Business Contract Validation Project Documentation

Team Members

1. **Luvkush Sharma**
2. **Dheeraj Sharma**
3. **Anik Roy**
4. **Muskan Garg**
5. **Uday Pratap Singh**

Table of Contents

Introduction

1.1 Overview

The Business Contract Validation project is an innovative solution to automate the process of analysing and validating legal business contracts. It leverages machine learning, generative AI, and advanced front-end development techniques to classify contract contents, identify deviations from standard templates, and highlight these deviations for user review.

1.2 Objectives

- **Automated Parsing**: Extract structured data from unstructured contract documents.
- **Content Classification**: Classify the extracted text into predefined clauses and sub-clauses.
- **Deviation Detection**: Identify and highlight deviations from standard contract templates.
- **Enhanced Accuracy**: Utilize advanced NLP and machine learning models.
- **User-Friendly Interface**: Develop an intuitive front-end for easy interaction and review.
- **Scalability and Performance**: Ensure the solution is scalable and performs efficiently.

1.3 Importance

Business contracts are critical legal documents that outline the terms and conditions of business relationships. Ensuring the accuracy and compliance of these contracts is essential to minimize legal risks and maintain business integrity. This project addresses the challenges of manual contract review by automating the validation process, thus enhancing efficiency and accuracy.

1.4 Benefits

- **Efficiency**: Reduces the time required to review and validate contracts.
- **Accuracy**: Minimizes human errors and ensures consistent validation criteria.
- **Compliance**: Helps maintain adherence to legal standards by automatically identifying deviations.
- **Cost Savings**: Lowers operational costs associated with manual contract review.
- **Scalability**: Capable of handling a high volume of contracts, making it suitable for organizations of all sizes.

1.5 Target Audience

The project is designed for businesses, legal firms, and other organizations that regularly manage large volumes of contracts. It is also a valuable learning opportunity for students and professionals interested in machine learning, generative AI, and full-stack development.

Project Description

2.1 Problem Statement

The project focuses on classifying content within business contract clauses, determining deviations from templates, and highlighting these deviations. Business contracts are legal documents that need to be parsed and structured, with key details and clauses identified. The goal is to classify the contents into these clauses and sub-clauses, and compare them against associated templates to highlight any deviations.

2.2 Project Goals

- **Parse contract documents**: Extract text and structure from PDF contracts.
- **Classify content**: Categorize text into predefined clauses and sub-clauses.
- **Identify deviations**: Compare classified content against standard templates to detect deviations.
- **Highlight deviations**: Highlight differences for easy review.
- **Integrate additional features**: Include functionalities like personal information masking, profanity filtering, and sentiment classification.

2.3 Key Components

- **PDF Parsing**: Extracts text from contract documents.
- **Named Entity Recognition (NER)**: Identifies key entities within the text.
- **Text Classification**: Classifies text into clauses and sub-clauses.
- **Text Comparison**: Compares text against templates to identify deviations.
- **PDF Highlighting**: Highlights deviations for user review.
- **Summarization**: Provides concise summaries of contract content.
- **Optical Character Recognition (OCR)**: Processes scanned contract documents.

2.4 Workflow

- **Document Upload**: Users upload contract documents via the front-end.
- **PDF Parsing**: The system parses documents to extract text and structure.
- **Classification and NER**: Extracted text is classified and key entities are identified.
- **Template Comparison**: Classified text is compared against templates to detect deviations.
- **PDF Highlighting**: Deviations are highlighted for user review.
- **Summarization and OCR**: Summarize the document and process any scanned images.
- **Display Results**: Results are presented on the front-end interface.

Intel AI Workbench Components

3.1 Application Components

3.1.1 PDF-Parser

Functionality:

- Extracts text from PDF documents to transform unstructured data into a structured format.

Key Features:

- Text extraction from PDFs.
- Metadata extraction including author, creation date, and other properties.
- Handling of corrupted or encrypted PDFs.

Implementation:

- Utilize libraries such as PyMuPDF (Fitz), PDFMiner, or Tika for robust PDF parsing.
- Implement pre-processing steps to clean and structure extracted text.

3.1.2 NER (Named Entity Recognition)

Functionality:

- Identifies and classifies entities within the text, such as names, dates, locations, and organizations.

Key Features:

- Entity detection within text.
- Categorization of entities into predefined types.
- Contextual analysis to improve entity recognition accuracy.

Implementation:

- Use NLP libraries like spaCy or Stanford NER.
- Train NER models on domain-specific data for improved accuracy.

3.1.3 Text Classifier

Functionality:

- Categorizes text into predefined categories or labels, such as identifying the type of contract, urgency, or relevant departments.

Key Features:

- Classification of text into multiple categories.
- Support for multi-label classification.
- Contextual understanding to improve classification accuracy.

Implementation:

- Utilize NLP libraries like scikit-learn, spaCy, or Hugging Face Transformers.
- Train classification models on labeled datasets specific to the domain.
- Implement pre-processing steps to clean and tokenize text for better model performance.

3.1.4 Text-Compare

Functionality:

- Analyzes text differences between contract documents and their templates.

Key Features:

- Text comparison algorithms to detect deviations.
- Customizable rules for comparison criteria.
- Highlighting or flagging differences for easy identification.

Implementation:

- Utilizes text comparison libraries and algorithms like Difflib.
- Implements custom rules and thresholds to determine significant deviations.

3.1.5 PDF-Highlighter

Functionality:

- Visually emphasizes deviations in the text by highlighting differences between the contract and its template.

Key Features:

- Text highlighting to mark differences with colors or annotations.
- Interactive interface for users to interact with highlighted text.
- Customizable settings for highlighting criteria and colors.

