Natural Language Processing and Text Analysis

A PROJECT REPORT

ON

Samvidhan AI (AI-based Legal Contract Generator)

Submitted in partial fulfilment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE AND ENGINEERING

Submitted By Group-10:

Akankshi Gera	220663
Lakshya Bindal	220416
Manan Khandelwal	220646
Sagar Sindhu	220372

Under The Supervision Of:
Dr. Atul Mishra

SCHOOL OF ENGINEERING AND TECHNOLOGY
BML MUNJAL UNIVERSITY, GURUGRAM

MAY 2025

Table of Contents

INTRODUCTION

Modern digital transformation has exposed legal documentation especially legally binding contracts as one of the few areas where manual intervention prevails predominantly. Creating vital agreements normally requires both extensive specialized legal insight as well as detailed precision and demands substantial amounts of precious time for completion. Notwithstanding the pervasive digital evolution witnessed across diverse industries, the fundamental process of creating essential agreements – ranging from sensitive non-disclosure agreements safeguarding proprietary information to foundational employment contracts defining working relationships and comprehensive service terms outlining obligations – continues to rely heavily on legal professionals laboriously tailoring each document to address the unique nuances of every new client or specific business need. The continued manual drafting process leads to operational inefficiencies and delays which increases the chances of human errors that could eventually result in both price spikes through disputes and compliance-related complications.

The proposed research analysis works toward creating and assessing a semi-automated system which generates legal documents through the use of natural language processing (NLP) technology to optimize efficiency. The overarching objective of this study is threefold: first, to explore the inherent feasibility and practical limitations of a rule-based approach to contract generation, predicated on the structured and accurate extraction of key entities ; second, to develop a tangible and functional prototype that effectively demonstrates this capability through the provision of illustrative and relevant practical examples; and third, to conduct a thorough and insightful comparison of its performance characteristics and overall utility against both traditional, time-intensive manual drafting practices and contemporary AI-heavy alternatives, particularly those employing computationally intensive and often less transparent large language models (LLMs). The proposed method features explicit clarity during operations along with resource-saving design while enabling users to actively participate during contract creation processes.

The key value of this project is its capability to make legal document development accessible to all users. Businesses together with startups along with human resources teams and independent contractors frequently do not have sufficient budget to obtain legal counsel for standard agreements. The project delivers an efficient tool that enables users without technical expertise or legal background to produce legally valid documents at affordable prices. Legal firms receive advantages from this system because it enables them to focus on intensive advisory work while the system completes repetitive documentation tasks.

The complex nature of legal language proves to be the most significant obstacle when generating legal contracts automatically. Contract written materials present domain-specific terms which machines face trouble interpreting due to standard precise language constructions. Multiple legal requirements pose ongoing barriers to developing a single solution because they vary according to jurisdiction type and contract type and agreement nature. System operations

face a significant challenge regarding the correct identification of essential legal entities within user material that includes both party names and dates together with payment details and governing law specifications. When incorrect entity identification or placement occurs, it damages both the legal validity and enforceability of contractual contracts. A reliable rule-based system needs to maintain logical clause consistency together with managing exceptional terms and conditional elements while increasing its technical complexity.

The research develops a strategic rule-based system through NLP that utilizes SpaCy library NER capabilities with precision to detect essential legal elements in contracts with high precision. The system employs Jinja2 templating engine operations to execute powerful and flexible automation of entity detection population into legal templates. The implemented system architecture processes various contract models without losing logical clarity and keeps documents consistent without requiring extensive computational power or unpredictable results that appear when generative language models function. This study has designed a method which obtains vital legal reasoning information from the large CaseHOLD database to produce contractual text that respects existing legal frameworks and past cases.

The document presents an organized format dedicated to supplying detailed information about the planned system alongside its essential contributions. **Section 2** of this work provides an extensive review of research in legal text processing that concentrates on relevant datasets and algorithms for developing contracts and legal reasoning systems. This study presents its research methodology through **Section 3** by showing system visual diagrams along with architectural model information and theoretical explanations of the NLP rule-based framework. The research assesses system performance by using precision, recall and F-score metrics while showing enhanced capabilities compared to previous approaches in **Section 4**. In **Section 5** the paper delivers a summary of primary accomplishments along with discussions about confronted hurdles and offers recommendations for future advancements and practical applications.

LITERATURE REVIEW

The research aims to evaluate previous technological solutions which handle automated legal document processing in addition to natural language understanding within legal domains. The section analyzes previously used algorithms together with methodologies and datasets which researchers deployed to solve comparable issues in legal text analysis, contract generation and case law interpretation. Knowledge of present methods enables our proposed system to grasp its position in legal NLP applications and identify optimization spaces for better execution.

