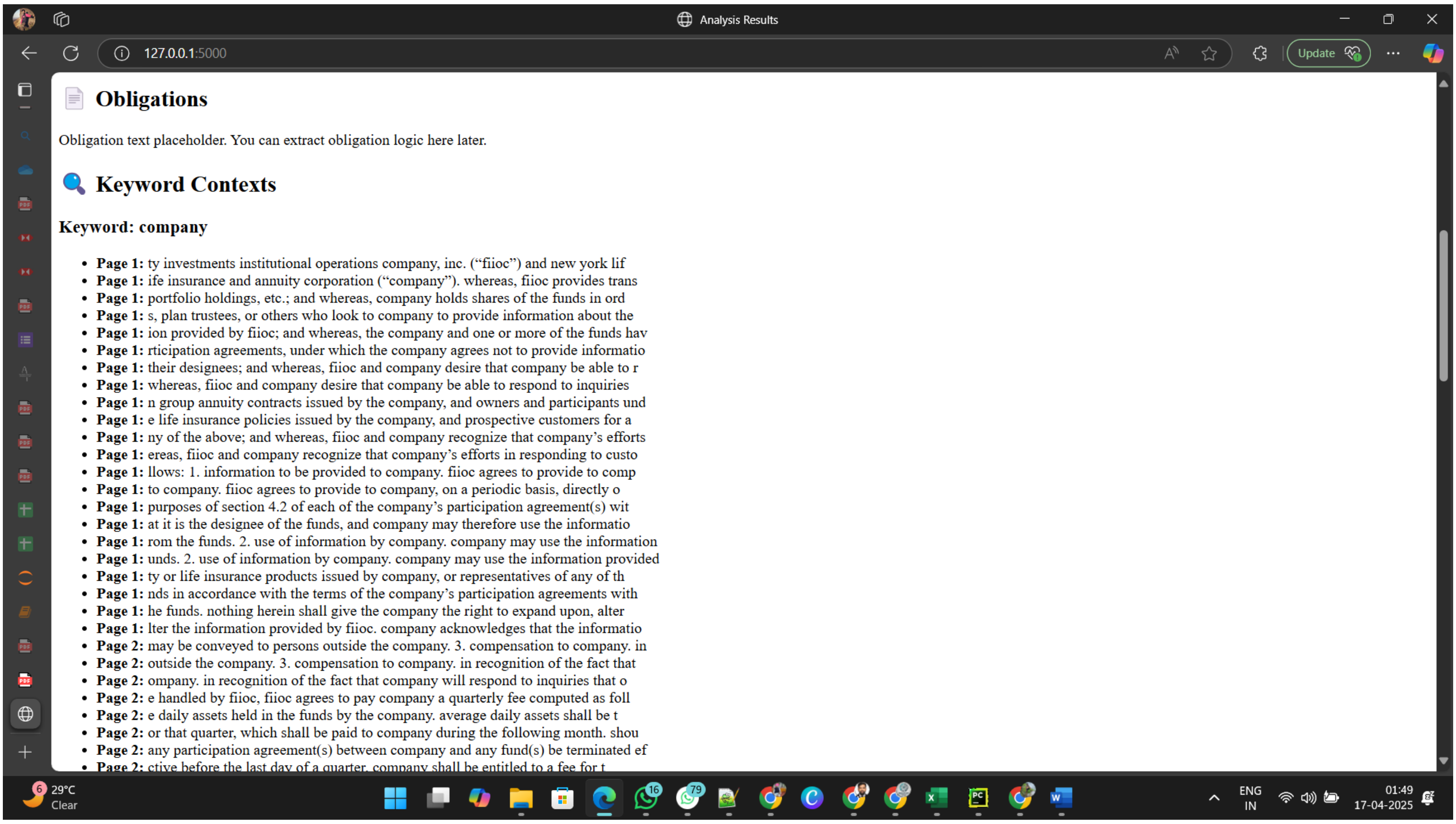
Additionally:

- All parties are extracted
- Obligations are listed
- Top keywords are shown with contextual usage