Implementation:

- Utilize PDF manipulation libraries such as PyMuPDF (Fitz) for text extraction and annotation.
- Implement algorithms to compare and highlight differences between documents.

3.1.6 Summarizer

Functionality:

- Generates concise summaries of contract documents, highlighting key differences and critical information.

Key Features:

- Text summarization techniques to condense lengthy documents.
- Emphasis on important deviations and changes.
- Adjustable settings for summary length and detail level.

Implementation:

- Employs text summarization models such as Gensim or BERT-based approaches.
- Trains summarization models on contract-specific datasets for accuracy.

3.1.7 OCR (Optical Character Recognition)

Functionality:

- Extracts text from scanned images and converts it into editable text format.

Key Features:

- Image to text conversion.
- Support for multiple languages and character sets.
- Pre-processing techniques to enhance accuracy.

Implementation:

- Utilizes OCR libraries like Tesseract or OpenCV.
- Incorporates image pre-processing steps such as denoising and binarization for improved text extraction.

3.2 Levels of project

3.2.1 Individual Components :- Each component of the system is developed and tested independently. This modular approach ensures that each functionality works correctly in isolation before integration.

- **PDF-Parser**: Extracts text and metadata from PDF documents.
- **NER**: Identifies and classifies entities within the text.
- **Text Classifier**: Categorizes text into predefined categories.
- **Text Compare**: Analyzes differences between contract documents and their templates.
- **PDF Highlighter**: Visually highlights deviations in the text.
- **Summarizer**: Generates concise summaries of contract documents.
- **OCR**: Extracts text from scanned images.

3.2.2 Flow :- The components work together in a defined sequence to process and validate contract documents.

- **Document Upload**: Users upload contract documents via the front-end.
- **PDF Parsing**: The system extracts text and structure from uploaded PDFs.
- **Classification and NER**: Extracted text is classified, and key entities are identified.
- **Template Comparison**: Classified text is compared against templates to detect deviations.
- **PDF Highlighting**: Deviations are highlighted for user review.
- **Summarization and OCR**: The document is summarized, and scanned images are processed.
- **Display Results**: Results are presented on the front-end interface.

3.2.3 Streamlit (Frontend) :- Streamlit is a powerful framework that allows for the rapid development of web applications with Python. It is particularly suited for creating data-driven applications and provides seamless integration with machine learning models and data visualization libraries. The Streamlit frontend serves as the user interface for the Business Contract Validation project. It provides an intuitive and interactive platform for users to upload documents, view processing results, and interact with highlighted deviations.

- **Document Upload :-** Users can upload PDF contract documents through a file uploader widget. Support for multiple file uploads simultaneously.
- **Real-time Processing Feedback :-** Display progress indicators while the document is being processed. Notify users of the processing status (e.g., parsing, classification, comparison).
- **Summarization and Key Points** :- Display a concise summary of the contract, highlighting key differences and critical information. Allow users to expand or collapse sections of the summary for detailed views.

3.2.4 Docker (Containerized) :- Docker containers encapsulate the application code, runtime, system tools, libraries, and settings. This isolation helps in maintaining consistency across various stages of development and deployment, reducing the "it works on my machine" problem. This ensures that the application runs consistently across different environments, from development to production. For the Business Contract Validation project, Docker will be used to containerize all components, making the system portable, scalable, and easy to deploy.

- **Frontend (Streamlit)**: The user interface for document upload and result display.
- **Backend (Flask)**: The Flask service for processing requests and interacting with the ML models.
- **Machine Learning Models**: The models for PDF parsing, NER, text classification, etc.
- **Database**: If the project requires persistent storage for user data, documents, or results.
- **Supporting Services**: Any other services like caching, message queues, etc.

3.2.5 Flask :- Flask is a modern, fast (high-performance), web framework for building APIs with Python 3.7+ based on standard Python type hints. It is designed to be easy to use, easy to deploy, and provides a powerful framework for building robust APIs. The Flask backend will serve as the core of the Business Contract Validation project, handling all the API requests, processing data, and interfacing with the machine learning models. The backend will be responsible for tasks such as parsing PDFs, performing Named Entity Recognition (NER), classifying text, comparing texts, and more.

- **API Endpoints** :- Endpoints for uploading and processing documents.Endpoints for retrieving results such as classified text, detected entities, and highlighted deviations.
- **Integration with Machine Learning Models :-** Load and utilize pre-trained ML models for various tasks. Ensure efficient model inference for real-time processing.
- **Database Interaction :-** Store and retrieve data such as processed documents, user annotations, and results.
- **Error Handling and Logging :-** Implement comprehensive error handling and logging to monitor and debug the application effectively.

3.2.6 Performance Engineering :- Performance engineering focuses on ensuring that the application meets required performance criteria such as response time, throughput, scalability, and resource usage. For the Business Contract Validation project, performance engineering involves a combination of proactive design, continuous monitoring, and iterative optimization to ensure the system operates efficiently under various loads.

- **Profiling and Benchmarking** :- Identify performance bottlenecks in the application.Measure and compare the performance of various components.
- **Load Testing :-** Simulate real-world load on the application to ensure it can handle expected user traffic. Identify the system’s breaking point and optimize accordingly.
- **Monitoring and Logging :-** Continuously monitor the application’s performance in real-time. Collect and analyze logs to identify and resolve performance issues.

- **Optimization :-** Optimize code, database queries, and infrastructure to improve performance. Implement caching mechanisms to reduce redundant processing.
- **Scalability Planning :-** Design the system architecture to support horizontal and vertical scaling. Implement auto-scaling features to adjust resources based on load.

3.2.7 Serving :- Serving refers to deploying the Business Contract Validation application in a way that makes it accessible to users. This includes setting up the infrastructure, deploying the application, managing updates, and ensuring the application runs smoothly in a production environment. The aim is to ensure that users can interact with the system efficiently and reliably.

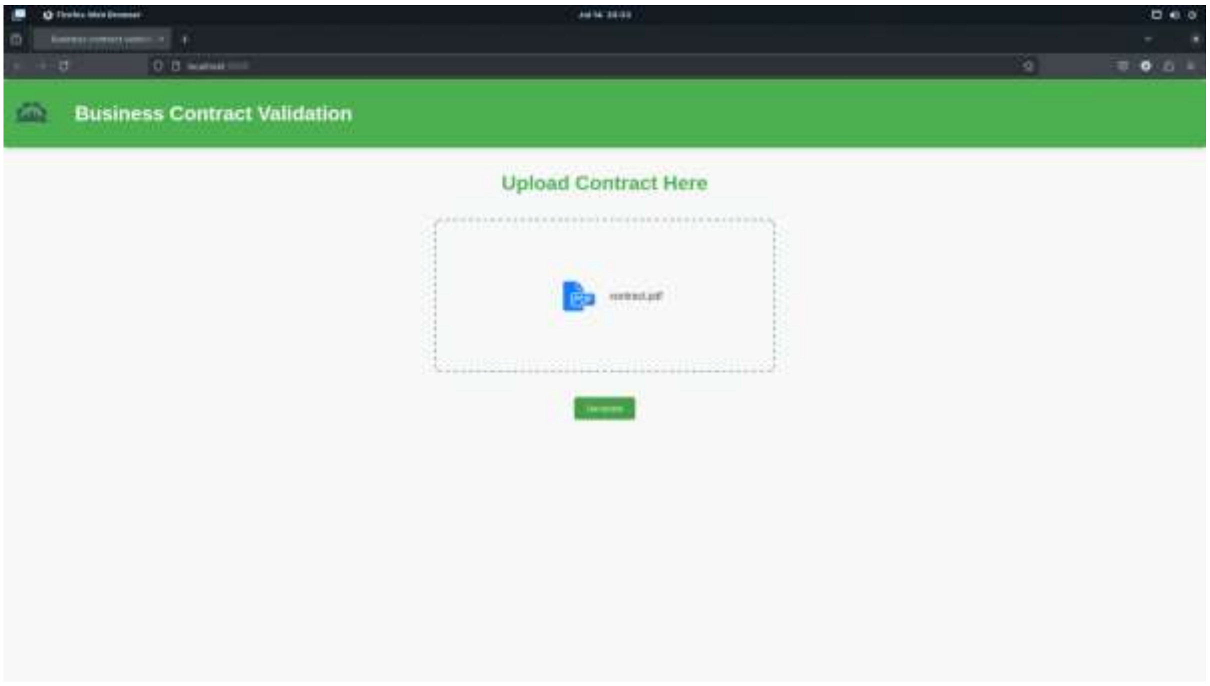
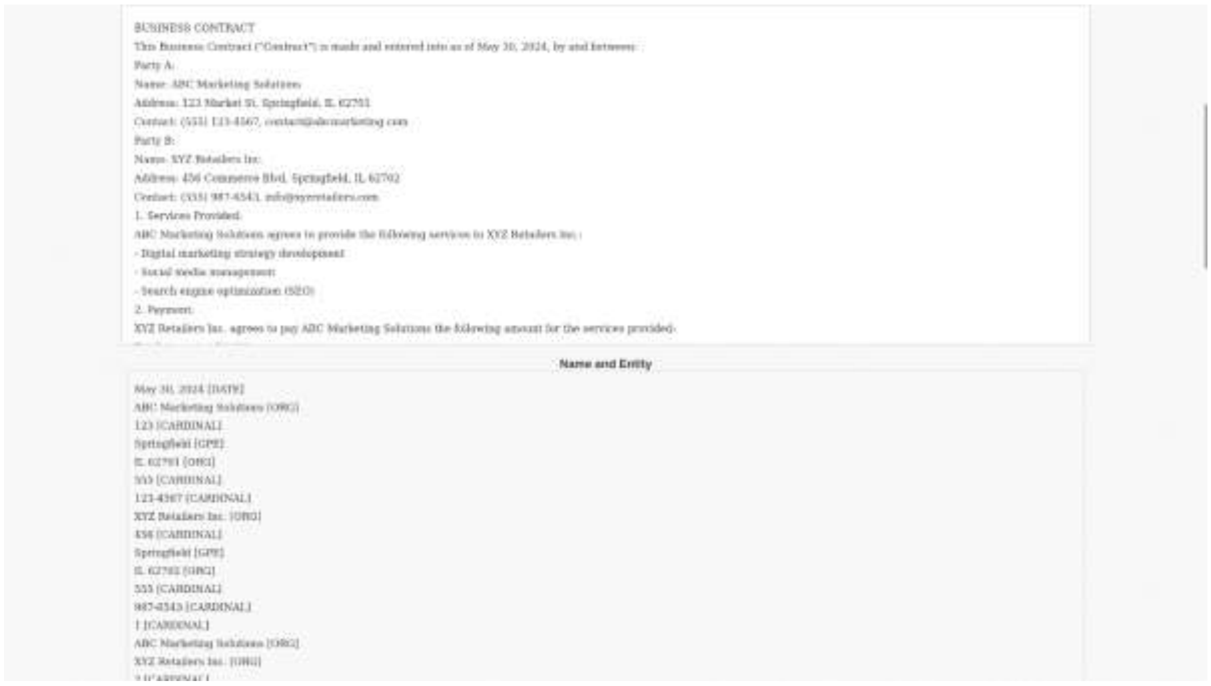


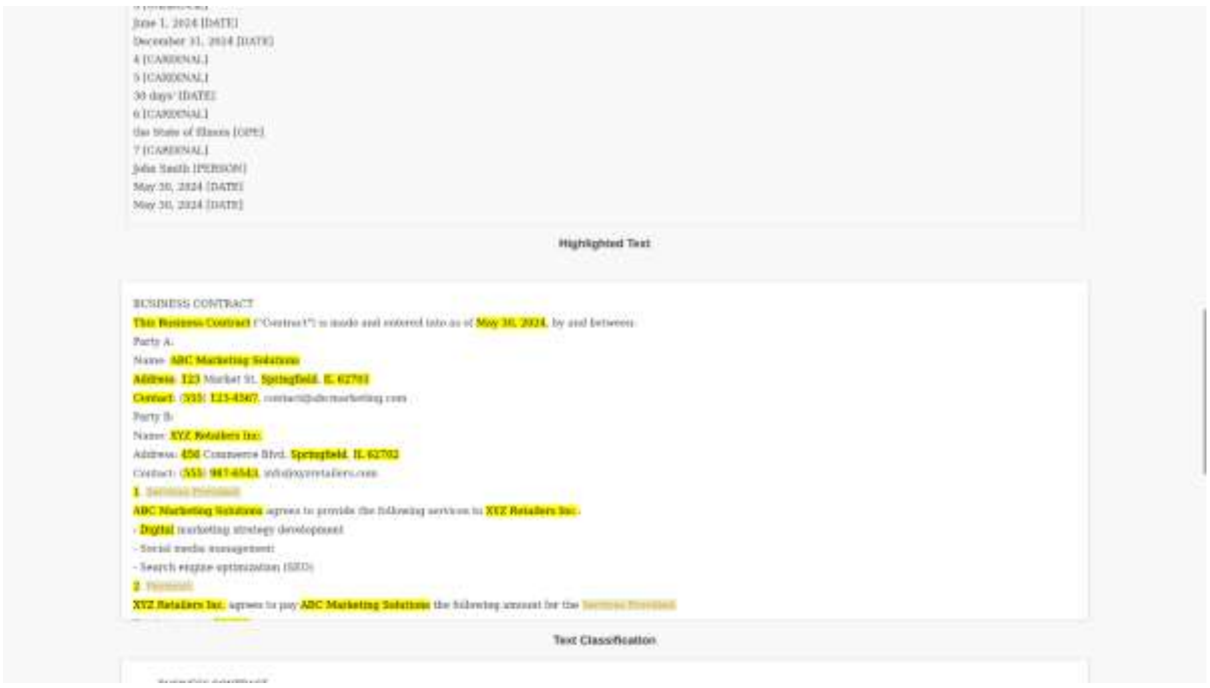
Figure 1 Home Page


BUSINESS CONTRACT
This Business Contract ("Contract") is made and entered into as of May 30, 2024, by and between:
Party A:
Name: ABC Marketing Solutions
Address: 123 Market St, Springfield, IL 62701
Contact: (555) 123-4567, contact@abcmarketing.com
Party B:
Name: XYZ Retailers Inc.
Address: 456 Commerce Blvd, Springfield, IL 62702
Contact: (555) 987-6543, info@xyzretailers.com
1. Services Provided:
ABC Marketing Solutions agrees to provide the following services to XYZ Retailers Inc.:
- Digital marketing strategy development
- Social media management
- Search engine optimization (SEO)
2. Payment:
XYZ Retailers Inc. agrees to pay ABC Marketing Solutions the following amount for the services provided:

Name and Entity

May 30, 2024 [DATE]
ABC Marketing Solutions [ORG]
123 [CARDINAL]
Springfield [GPE]
IL 62701 [ORG]
555 [CARDINAL]
123-4567 [CARDINAL]
XYZ Retailers Inc. [ORG]
456 [CARDINAL]
Springfield [GPE]
IL 62702 [ORG]
555 [CARDINAL]
987-6543 [CARDINAL]
1 [CARDINAL]
ABC Marketing Solutions [ORG]
XYZ Retailers Inc. [ORG]

Figure 2 Extracted Text



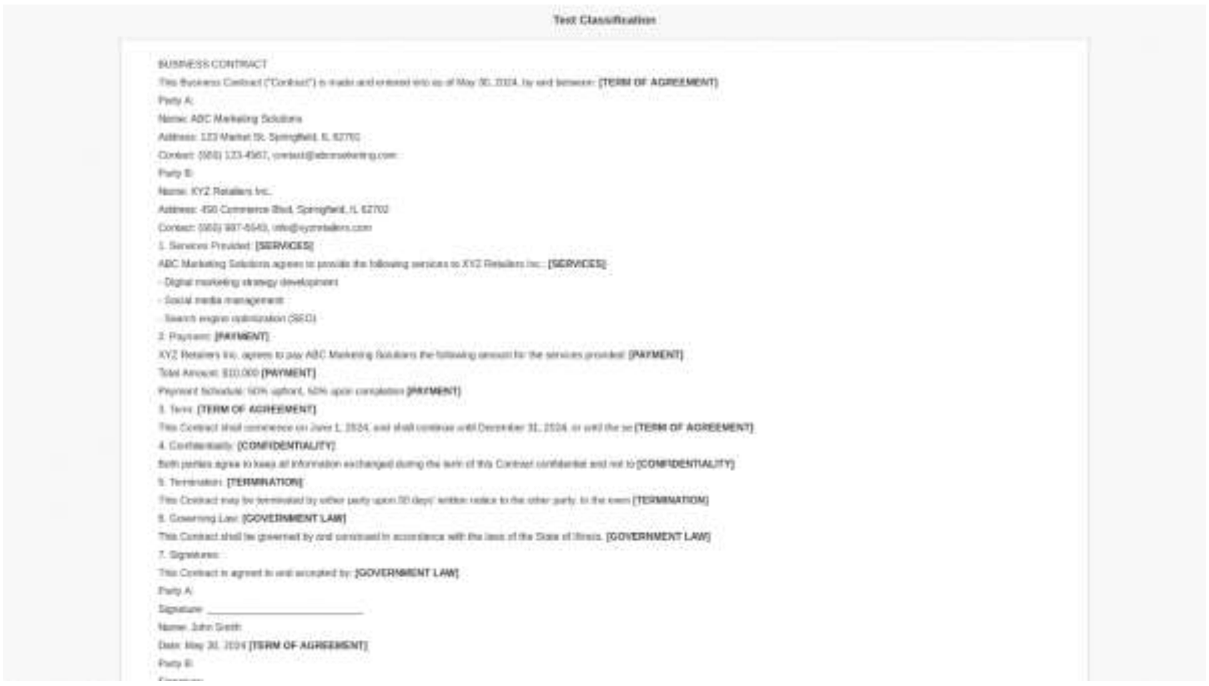
June 1, 2024 [DATE]
December 31, 2024 [DATE]
4 [CARDINAL]
5 [CARDINAL]
30 days' [DATE]
6 [CARDINAL]
the State of Illinois [GPE]
7 [CARDINAL]
John Smith [PERSON]
May 30, 2024 [DATE]
May 30, 2024 [DATE]

Highlighted Text

BUSINESS CONTRACT
This Business Contract ("Contract") is made and entered into as of May 30, 2024, by and between:
Party A:
Name: ABC Marketing Solutions
Address: 123 Market St, Springfield, IL 62701
Contact: (555) 123-4567, contact@abcmarketing.com
Party B:
Name: XYZ Retailers Inc.
Address: 456 Commerce Blvd, Springfield, IL 62702
Contact: (555) 987-6543, info@xyzretailers.com
1. Services Provided
ABC Marketing Solutions agrees to provide the following services to XYZ Retailers Inc.:
- Digital marketing strategy development
- Social media management
- Search engine optimization (SEO)
2. Payment
XYZ Retailers Inc. agrees to pay ABC Marketing Solutions the following amount for the Services Provided

Text Classification

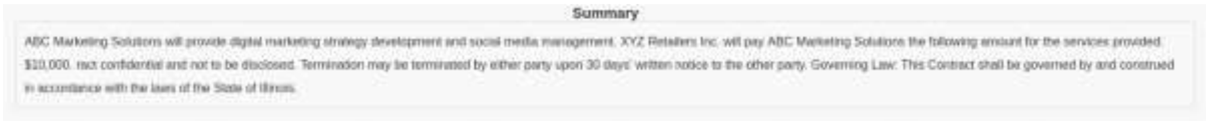
Figure 3 Highlighted Text



Text Classification

BUSINESS CONTRACT
This Business Contract ("Contract") is made and entered into as of May 30, 2024, by and between: **[TERM OF AGREEMENT]**
Party A:
Name: ABC Marketing Solutions
Address: 123 Market St, Springfield, IL 62701
Contact: (555) 123-4567, contact@abcmarketing.com
Party B:
Name: XYZ Retailers Inc.
Address: 456 Commerce Blvd, Springfield, IL 62702
Contact: (555) 987-6543, info@xyzretailers.com
1. Services Provided: **[SERVICES]**
ABC Marketing Solutions agrees to provide the following services to XYZ Retailers Inc.: **[SERVICES]**
- Digital marketing strategy development
- Social media management
- Search engine optimization (SEO)
2. Payment: **[PAYMENT]**
XYZ Retailers Inc. agrees to pay ABC Marketing Solutions the following amount for the services provided: **[PAYMENT]**
Total Amount: $10,000 **[PAYMENT]**
Payment Schedule: 50% upfront, 50% upon completion **[PAYMENT]**
3. Term: **[TERM OF AGREEMENT]**
This Contract shall commence on June 1, 2024, and shall continue until December 31, 2024, or until the se **[TERM OF AGREEMENT]**
4. Confidentiality: **[CONFIDENTIALITY]**
Both parties agree to keep all information exchanged during the term of this Contract confidential and not to **[CONFIDENTIALITY]**
5. Termination: **[TERMINATION]**
This Contract may be terminated by either party upon 30 days' written notice to the other party. In the even **[TERMINATION]**
6. Governing Law: **[GOVERNMENT LAW]**
This Contract shall be governed by and construed in accordance with the laws of the State of Illinois. **[GOVERNMENT LAW]**
7. Signatures:
This Contract is agreed to and accepted by: **[GOVERNMENT LAW]**
Party A:
Signature: ______________________
Name: John Smith
Date: May 30, 2024 **[TERM OF AGREEMENT]**
Party B:

Figure 4 Classification



Summary

ABC Marketing Solutions will provide digital marketing strategy development and social media management. XYZ Retailers Inc. will pay ABC Marketing Solutions the following amount for the services provided: $10,000. ract confidential and not to be disclosed. Termination may be terminated by either party upon 30 days' written notice to the other party. Governing Law: This Contract shall be governed by and construed in accordance with the laws of the State of Illinois.

Figure 5 Summary

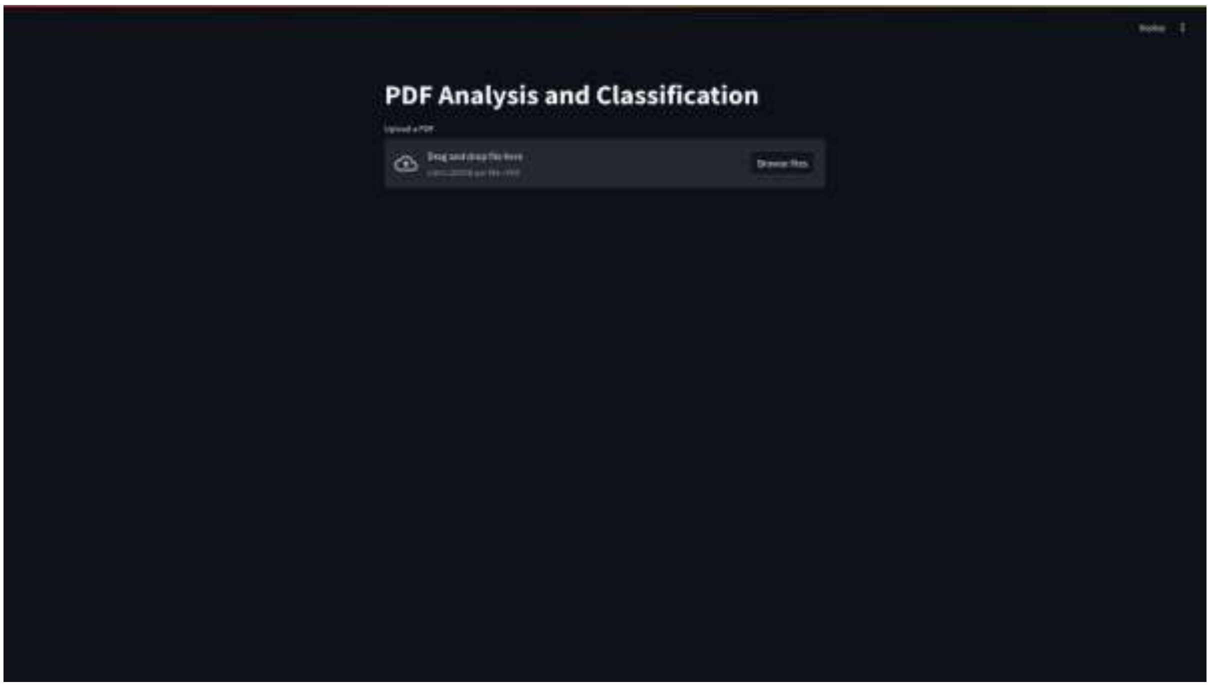
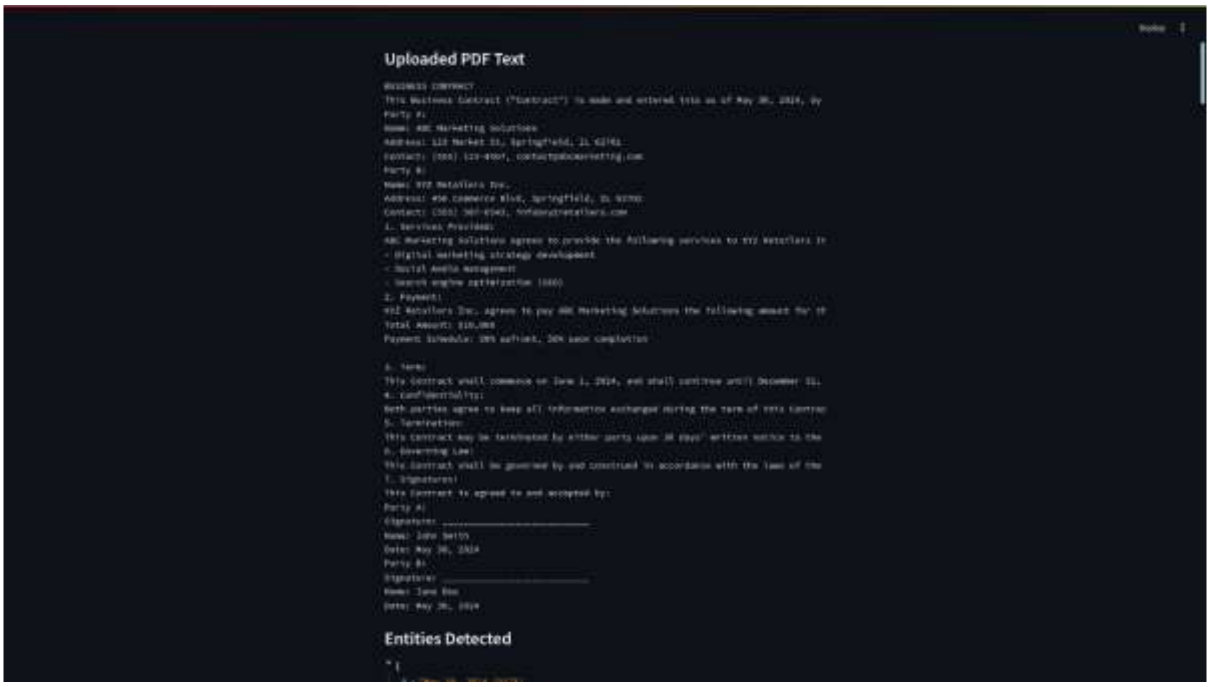