There exists numerous machine learning and NLP methods for classifying legal texts and contract analysis jobs. Named Entity Recognition (NER) stands as a primary approach for extracting legal entities including parties along with dates and jurisdictions. The flexibility together with customizable naming processes makes SpaCy and Stanford NLP popular applications in legal named entity recognition operations. SVMs and Random Forests are seen to use in document classification applications where they delivered high performance for many structured legal text analyses. The recent trend in legal text analysis employs BERT (Bidirectional Encoder Representations from Transformers) with its specialized variant LegalBERT to understand both linguistic directions and semantic relationships present in legal written documents. The employed models demonstrate notable enhancement in legal operations related to entailment detection and classification together with similarity measures. Universal rule-based systems maintain importance when generating contracts because they guarantee precise legal compliance in addition to being easy to understand.

More advanced models in the form of BERT, RoBERTa, and CaseLaw-BERT have recently been released that capture both the representation of a word in context as well as the meaning of the words in terms of surrounding context. These models achieve the state-of-the-art results for the legal NLP tasks like CaseHOLD but do so at a high computational cost and the price of interpretability. For instance, the work by Chalkidis et al. (2020) established that domain specific pretraining helps in legal tasks. But hybrid models combining classical models (TF IDF) and dense embeddings (Glove) have been viable trade-offs in accuracy, readability, and deployment cost, e.g., legal document generation platforms for which readability along with ease of deployment to a new location constitute a requirement.

Document similarities and retrieval are yet another vital aspect in legal text analysis that has traditionally been driven by models like TF-IDF (Term Frequency-Inverse Document Frequency). In fact, TF-IDF provides a statistical representation of text illustrating the top-frequency words in a document (document by document), so it comes in very handy to locate holding or precedent for a legal matter. However, the TF-IDF score falls through when it comes to addressing the semantic relationships because several words connoting the same legal concept are involved. For this reason, word embedding models known as GloVe (Global Vectors for Word Representation), that map the words to dense vector spaces in which the semantic similarities are reflected in the geometry of the vectors for the words were developed.

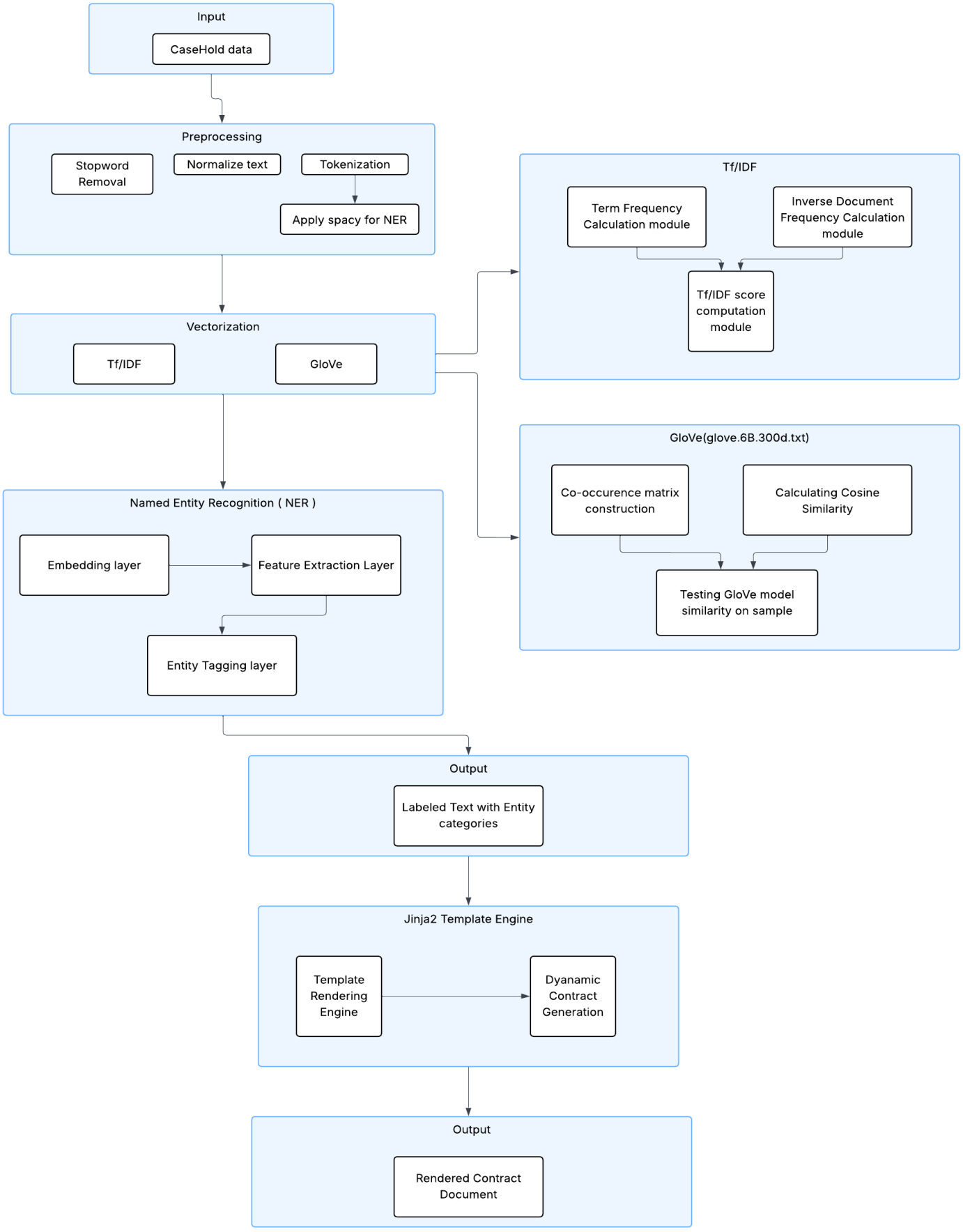
METHODOLOGY

This section offers an in-depth description of the automated system of generating legal contracts, including its architecture and functional units. The system is created to produce legally compliant and editable contracts by combining rule-based Natural Language Processing (NLP) and dynamic templating capabilities. The heart of the approach is the unity of pre-existing legal contract models and strong data extraction methodologies. The process starts with preprocessing the input from the user and then conducting Named Entity Recognition (NER) through the use of spaCy patterns to collect essential contractual details including names of the parties, dates, payment conditions, and jurisdiction. These entities are vital to maintaining legal precision and consistency.

Upon extraction, the recognized entities are directly inserted into the structured legal templates using the Jinja2 rendering engine. The engine replaces the placeholders in the contract template at runtime, and the automatically generated document is customized to the input parameters while preserving correctness in law. The method has an edge over the manual or black-box approach since it provides the benefit of interpretability, modularity, and accuracy.

In addition, the system is built for scalability and real-world applicability. From producing Non-Disclosure Agreements, employment contracts, and lease agreements, the pipeline is formatted to meet standard legal requirements and terminology. Through marrying automation with rule-based controls, the system minimizes errors in drafts, maximizes the pace, and decreases the cost of law—making it very convenient for startups, freelancers, HR departments, and law firms handling high-volume repetitive law paperwork.