9. Sample Output Screenshots







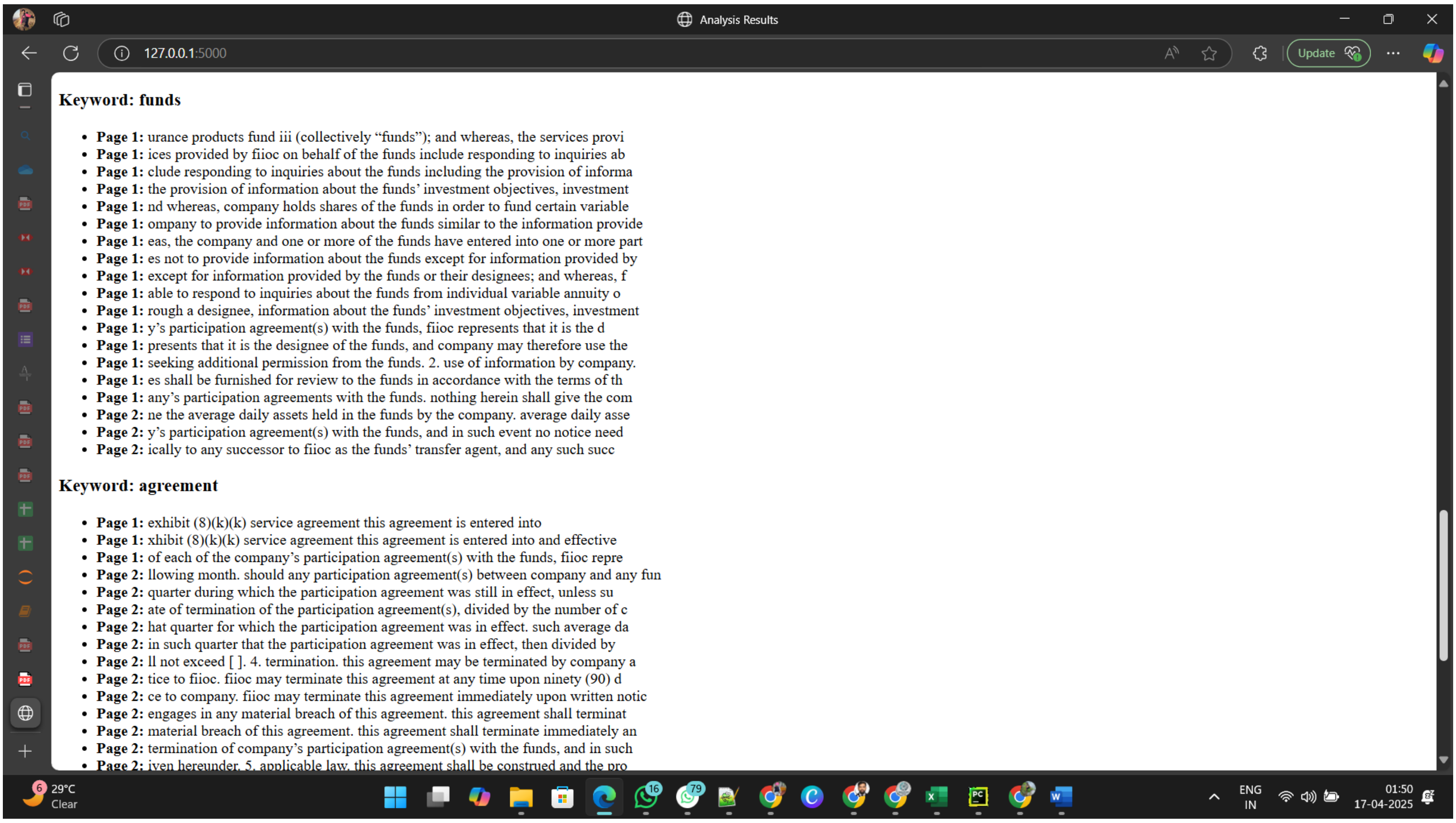
Obligations

Obligation text placeholder. You can extract obligation logic here later.

Keyword Contexts

Keyword: company

- **Page 1:** ty investments institutional operations company, inc. ("fiioc") and new york lif
- **Page 1:** ife insurance and annuity corporation ("company"). whereas, fiioc provides trans
- **Page 1:** portfolio holdings, etc.; and whereas, company holds shares of the funds in ord
- **Page 1:** s, plan trustees, or others who look to company to provide information about the
- **Page 1:** ion provided by fiioc; and whereas, the company and one or more of the funds hav
- **Page 1:** rticipation agreements, under which the company agrees not to provide informatio
- **Page 1:** their designees; and whereas, fiioc and company desire that company be able to r
- **Page 1:** whereas, fiioc and company desire that company be able to respond to inquiries
- **Page 1:** n group annuity contracts issued by the company, and owners and participants und
- **Page 1:** e life insurance policies issued by the company, and prospective customers for a
- **Page 1:** ny of the above; and whereas, fiioc and company recognize that company's efforts
- **Page 1:** ereas, fiioc and company recognize that company's efforts in responding to custo
- **Page 1:** llows: 1. information to be provided to company. fiioc agrees to provide to comp
- **Page 1:** to company. fiioc agrees to provide to company, on a periodic basis, directly o
- **Page 1:** purposes of section 4.2 of each of the company's participation agreement(s) wit
- **Page 1:** at it is the designee of the funds, and company may therefore use the informatio
- **Page 1:** rom the funds. 2. use of information by company. company may use the information
- **Page 1:** unds. 2. use of information by company. company may use the information provided
- **Page 1:** ty or life insurance products issued by company, or representatives of any of th
- **Page 1:** nds in accordance with the terms of the company's participation agreements with
- **Page 1:** he funds. nothing herein shall give the company the right to expand upon, alter
- **Page 1:** lter the information provided by fiioc. company acknowledges that the informatio
- **Page 2:** may be conveyed to persons outside the company. 3. compensation to company. in
- **Page 2:** outside the company. 3. compensation to company. in recognition of the fact that
- **Page 2:** ompany. in recognition of the fact that company will respond to inquiries that o
- **Page 2:** e handled by fiioc, fiioc agrees to pay company a quarterly fee computed as foll
- **Page 2:** e daily assets held in the funds by the company. average daily assets shall be t
- **Page 2:** or that quarter, which shall be paid to company during the following month. shou
- **Page 2:** any participation agreement(s) between company and any fund(s) be terminated ef
- **Page 2:** ctive before the last day of a quarter, company shall be entitled to a fee for t