Figure 6 StreamLit Upload



Uploaded PDF Text

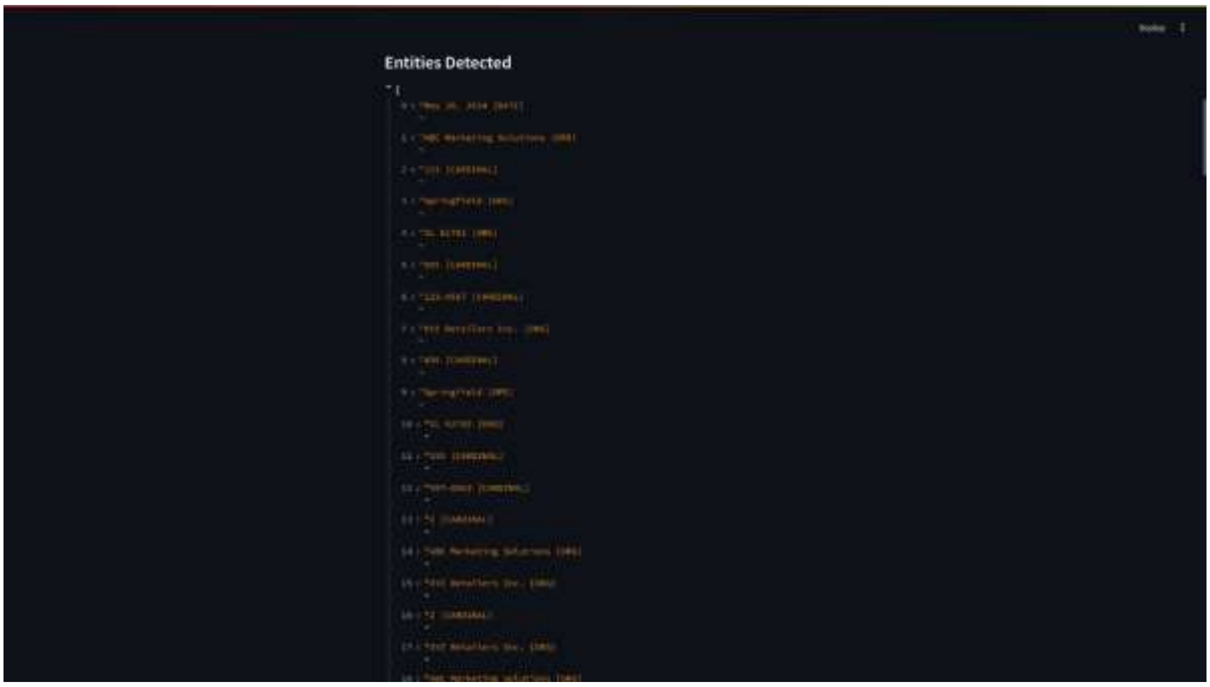
```
BUSINESS CONTRACT
This Business Contract ("Contract") is made and entered into as of May 30, 2024, by
Party A:
Name: ABC Marketing Solutions
Address: 123 Market St, Springfield, IL 62701
Contact: (555) 123-4567, contact@abcmarketing.com
Party B:
Name: XYZ Retailers Inc.
Address: 456 Commerce Blvd, Springfield, IL 62702
Contact: (555) 987-6543, info@xyzretailers.com
1. Services Provided:
ABC Marketing Solutions agrees to provide the following services to XYZ Retailers In
- Digital marketing strategy development
- Social media management
- Search engine optimization (SEO)
2. Payment:
XYZ Retailers Inc. agrees to pay ABC Marketing Solutions the following amount for th
Total Amount: $10,000
Payment Schedule: 50% upfront, 50% upon completion

3. Term:
This Contract shall commence on June 1, 2024, and shall continue until December 31,
4. Confidentiality:
Both parties agree to keep all information exchanged during the term of this Contrac
5. Termination:
This Contract may be terminated by either party upon 30 days' written notice to the
6. Governing Law:
This Contract shall be governed by and construed in accordance with the laws of the
7. Signatures:
This Contract is agreed to and accepted by:
Party A:
Signature: ________________________
Name: John Smith
Date: May 30, 2024
Party B:
Signature: ________________________
Name: Jane Doe
Date: May 30, 2024
```