SYSTEM DIAGRAM

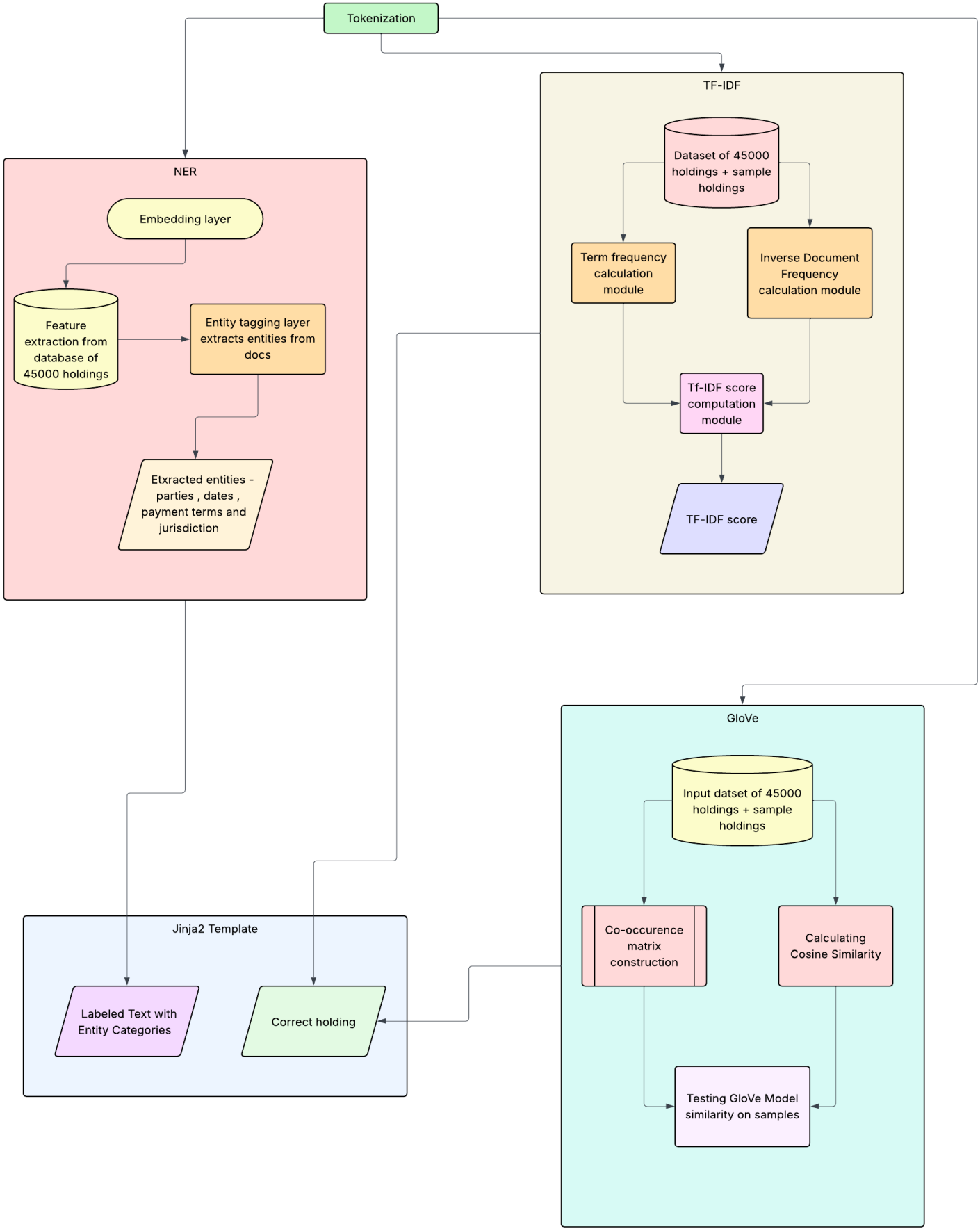


This system design is the end-to-end processing path for the automation of the creation of legal contracts with the help of rule-based NLP and template processing. It starts from the case data input and is followed by the preprocessing stage. This phase consists of stopwords elimination, textual normalization (for example, lowercasing), and tokenization. The next step is to prepare the text for Named Entity Recognition with spaCy. Preprocessing is done to clean and standardize the data to prepare it for downstream vectorization and entity extraction.

The second phase uses two main techniques—TF-IDF and GloVe—to transform textual data into numerical values. TF-IDF calculates the term frequency and inverse document frequency scores to emphasize critical words within the context of the data set. GloVe uses the co-occurrence matrix and uses the cost function during training to produce word embeddings of high density. These values are fed into the NER model, which has an embedding layer, feature extraction layer, and entity tagging layer to detect major elements such as party names, dates, payment terms, and jurisdiction.

The marked output from the NER layer in the form of labelled entity categories is then passed to the Jinja2 Template Engine. This is made up of the template rendering module and the dynamic contract generation unit. It inserts the entity data into pre-existing legal templates to generate tailored and compliant documents. The end product is the ready-to-be-used legal contract combining the accuracy of rule-based NLP and the flexibility of dynamic content creation.

ARCHITECTURE DESIGN



The data processing phase begins by receiving information from the CaseHold system after which it executes stopwords removal and text normalization and tokenization procedures. The preprocessed information receives entity recognition through SpaCy's Natural Entity Recognition (NER) software. Two methods of text vectorization follow the text cleaning process: TF/IDF (Term Frequency-Inverse Document Frequency) alongside GloVe (Global Vectors for Word Representation). The term significance is determined through the TF/IDF module by its built-in modules for term counts and reciprocal frequency calculations whereas GloVe analyzes semantic relationships via occurring word data training for vector representations.

TF Formula

$$TF(t,d) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ document\ d}{Total\ number\ of\ terms\ in\ doocument\ d}$$

IDF Formula

$$IDF(t) = log\frac{Total\ number\ of\ documents}{1 + Number\ of\ documents\ containing\ term\ t}$$

TF-IDF Formula

$$TF - IDF(t,d) = TF(t,d) * IDF(t)$$

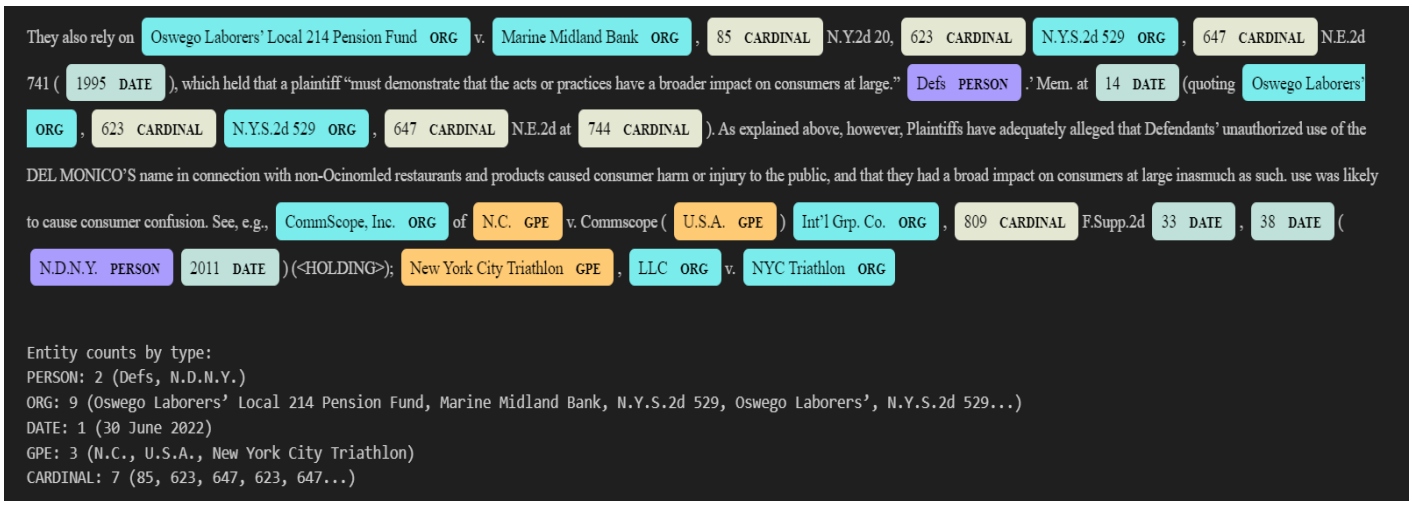
Cosine Similarity Function

$$cosine\ similarity = \frac{A.B}{||A||\ ||B||}$$

Where:

A and B are the vectors being compared.
(.) denotes the dot product of the vectors and $||A||\ ||B||$ are the magnitudes of the vectors.

A **Named Entity Recognition (NER)** model operates with three components to process extracted data features consisting of Embedding Layer followed by Feature Extraction Layer before reaching Entity Tagging Layer. All these layers work jointly to convert the input data into structured results while marking text spans with entity labels that span entities like Parties, Dates, Payment Terms and Jurisdiction. TF/IDF and GloVe provide embedding process enhancement which significantly boosts the entity extraction accuracy allowing the maintenance of legal relevance and output integrity.



The Jinja2 Template Engine receives successfully labeled entities. The Template Rendering Engine automatically inserts data into established contractual templates which then creates a complete document with legal format. Users obtain a complete contract document prepared for legal review following template processing and output generation.

TOOLS USED:

NLTK, SpaCy for NLP tasks like preprocessing and NER. TF/IDF and GloVe is used for vectorization. Jinja2 is used for template generation. And Streamlit is the platform used to display the frontend output.

SUMMARY:

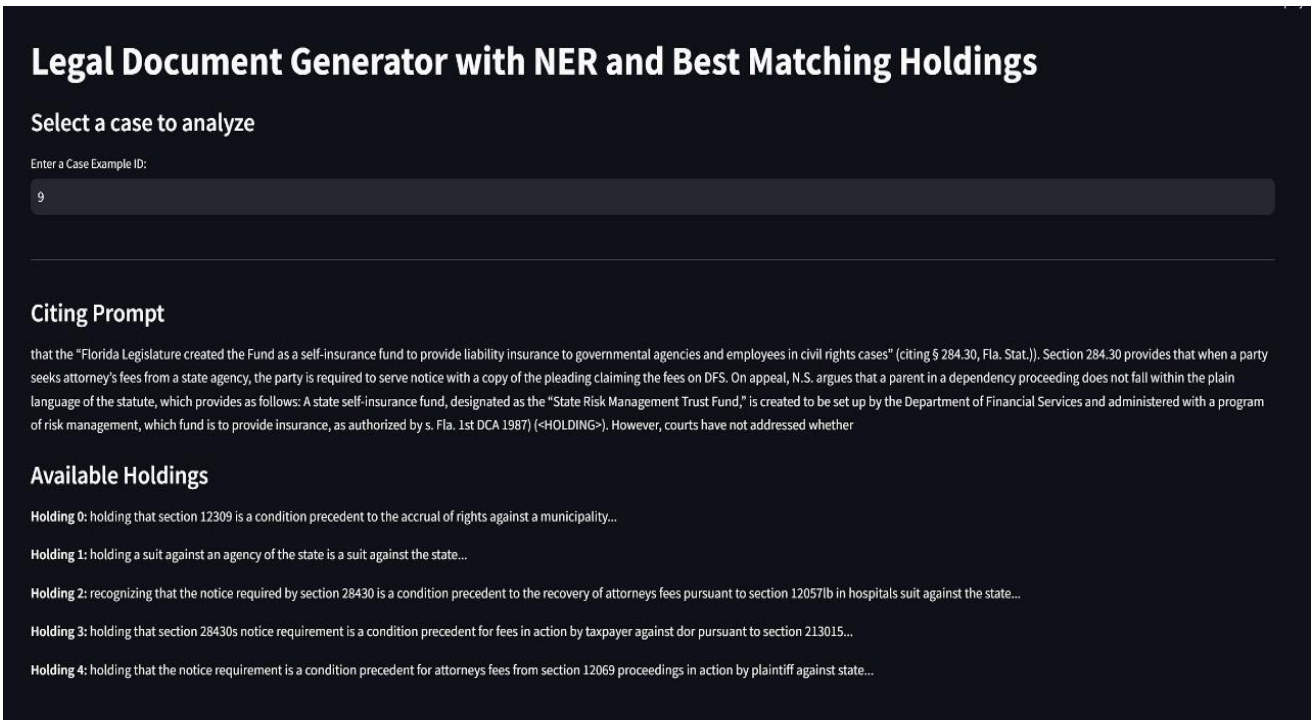
The system begins by acquiring user-entered information or uploaded documents while obtaining necessary data elements including names and dates as well as payment terms and jurisdictional constraints. The data processing phase utilizes spaCy and NLTK to perform tokenization then stops the words and converts text to its base form. The document classification step dictates which specific handling processes will apply to different document types. The text cleanup ensures proper vectorization abilities using TF/IDF or GloVe which allows semantic analysis to proceed in the following operations. The Named Entity Recognition (NER) module detects essential legal entities that include parties and their corresponding details together with pay periods and financial amounts. A ready-to-review document originates when the system applies predefined Jinja2 templates to net new legal entities extracted dynamically from the input text. Users can access a preview mode to view the file and a download function enables them to obtain files in either PDF or DOCX formats. The evaluation module relies on precision and recall for metric measurements during NER operations while incorporating user assessment to improve the system's accuracy and quality of contract.

PERFORMANCE AND RESULTS

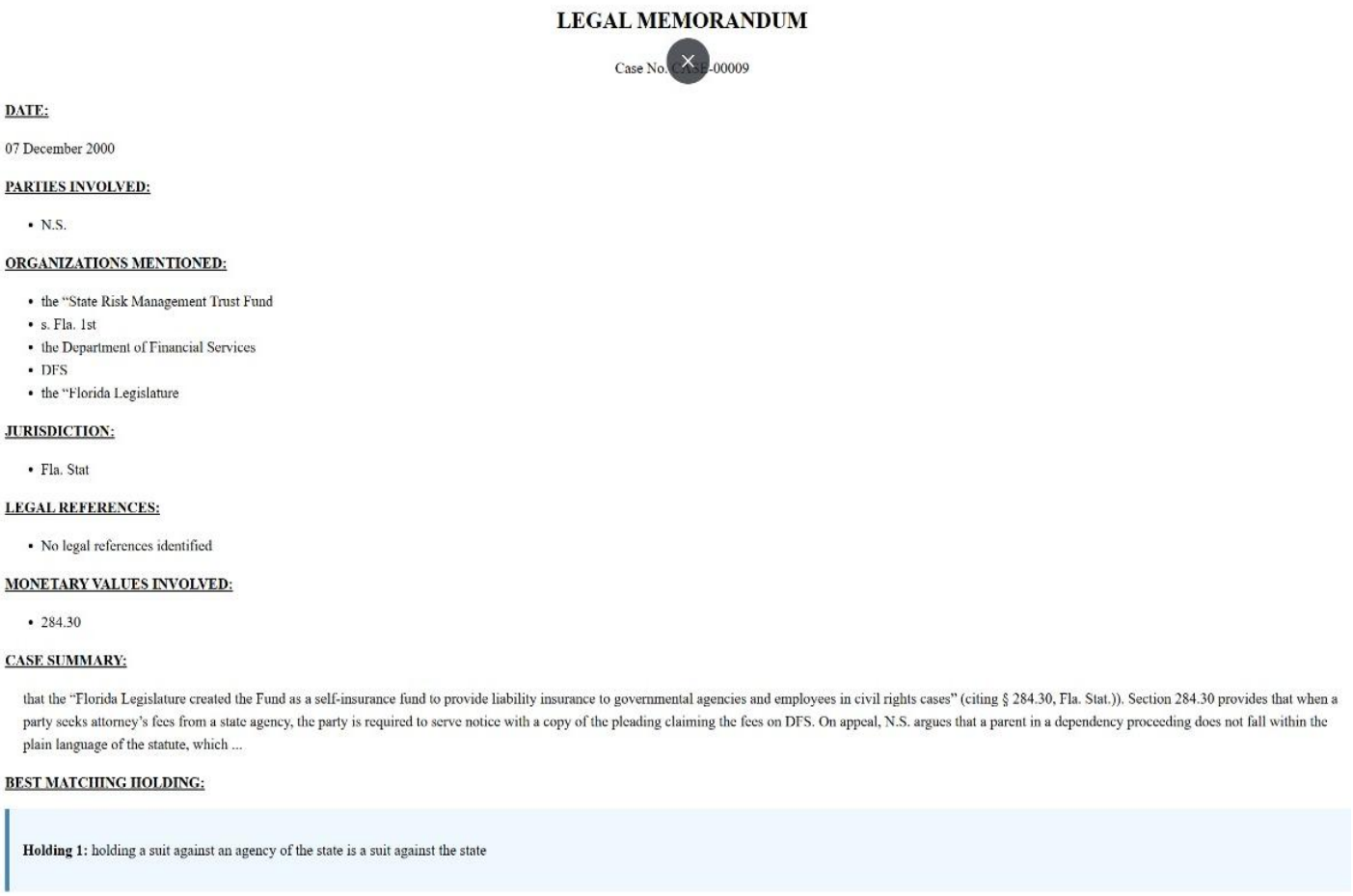
The performance evaluation of AI-based legal contract generation required testing two models which applied TF-IDF vectorization and GloVe word embeddings to process the CaseHOLD dataset. The project analyzed two text feature representations to identify which method provided the best support for classification tasks that lead to clause extraction in contract automation.

Through Streamlit users access a user interface to enter case IDs and examine citing prompts and holdings while automatically creating downloadable formatted legal memoranda for review.

The system composed of Named Entity Recognition (NER) and similarity-based matching models (TF-IDF and GloVe) produces legal memorandums from case requests. An application accepts a case ID input which triggers the retrieval process for displaying both citing prompt content with available holdings. A NER model in the system performs entity extraction to identify parties alongside organizations along with legal jurisdictions and monetary values and references from case documents.



The system utilizes TF-IDF together with GloVe models to determine the most fitting holding by comparing the citing prompt against available holdings. A structured legal memorandum with extracted entities and the strongest matching holding emerges from this system.



LEGAL MEMORANDUM

Case No. CASE-00009

DATE:

07 December 2000

PARTIES INVOLVED:

- N.S.

ORGANIZATIONS MENTIONED:

- the "State Risk Management Trust Fund
- s. Fla. 1st
- the Department of Financial Services
- DFS
- the "Florida Legislature

JURISDICTION:

- Fla. Stat

LEGAL REFERENCES:

- No legal references identified

MONETARY VALUES INVOLVED:

- 284.30

CASE SUMMARY:

that the "Florida Legislature created the Fund as a self-insurance fund to provide liability insurance to governmental agencies and employees in civil rights cases" (citing § 284.30, Fla. Stat.)). Section 284.30 provides that when a party seeks attorney's fees from a state agency, the party is required to serve notice with a copy of the pleading claiming the fees on DFS. On appeal, N.S. argues that a parent in a dependency proceeding does not fall within the plain language of the statute, which ...

BEST MATCHING HOLDING:

Holding 1: holding a suit against an agency of the state is a suit against the state

TF-IDF outperformed GloVe by posting performance results of 51.8% accuracy and 51.76% weighted F1-score while GloVe achieved 37.2% accuracy and 37.33% weighted F1-score. The classification reports also revealed TF-IDF yielded higher levels of precision and recall across a range of holding classes while GloVe reported lower recall along with precision levels in several classes.

TIER-1 Results

TF-IDF MODEL

Class (Holding)	Precision	Recall	F1-Score	Support
Holding 0	0.45	0.56	0.50	85
Holding 1	0.51	0.41	0.46	97
Holding 2	0.55	0.61	0.58	96
Holding 3	0.46	0.52	0.49	102
Holding 4	0.63	0.49	0.55	120

The overall accuracy score reached 51.8% along with F1-score of 51.76% from the TF-IDF model.
Test results indicated the best outcomes occurred when applying TF-IDF features to analyze **Holding 2 (F1 = 0.58)** and **Holding 4 (F1 = 0.55)** samples. Lower recall for **Holding 1**(**0.41)** indicates some difficulties in identifying this class. The classes maintained similar levels of precision and recall which indicated no signs of extreme class imbalance.

GloVe MODEL

Class (Holding)	Precision	Recall	F1-Score	Support
Holding 0	0.26	0.29	0.28	85
Holding 1	0.36	0.37	0.36	97
Holding 2	0.43	0.45	0.44	96
Holding 3	0.36	0.38	0.37	102
Holding 4	0.44	0.36	0.40	120

The GloVe-based model achieved a lower accuracy of 37.2% and a weighted F1-score of 37.33%**.**
The highest F1-Score was for **Holding 2 (0.44)** and **Holding 4 (0.40)**. Very low performance for **Holding 0 (F1 = 0.28)** indicates poor separability for this class using GloVe. Both precision and recall are notably lower than the TF-IDF model across all classes.

TIER-2 Results

Model	Accuracy (%)	F1-Score (%)
TF-IDF (Used)	51.8	51.76
GloVe (Used)	37.2	37.33
BERT	70.8	70.8
RoBERTa	71.4	71.4
CaseLaw-BERT	75.4	75.4

In contrast to TF-IDF (51.8% accuracy) and GloVe (37.2%), the transformer-based models BERT (70.8%), RoBERTa (71.4%), and CaseLaw-BERT (75.4%) have a 19%–24% improvement in accuracy when compared to TF-IDF and 33%–38% improvement when compared to GloVe. The best accuracy and F1-score are achieved by the domain pretrained

CaseLaw-BERT (75.4%). This verifies that domain-specific pretraining results in improved performance on legal NLP tasks as opposed to general-purpose models such as BERT or RoBERTa.

While not domain specific, pretraining does not aid, dynamic masking and longer training even enhance performance slightly (RoBERTa performs 0.6% better than BERT accuracy/F1). For TF-IDF, this project even outperforms GloVe (37.2%) with 51.8%, but both of these are a clear way behind the transformer approaches. Thus, this indicates TF-IDF still performs a good baseline for similarity but does not have the in-depth semantic knowledge contained in context embeddings.

TIER-3 Results

COMPARITIVE ANALYSIS

Metric	TF-IDF	GloVe	Difference
Accuracy	0.5180	0.3720	+0.1460
Precision (Weighted)	0.5268	0.3773	+0.1495
Recall (Weighted)	0.5180	0.3720	+0.1460
F1-Score (Weighted)	0.5176	0.3733	+0.1443

The TF-IDF approach delivers a steady boost of 14-15% throughout all measures of evaluation. Pretrained GloVe embeddings without fine-tuning demonstrate inferior performance in legal domain classification than the TF-IDF features that come from the dataset corpus. The TF-IDF method shows better performance in identifying legal specific terminology and context-related keywords that matters for legal document generation.

When it comes to legal case holdings TF-IDF represents term importance better than GloVe embeddings because it adapts to unique legal vocabulary. Some holdings such as Holding 4 and Holding 2 provided easier predictability results compared to others. A potential issue with class distribution and class feature overlap seems to exist according to these evaluation results. A higher precision and recall rate makes the TF-IDF model more dependable as the first step of NLP processing for legal document generation.

CONCLUSION AND FUTURE SCOPE

Experiment data showed that TF-IDF surpassed GloVe as it reached 51.8% accuracy accompanied by a weighted F1-score of 51.76%. Output data from the GloVe model revealed a 37.2% accuracy level along with a weighted F1-score of 37.33%. According to research findings TF-IDF demonstrated better performance than GloVe embeddings since it proved superior in recognizing legal domain terminology although it uses a basic statistical approach. Statistical vectorization techniques which use TF-IDF have proven significantly suitable for legal NLP domains whenever large pretrained models or extensive computational capabilities are available. The project achieved its target by developing a system that pulled essential legal entities from the text and produced dynamic contract templates by filling templates automatically.

The project proves the possibility of building an affordable, interpretable AI system which automates legal contract creation to decrease preparation time and eliminate drafting errors while expanding support for small enterprises and startup operations and individual users. The Samvidhan AI current deployment brings practical interpretability to automated legal contract writing yet its performance can still be substantially increased. Samvidhan AI needs enhancement by incorporating specialized language models like LegalBERT or CaseLawBERT in its development framework. Legal domain-specific pretraining on legal corpora enables these models to improve their semantic understanding of legal terminology plus structure which enhances their accuracy and contextual applicability in contract drafting.

The application of transformer-based architectures including BERT, RoBERTa and GPT-based models would provide better contextual embedding of legal clauses than traditional static embeddings like TF-IDF or GloVe. Transformer-based models provide better capabilities for dealing with legal language complexities during the process of both contract interpretation and clause generation for complex documents. Introducing more diverse contracts and more court decisions and jurisdictional documents into the dataset will make the model effective in wider legal frameworks and market segments. The system's suitability expands when presented with different ranges of contract types and legal requirements through an enlarged dataset that allows it to function internationally and across multiple jurisdictions.

These upcoming improvements will transform SamvidhanAI into an advanced, intelligent user-friendly solution for the developing requirements of legal contract generation.