Business Contract Validation - To Classify Content within the Contract Clauses and Determine Deviations from Templates and highlight them

Team Neuro Nexus

Institutional Affiliation(s)**:** NITTE Meenakshi Institute of Technology

Intel Unnati Industrial Training

I. Introduction:

The complexity and dynamic nature of business contracts pose significant challenges in ensuring their accuracy and compliance. As business practices continuously evolve, these contracts must be meticulously reviewed to capture every detail accurately and ensure they adhere to standardized templates. This necessity underscores the importance of developing sophisticated tools that can parse, classify, and validate the content within contracts efficiently. Effective contract management is essential for mitigating legal risks, ensuring regulatory compliance, and maintaining the integrity of business agreements. Without reliable validation methods, the potential for misunderstandings, disputes, or legal issues increases significantly, making the development of advanced contract validation solutions crucial for modern businesses.

Business contracts are formal legal documents composed of numerous clauses and sub-clauses that outline the rights, obligations, and responsibilities of the parties involved. Parsing these documents to extract key details is the first critical step in structuring them for further analysis. Each contract typically follows an associated template, and deviations from this template can lead to misunderstandings, disputes, or even legal issues.

The main challenge lies in developing a method to accurately classify the contents of these parsed documents according to their clauses and sub-clauses, and then determining any deviations from the standard templates. This task is crucial for maintaining the integrity and enforceability of contracts. Therefore, our focus is on creating an AI-driven solution to streamline this process.

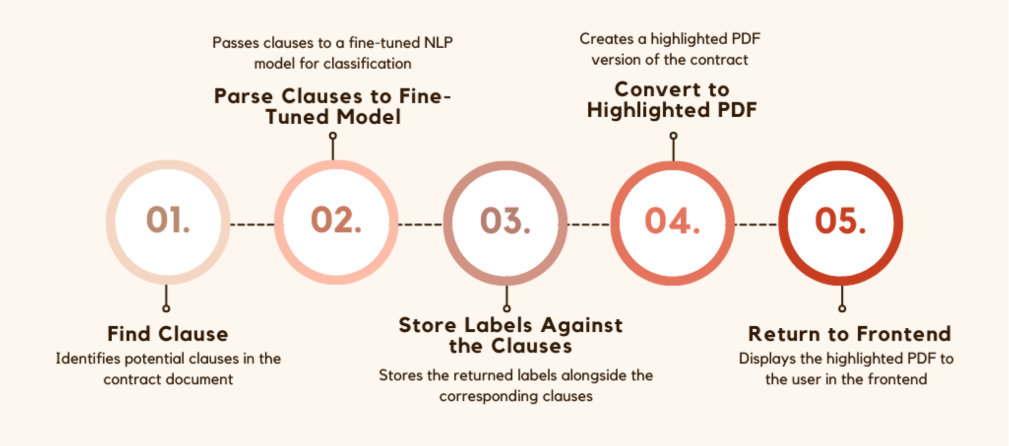
II. Literature Survey and Related Work:

- **BERT and Beyond:** The Evolution of Legal NLP Models While BERT has been a pivotal model in NLP for various applications, newer models like Phi 3 are being explored for specialized tasks such as clause identification in contracts. Phi 3, specifically designed to handle complex legal texts, offers improved accuracy and efficiency.
- **Phi 3:** Advanced Transformer Models for Legal Text: Phi 3 a more recent model, leverages advancements in transformer architectures to enhance the understanding

of legal language. This model has been fine-tuned on extensive legal corpora, making it adept at identifying and categorizing contract clauses with high precision.

- **Comparative Analysis of Phi 3 and BERT:** Studies comparing Phi 3 to BERT in legal contexts have shown that Phi 3 outperforms BERT in tasks requiring deep understanding of contractual language. Phi 3's architecture incorporates additional layers and attention mechanisms tailored for legal texts, resulting in better performance in identifying nuanced legal clauses.

III. Methodology


Passes clauses to a fine-tuned NLP model for classification
Parse Clauses to Fine-Tuned Model
Creates a highlighted PDF version of the contract
Convert to Highlighted PDF
01.
02.
03.
04.
05.
Find Clause
Identifies potential clauses in the contract document
Store Labels Against the Clauses
Stores the returned labels alongside the corresponding clauses
Return to Frontend
Displays the highlighted PDF to the user in the frontend

1. Finding Clauses

- **Description:** The initial step is to identify potential clauses within the contract documents. This is done using a find-clauses function which scans the document and marks sections that fit predefined criteria for clauses.

- **Tools Used:** Natural Language Processing (NLP) techniques and regular expressions to detect and extract clauses**.**

2. Parsing Clauses to Fine-Tuned Model

- **Description:** The extracted clauses are then passed to a fine-tuned language model (e.g., Phi 3) for classification. The model has been trained on a labeled dataset of contract clauses and is capable of identifying specific types of clauses such as "exclusivity."

- **Special Handling:** If a clause is identified as an exclusivity clause, it is passed to another specialized model for further detailed analysis.

- **Output:** The model returns labels that categorize each clause.

3. Storing Labels Against the Clauses

- **Description:** The labels returned by the model are stored alongside their corresponding clauses in the document. This step ensures that each clause is properly annotated with its classification for easy reference.

- **Tools Used:** Data structures such as dictionaries or databases to store clause-label pairs.

4. Converting to Highlighted PDF

- **Description:** The labeled clauses are used to generate a highlighted PDF version of the contract. Each type of clause is highlighted in a different color or style, which helps in visually distinguishing between various clause types.
- **Tools Used:** PDF libraries (e.g., PyPDF2) to generate and manipulate PDF documents.

5. Returning to Frontend

- **Description:** The final highlighted PDF is then returned to the frontend application. This allows users to interact with the document, view the highlighted clauses, and gain a better understanding of the contract's structure.

- **Tools Used:** Web frameworks and APIs to handle the communication between the backend processing and the frontend display.

IV. Dataset Selection:

Data Sources

For this project, we utilized a dataset containing legal contract clauses. The dataset includes various types of clauses such as confidentiality, indemnification, and termination clauses. These datasets were sourced from publicly available legal documents and annotated by legal experts to ensure the accuracy and relevance of the clause classifications.

Data Preprocessing

Data preprocessing involved several key steps:

- **Data Cleaning**: Removing any extraneous text, correcting formatting issues, and ensuring consistency in the data.

- **Tokenization**: Breaking down the text into individual tokens or words.

- **Stopword Removal**: Filtering out common words that do not contribute to clause identification (e.g., 'and', 'the').

- **Stemming/Lemmatization**: Reducing words to their base or root forms to ensure uniformity.

Data Augmentation

To enhance the model's performance, we applied data augmentation techniques such as:

- **Synonym Replacement**: Replacing certain words with their synonyms to create variations in the dataset.

- **Back Translation**: Translating clauses to another language and back to introduce diversity while maintaining meaning.

Dataset Split

The dataset was divided into training, validation, and test sets in the ratio of 70:15:15 to ensure that the model is evaluated on unseen data and to prevent overfitting.

V. Metric & Model Selection:

Metrics

1. **Accuracy**
   - **Definition**: The ratio of correctly predicted clause labels to the total number of clauses.
   - **Importance**: Measures overall correctness but may not be sufficient alone, especially in imbalanced datasets.
2. **Precision**
   - **Definition**: The ratio of correctly predicted positive observations to the total predicted positives.
   - **Formula**: Precision = $\frac{TP}{TP+FP}$
   - **Importance**: High precision indicates a low false positive rate, crucial for correctly identifying specific clause types.
3. **Recall (Sensitivity)**
   - **Definition**: The ratio of correctly predicted positive observations to all observations in the actual class.
   - **Formula**: Recall=$\frac{TP}{TP+FN}$
   - **Importance**: High recall indicates a low false negative rate, important for ensuring all relevant clauses are identified.
4. **F1 Score**
   - **Definition**: The harmonic mean of precision and recall.
   - **Formula:** F1 = 2* $\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$
   - **Importance**: Balances precision and recall, providing a single metric that accounts for both false positives and false negatives.
5. **Confusion Matrix**
   - **Definition**: A table that shows the actual vs. predicted classifications.
   - **Importance**: Provides a detailed breakdown of true positives, false positives, true negatives, and false negatives, helping to understand model performance in a more granular way.
6. **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)**

   - **Definition**: Measures the ability of the model to distinguish between classes.
   - **Importance**: A higher AUC indicates better model performance in differentiating clause types.

Model Evaluation

1. **Training and Validation Split**
   - **Process**: The dataset is split into training and validation sets to evaluate the model's performance on unseen data.
   - **Importance**: Helps to ensure that the model generalizes well and is not overfitting to the training data.
2. **Cross-Validation**
   - **Process**: The data is divided into multiple folds, and the model is trained and evaluated on each fold.
   - **Importance**: Provides a more robust estimate of model performance by using multiple training and validation splits.
3. **Hyperparameter Tuning**
   - **Process**: Optimization of model parameters using techniques such as Grid Search or Random Search.
   - **Importance**: Enhances model performance by finding the best set of parameters for the task.
4. **Error Analysis**
   - **Process**: Analysis of misclassified clauses to understand and rectify the model's weaknesses.
   - **Importance**: Provides insights into specific types of errors, helping to improve the model further.
5. **Real-World Testing**
   - **Process**: Evaluation of the model on real-world contracts and documents.
   - **Importance**: Ensures the model performs well in practical, everyday scenarios beyond the curated datasets.

By applying these metrics and evaluation methods, we ensure that our clause identification and classification model is both accurate and reliable, providing significant value in legal document analysis.

VI. References:

- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review. arXiv preprint arXiv:2103.06268.