Entities Detected

```
▾ [
```

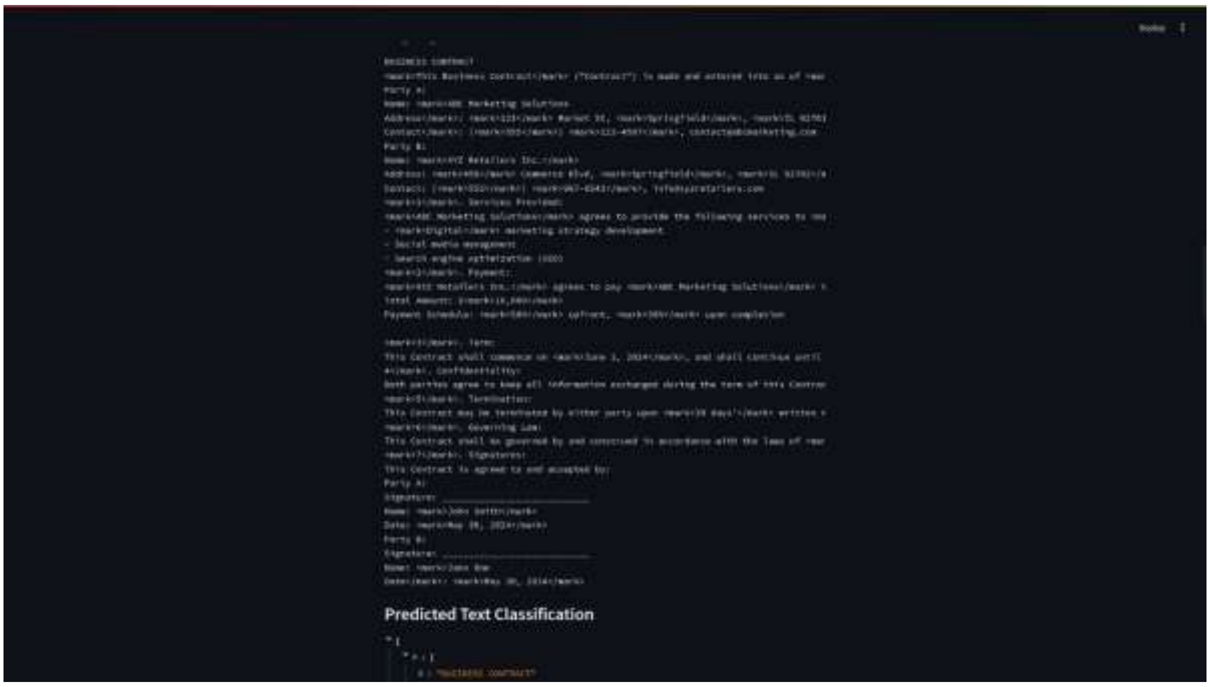
Figure 7 PDF Extract



Entities Detected

```
▾[
  0 : "May 20, 2024 [DATE]"
  1 : "ABC Marketing Solutions [ORG]"
  2 : "123 [CARDINAL]"
  3 : "Springfield [GPE]"
  4 : "IL 62701 [ORG]"
  5 : "555 [CARDINAL]"
  6 : "123-4567 [CARDINAL]"
  7 : "XYZ Retailers Inc. [ORG]"
  8 : "456 [CARDINAL]"
  9 : "Springfield [GPE]"
  10 : "IL 62702 [ORG]"
  11 : "555 [CARDINAL]"
  12 : "987-6543 [CARDINAL]"
  13 : "1 [CARDINAL]"
  14 : "ABC Marketing Solutions [ORG]"
  15 : "XYZ Retailers Inc. [ORG]"
  16 : "2 [CARDINAL]"
  17 : "XYZ Retailers Inc. [ORG]"
  18 : "ABC Marketing Solutions [ORG]"
```

Figure 8 Entities



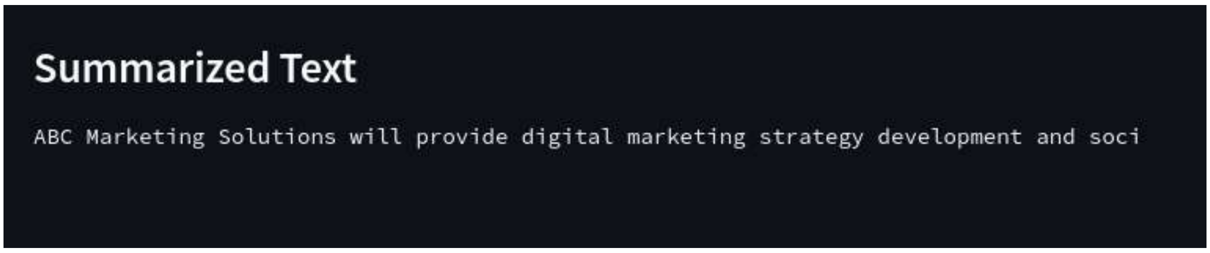
```
BUSINESS CONTRACT
<mark>This Business Contract</mark> ("Contract") is made and entered into as of <mar
Party A:
Name: <mark>ABC Marketing Solutions
Address</mark>: <mark>123</mark> Market St, <mark>Springfield</mark>, <mark>IL 62701
Contact</mark>: (<mark>555</mark>) <mark>123-4567</mark>, contact@abcmarketing.com
Party B:
Name: <mark>XYZ Retailers Inc.</mark>
Address: <mark>456</mark> Commerce Blvd, <mark>Springfield</mark>, <mark>IL 62702</m
Contact: (<mark>555</mark>) <mark>987-6543</mark>, info@xyzretailers.com
<mark>1</mark>. Services Provided:
<mark>ABC Marketing Solutions</mark> agrees to provide the following services to <ma
- <mark>Digital</mark> marketing strategy development
- Social media management
- Search engine optimization (SEO)
<mark>2</mark>. Payment:
<mark>XYZ Retailers Inc.</mark> agrees to pay <mark>ABC Marketing Solutions</mark> t
Total Amount: $<mark>10,000</mark>
Payment Schedule: <mark>50%</mark> upfront, <mark>50%</mark> upon completion

<mark>3</mark>. Term:
This Contract shall commence on <mark>June 1, 2024</mark>, and shall continue until
4</mark>. Confidentiality:
Both parties agree to keep all information exchanged during the term of this Contrac
<mark>5</mark>. Termination:
This Contract may be terminated by either party upon <mark>30 days'</mark> written n
<mark>6</mark>. Governing Law:
This Contract shall be governed by and construed in accordance with the laws of <mar
<mark>7</mark>. Signatures:
This Contract is agreed to and accepted by:
Party A:
Signature: ______________________
Name: <mark>John Smith</mark>
Date: <mark>May 20, 2024</mark>
Party B:
Signature: ______________________
Name: <mark>Jane Doe
Date</mark>: <mark>May 20, 2024</mark>
```

Predicted Text Classification

```
▾[
  ▾0 : [
    0 : "BUSINESS CONTRACT"
```

Figure 9 Highlighted (using <mark></mark>)



Summarized Text

```
ABC Marketing Solutions will provide digital marketing strategy development and soci
```

Figure 10 Summarize

Conclusion

5.1 Summary

The Business Contract Validation project automates the process of parsing, classifying, and validating business contracts. By leveraging machine learning, generative AI, and advanced front-end development techniques, it enhances efficiency, accuracy, and compliance in contract review processes.

5.2 Future Work

- **Additional Features**: Incorporate additional functionalities like personal information masking, profanity filtering, and sentiment classification.
- **Scalability**: Enhance the scalability of the application to handle larger volumes of contracts.
- **User Interface**: Improve the user interface for better user experience.
- **Integration**: Integrate with other business systems for seamless contract management.

References

- spaCy NER Documentation
- nltk
- pymupdf
- streamlit
- flask
- docker
- pytesseract