Keyword: funds

- **Page 1:** urance products fund iii (collectively "funds"); and whereas, the services provi
- **Page 1:** ices provided by fiioc on behalf of the funds include responding to inquiries ab
- **Page 1:** clude responding to inquiries about the funds including the provision of informa
- **Page 1:** the provision of information about the funds' investment objectives, investment
- **Page 1:** nd whereas, company holds shares of the funds in order to fund certain variable
- **Page 1:** ompany to provide information about the funds similar to the information provide
- **Page 1:** eas, the company and one or more of the funds have entered into one or more part
- **Page 1:** es not to provide information about the funds except for information provided by
- **Page 1:** except for information provided by the funds or their designees; and whereas, f
- **Page 1:** able to respond to inquiries about the funds from individual variable annuity o
- **Page 1:** rough a designee, information about the funds' investment objectives, investment
- **Page 1:** y's participation agreement(s) with the funds, fiioc represents that it is the d
- **Page 1:** presents that it is the designee of the funds, and company may therefore use the
- **Page 1:** seeking additional permission from the funds. 2. use of information by company.
- **Page 1:** es shall be furnished for review to the funds in accordance with the terms of th
- **Page 1:** any's participation agreements with the funds. nothing herein shall give the com
- **Page 2:** ne the average daily assets held in the funds by the company. average daily asse
- **Page 2:** y's participation agreement(s) with the funds, and in such event no notice need
- **Page 2:** ically to any successor to fiioc as the funds' transfer agent, and any such succ

Keyword: agreement

- **Page 1:** exhibit (8)(k)(k) service agreement this agreement is entered into
- **Page 1:** xhibit (8)(k)(k) service agreement this agreement is entered into and effective
- **Page 1:** of each of the company's participation agreement(s) with the funds, fiioc repre
- **Page 2:** llowing month. should any participation agreement(s) between company and any fun
- **Page 2:** quarter during which the participation agreement was still in effect, unless su
- **Page 2:** ate of termination of the participation agreement(s), divided by the number of c
- **Page 2:** hat quarter for which the participation agreement was in effect. such average da
- **Page 2:** in such quarter that the participation agreement was in effect, then divided by
- **Page 2:** ll not exceed [ ]. 4. termination. this agreement may be terminated by company a
- **Page 2:** tice to fiioc. fiioc may terminate this agreement at any time upon ninety (90) d
- **Page 2:** ce to company. fiioc may terminate this agreement immediately upon written notic
- **Page 2:** engages in any material breach of this agreement. this agreement shall terminat
- **Page 2:** material breach of this agreement. this agreement shall terminate immediately an
- **Page 2:** termination of company's participation agreement(s) with the funds, and in such
- **Page 2:** iven hereunder. 5. applicable law. this agreement shall be construed and the pro

10. Conclusion:

This project demonstrates how domain-specific NLP models can enhance the speed and accuracy of legal contract analysis. Using Legal-BERT and LegalT5, we successfully:

- Identified and labeled clauses
- Summarized legal language
- Highlighted risky content
- Delivered outputs in a clean, structured format
- The tool provides meaningful assistance to legal teams, saving time and effort.

11. Future Scope:

- Add OCR for scanned contracts
- Support multi-language contracts
- Train on user-provided templates
- Integrate with enterprise CMS (e.g., SharePoint, Salesforce)
- Add clause editing and clause comparison features

12. References

- https://huggingface.co/nlpaueb/legal-bert-base-uncased
- https://huggingface.co/SEBIS/legal_t5_small_summ_en
- CUAD Dataset: https://huggingface.co/datasets/cuad
- PyMuPDF: https://pymupdf.readthedocs.io/
- T5 paper: "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer"