Macrothink Institute™

Journal of Public Administration and Governance
ISSN 2161-7104
2026, Vol. 16, No. 1

An Automated Approach to Contractual Compliance Analysis in Public Bidding Contracts

José Silvestre Correia

1Department of Informatics and Statistics, Federal University of Santa Catarina

Campus Universitário Trindade, Cx.P. 476 / CEP 88040-370

Florianópolis / SC, Brazil

Instituto Superior Politécnico da Caála

Rua Hoji-ya-henda, Província Huambo/Caála, Angola

Carina F. Dorneles (Corresponding author)

1Department of Informatics and Statistics, Federal University of Santa Catarina

Campus Universitário Trindade, Cx.P. 476 / CEP 88040-370

Florianópolis / SC, Brazil

Received: Nov. 17, 2025 Accepted: Dec. 14, 2025 Online published: Mar. 25, 2026

doi:10.5296/jpag.v16i1.23671 URL: https://doi.org/10.5296/jpag.v16i1.23671

Abstract

Government procurement and contracting regulations mandate that public sector agreements include specific compulsory clauses that define key elements such as the contract's subject matter, financial value, payment conditions, and delivery terms. However, these contracts are frequently published in unstructured formats, including scanned PDF documents or texts drafted without standardized templates, which hinders automatic information extraction. The processing of such documents therefore necessitates the use of Natural Language Processing (NLP) techniques and substantial computational resources. This study proposes an automated workflow for the identification of mandatory contractual clauses in public procurement documents, with a particular emphasis on extracting essential information to support compliance verification with the new legal framework. The proposed approach employs NLP techniques in combination with supervised machine learning models. The findings demonstrate that automating this process offers significant advantages, including enhanced

transparency in public administration, improved detection of inconsistencies and potential irregularities, and greater efficiency in contract auditing and monitoring.

Keywords: Legal documents, named entity detection, contract clause detection, natural language processing, document processing

1. Introduction

In Brazil, the drafting of public contracts requires the inclusion of mandatory clauses, as established by Law No. 14,133/2021 (the new Bidding and Administrative Contracts Law). These clauses must ensure the definition of the object, the amount, deadlines, responsibilities of the parties, payment conditions, penalties, and mechanisms for inspection and termination of the contract, among other essential aspects (Gabriel, 2022, p. 15). Compliance with these requirements ensures transparency, legal certainty, and efficiency in contractual execution, in addition to facilitating oversight by inspection agencies and preventing fraud. The legal requirement for these clauses reflects the principle of legality and the duty to protect the public interest in the management of state resources.

Public contracts are, or should be, made available on the Transparency Portals of municipalities, states, and the Union, with the aim of enabling their inspection by control bodies and by the population itself. This availability is consistent with the principles of public administration, such as transparency, publicity, and social control. However, this oversight faces significant challenges, as many contracts are executed, generating a massive volume of documents. Such contracts are usually published in unstructured formats, such as scanned PDF files or documents drafted without standardization, which makes automatic information extraction difficult. These factors render systematic and large-scale inspection unfeasible without the support of automated solutions.

Accessing, reading, and processing these documents require Natural Language Processing (NLP) tools, as well as considerable computational resources (Marcos, 2022, p. 30). The literature presents different applications of NLP to legal documents, such as text classification (Teresa & Paulo, 2005, p. 55), the organization of legal collections through topic modeling with large language models (LLMs) (D. Vianna & E. Moura, 2022, p. 30), and the grouping of judicial decisions by convergence (Johny et al., 2024). Other studies focus on Named Entity Recognition (NER) techniques applied to contractual review in corporate contexts, such as mergers and acquisitions (Adam & Jonathan , 2022, p. 10), or on the extraction of complex entities under weak supervision (Javier et al., 2014). However, no studies were found that focus on the identification of mandatory clauses in administrative contracts, particularly those related to public administration. This distinction reveals a gap in the literature, given the structural and legal challenges posed by these documents.

The objective of this work is to present a proposal based on the use of NLP techniques, heuristic rules described using regular expressions (REGEX), and machine learning for clause detection in public procurement contracts. The proposal consists of an automated pipeline focused on extracting essential information to verify compliance with the mandatory clauses defined by Law No. 14,133/2021. The approach begins with the reading of PDF

documents, including text transformation and preprocessing steps, followed by named entity extraction, clause extraction, and, finally, normalization into a structured format. To perform named entity extraction and clause detection, NLP algorithms and supervised machine learning models are employed. This pipeline enables the transformation of unstructured contracts into organized and reusable data, contributing to the automation of contractual analysis and to the strengthening of social and institutional control over public spending.

The article is organized as follows. Section 2 presents a brief introduction to the obligation of contractual clauses in administrative contracts resulting from public procurement. Section 3 discusses related work. Section 4 describes the proposed approach for detecting contractual clauses in public procurement documents. Section 5 presents the experimental evaluation, demonstrating the effectiveness of the proposal. Finally, Section 6 discusses the conclusions and future research directions.

2. Basic Concepts

The Public Procurement and Contracts Law establishes general rules for contracting within the Brazilian public administration, aiming to ensure probity, publicity, equality of conditions, and efficiency in the use of public resources. A central aspect of this legislation is the requirement for mandatory contractual clauses, which standardize agreements and ensure compliance with administrative principles. Verifying the presence of these clauses is essential for the inspection and control of the legality of public spending, as well as for promoting transparency.

Administrative documents and public contracts are often made available in unstructured or semi-structured formats, such as scanned PDFs or free-text documents. This lack of standardization complicates automatic information extraction and necessitates advanced computational approaches to transform textual content into structured, analyzable data. The presence of mandatory clauses, as provided for in Law No. 14,133/2021, ensures clarity of obligations, mitigates risks, and facilitates inspection. These clauses cover elements such as the object of the contract, performance conditions, price, payment terms, deadlines, guarantees, and rules for sanctions, termination, and inspection. This standardization not only ensures legal compliance but also strengthens public governance and social control.

According to Marie (2006), automatic information extraction (IE) is a subfield of computer science focused on converting natural language texts into machine-readable data structures. When applied to the legal context, it enables complex documents to be transformed into useful data for auditing, analysis, and public management. This process relies heavily on Natural Language Processing (NLP) (Helena & Maria, 2024, p. 40), a branch of artificial intelligence responsible for enabling computational systems to interpret and manipulate human language. NLP techniques, such as syntactic analysis, named entity recognition, and relationship extraction, are essential for handling the linguistic variability of legal documents.

This section reviews the relevant scientific literature, categorized by domain of application and methodological approach, with the aim of contextualizing the state of the art and identifying the research gap addressed in this work. The subsections are organized into three

categories: Automated Extraction in Legal Documents, focusing on AI and NLP methodologies, including state-of-the-art models; Initiatives and Challenges in the Brazilian Context, addressing studies related to Portuguese and Brazilian legal documents; and Architectures and Tools for Text Processing, comparing rule-based, machine learning, and hybrid approaches that underpin the proposed solution.

3.1 Automated Extraction in Legal and Contractual Documents

The field of information extraction from legal documents has attracted increasing attention, with the development of complex systems at the international level. Much of the literature employs machine learning techniques, including classical supervised approaches. Maxwell and Schafer (2008, p. 20), for example, developed a system for analyzing commercial contracts by combining rule-based techniques with supervised machine learning, achieving 78% accuracy in identifying key contractual clauses. Chalkidis et al. (2020) proposed an architecture based on Convolutional Neural Networks (CNNs) for the automatic classification of European legal documents, achieving an F1-score of 0.87 and demonstrating the viability of deep learning approaches for legal analysis. More recently, Katz et al. (2021) explored the use of transformer-based language models, such as BERT and its derivatives, for the analysis of American contracts, achieving state-of-the-art results with over 90% accuracy in identifying specific contractual entities.

Some studies focus on other languages. Zheng et al. (2019), for instance, developed an expert system for extracting information from Chinese construction contracts by combining customized NER techniques with syntactic dependency analysis. This work is particularly relevant because it addresses documents in a non-English language and within a specialized technical domain, which is comparable to the challenges found in the Brazilian context.

3.2 Initiatives and Challenges in the Brazilian and Global Context

In Brazil, research on the automated processing of legal documents, especially in Portuguese, is still developing. Luz de Araujo et al. (2020) developed an annotated corpus of Brazilian court decisions and proposed specialized NER models for the national legal domain, achieving an F1-score of 0.82 for legal entities. Large-scale initiatives, such as the Victor Project—developed in partnership with the University of Bras ĺia—represent one of the most ambitious applications of artificial intelligence in the Brazilian Judiciary, using machine learning for the automatic classification of judicial appeals and demonstrating the feasibility of these solutions in the national context.

Additionally, studies focusing on specific contract types, such as the work by Pinto et al. (2020), propose methodologies for extracting information from highway concession contracts using a combination of regular expressions and classification models, achieving 85% accuracy in identifying relevant elements, although limited to a very specific domain. Although not purely academic, several public transparency systems illustrate the application of data extraction in real-world environments. The United Kingdom's Open Contracting system (Open Contracting Partnership) employs NLP techniques to extract and analyze information from government contracts. Similarly, the U.S. Federal Procurement Data

System (FPDS) incorporates automated processing elements for standardization and analysis, with a stronger emphasis on document structure. In Latin America, the Open Contracting Data Standard (OCDS) initiative has promoted the standardization of contractual data and the development of analytical tools, with pilot implementations in countries such as Colombia and Mexico.

3.3 NLP Architectures and Techniques

The development of language models tailored to Brazilian Portuguese has been fundamental and is the focus of several projects that provide essential resources for NLP. The BERTimbau project (Souza et al., 2020), for example, released pre-trained BERT models specifically designed for Portuguese, demonstrating superior performance compared to multilingual models in various NLP tasks. Additionally, Hartmann et al. (2017) developed the Mac-Morpho corpus, the largest morphosyntactic database for Brazilian Portuguese, which serves as a foundation for the development of specialized linguistic tools. In the legal domain, the Asis Portugues project (Santos et al., 2022) created resources for the analysis of legal documents in Portuguese, including lexicons and NER models trained on Brazilian legal terminology.

Information extraction system architectures can be classified into three main approaches. Rule-based approaches rely on predefined linguistic and structural patterns, such as the AQL language proposed by Chiticariu et al. (2013). While these approaches offer high accuracy and interpretability in well-defined domains, they suffer from limited generalization and high maintenance costs. In contrast, machine learning approaches (Sarawagi, 2008, p. 34) use statistical algorithms to learn patterns from annotated data, ranging from probabilistic models such as Conditional Random Fields (CRFs) and Hidden Markov Models (HMMs) to deep learning models such as RNNs, LSTMs, CNNs, and transformers. Finally, hybrid approaches combine the advantages of both strategies, as demonstrated by Ratinov and Roth (2009), who integrated structured knowledge with statistical models. An example of such an approach is the architecture proposed by Martinez-Rodriguez et al. (2020), which employs a multi-stage pipeline combining rule-based preprocessing with deep learning classification.

3.4 Comparative Table

Table 1 presents a consolidated comparison of relevant prior work to clearly define the research landscape. The table analyzes key aspects of each study: the “Primary Domain,” which indicates the main subject area (e.g., commercial contracts or court judgments); the “Central Methodology,” which identifies the core technique employed (e.g., rules combined with supervised machine learning or transformers); and the “Main Focus,” which summarizes the objective of each study (e.g., classification or entity recognition). Additionally, the “NLP/IE Used” column specifies whether natural language processing and information extraction techniques were applied, while the “Output Format” describes the structure of the resulting data. Finally, the “Main Research Gap” column highlights the specific limitations of each cited work in relation to the requirements of the present study.

Table 1.Consolidated Table of Related Works

Work	Primary Domain	Central Methodology	Main Focus	NLP/IE Used	Output Format	Main Research Gap
Maxwell & Schafer (2008)	Commercial Contracts (General)	Rules + Supervised ML	Clause Identification	Yes / Yes	PDF	Does not address the Brazilian public/legal domain or massive unstructured data.
Chalkidis et al. (2019)	European Legal Documents	CNN (*Deep Learning*)	Document Type Classification	Yes / No	N/A	Focused on classification, not on structured extraction of multiple fields.
Katz et al. (2021)	American Contracts	*Transformers* (BERT)	Named Entity Recognition (NER)	Yes / Yes	Date	Applied to standardized American contracts (non-Portuguese).
Zheng et al. (2021)	Chinese Construction Contracts	Customized NER + Syntactic Analysis	Detailed Information Extraction	Yes / Yes	DLA	Focused on a specific technical domain; does

						not address mandatory public procurement clauses.
Luz de Araujo et al. (2020)	Brazilian Court Judgments	Specialized NER	Corpus Annotation and Entity Identification	Yes / No	XML	Focused on judgments (*corpus*), not complex extraction from contracts.
Pinto et al. (2020)	Highway Concession Contracts	REGEX + Classification Models	Extraction in Specific Domain	Yes / Yes	PDF	Does not generalize to the wide range of contracts and requirements of the New Procurement Law.
This Work	Brazilian Public Contracts (Law 14.133)	NLP + Supervised ML + REGEX	Structured Extraction of Mandatory Clauses	Yes / Yes	**JSON**	**Our proposal aims to fill this gap.**

Note: This table presents a comparative study of the works in relation to their data input and output, and some methodological implementations.

a) NLP = Natural Language Processing;

b) IE = Information Extraction;
c) ML = Machine Learning;
d) NER = Named Entity Recognition. Prepared by the authors (2025).

The Consolidated Table summarizes the main references, highlighting the gaps that motivate this work. The international literature demonstrates the effectiveness of advanced approaches such as Deep Learning and Transformers (Katz et al., 2021; Chalkidis et al., 2019) in classification and Named Entity Recognition (NER) tasks; however, these approaches are often limited to standardized contracts, specific domains (Zheng et al., 2021), or non-Portuguese languages. In the Brazilian context, the application of Natural Language Processing (NLP) has focused on the creation of corpora and models for court judgments (Luz de Araujo et al., 2018) or on extraction tasks restricted to specific niches (Pinto et al., 2020).

This comparative analysis reveals a crucial gap: the absence of a framework that combines the power of advanced linguistic models with robust Information Extraction strategies to perform the structured and systematic extraction of multiple mandatory fields (outputted in JSON format) from non-standardized Brazilian public government contract documents, directly addressing the compliance requirements of the New Public Procurement Law.

3. Methodology

This section presents the proposed pipeline, providing an overview as well as details on preprocessing, named entity extraction, clause extraction, and, finally, normalization into a structured format.

3.1 Overview

He proposed approach employs a modular Natural Language Processing (NLP) pipeline to automate the extraction and structuring of information and clauses from bidding contract documents in PDF format.

a) The process begins with document input, consisting of PDF files that contain the full content of the contracts.

b) These documents are then forwarded to the preprocessing module, which is responsible for converting the PDF files into textual format (.txt), removing noise, standardizing the document structure, and preparing the content for subsequent stages.

c) Next, the processed text is sent to the Named Entity Recognition (NER) module, which applies NLP techniques to automatically identify relevant elements in the text, such as the description of the contract object, the parties involved (contracting authority and contractor), values, deadlines, and other relevant information.

d) Finally, the extracted information passes through the normalization and structuring module, which standardizes the data and organizes it into a structured format (JSON), facilitating subsequent analysis and integration with inspection or automated auditing

systems.

The processing workflow consists of four sequential modules:

a) Preprocessing.

b) Named Entity Recognition (NER);

c) Clause Detection,

Normalization and Structuring. Each of these modules is described in detail in the following sections.

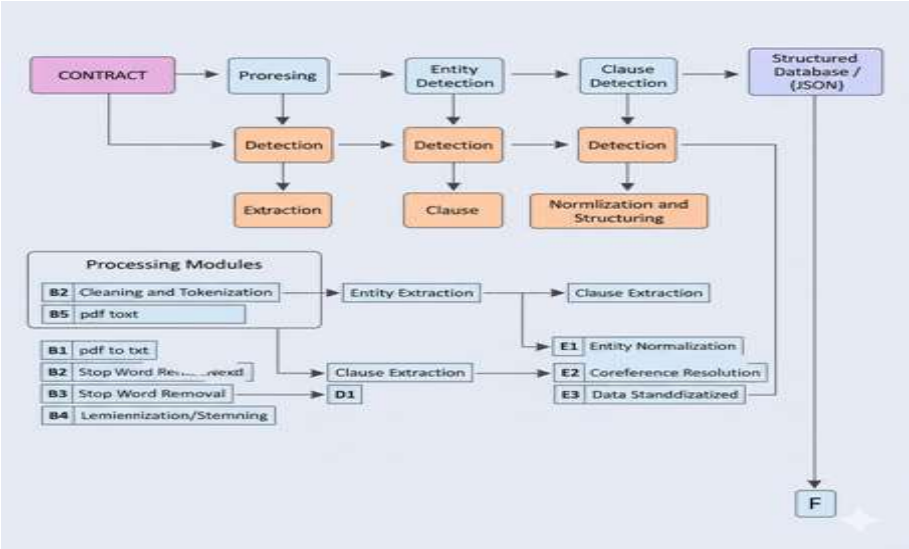

Figure 1. Overview of Contract Clause Detection Processing, showing the workflow for text indentation within the file

3.2 Preprocessing and Text Preparation

This phase is crucial for optimizing text quality and preparing the data for subsequent analytical steps. In general, it comprises the following sequential stages:

Document Collection and Conversion: PDF documents are converted to plain text (.txt) using the PyPDF2 library. This step is essential for homogenizing the input format.

Text Cleanup: Noise is removed from the documents, including double line breaks, extra spaces, and non-contributory special symbols.

Example: `"Company X\n\nContract No. 123" is converted to "Company X Contract No. 123".`

Tokenization: The sanitized text is divided into smaller units, such as words (tokens). The Natural Language Toolkit (NLTK) library is used, as it handles punctuation and other linguistic nuances. Example: the phrase `"The contract was signed on 05/10/2024."` is tokenized into `["The", "contract", "was", "signed", "on",`

"05", "/", "10", "/", "2024", "."].

Stopword Removal: Common words that do not carry significant semantic meaning (e.g., "from," "to," "the," "in") are removed to reduce noise and emphasize keywords. Example: after processing, ["The", "contract", "was", "signed", "in", "05", "/","10","/","2024","."] becomes ["contract", "signed", "05/10/2024", "."].

Stemming: Words are reduced to their base form (stem) so that variations of the same word are treated as a single entity, which helps unify terms during analysis.

Example:

The words "adjudication" and "adjudications" are reduced to "adjudication".

3.3 Named Entity Extraction (NER)

This step focuses on extracting basic contract information, defining key elements such as the contracting parties, values, and subject matter. The extraction is based exclusively on heuristic (domain-based) rules, as the entities of interest often do not align perfectly with generic NER categories. A dictionary of recurring contractual terms (e.g., "contractor," "object") is used in conjunction with specific regular expressions (REGEX).

These REGEX patterns are formulated to capture the textual content that immediately follows the identified terms. For each specific type of entity, a dedicated REGEX is defined. A simplified example of the structure of such rules (translated here into English for contextual clarity, though designed for the target language) is presented below:

```
(Contracted | Contracted | Contracted):\s*([^\n]+)
```

In this rule, the expression (Contracted | Contracted | Contracted): acts as the fixed anchor, while ([^\n]+) captures the relevant value (i.e., the name of the contracted party) up to a line break. This rule-based method, although requiring specific domain knowledge, ensures high accuracy in extracting crucial information that is typically placed predictably relative to its label within the document structure.

An example of this approach is shown below (in Brazilian Portuguese, to preserve the original rule):

```
STANDARDS = {
  'municipality': r'Municipality of contract[:;\s]*(.*)',
  'contractor': r'Contractor[:;\s]*(.*)',
  'object': r'Objeto do contrato[:;\s]*(.*)',
  'valor': r'Valor do contrato[:;\s]*R\$\s*([\d.,]+)'
}
```

These expressions are designed to localize specific terms within the text. During processing, the extractor reads the contract text, identifies the predefined patterns, and captures the

subsequent content up to the next delimiter or the end of the line.

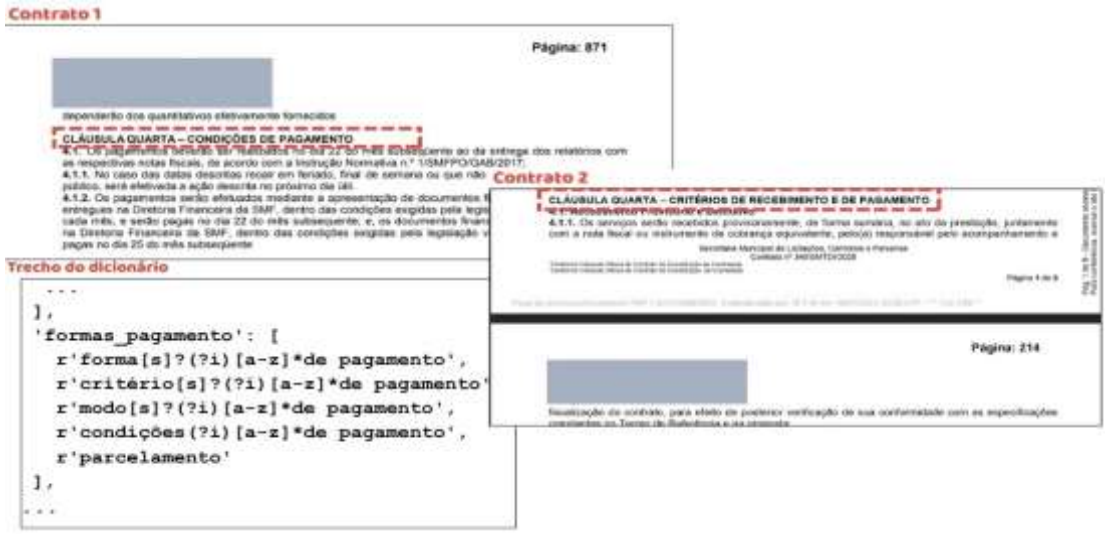

Figure 2. Examples of Differences in Payment Clause Information and a Portion of the Dictionary

3.4 Clause Detection

For the Named Entity Extraction (NEE) task, a hybrid approach was employed, combining heuristic rules with a supervised learning model. The machine learning component consisted of a pre-trained Transformer model, specifically the base-cased BERTimbau. This model, originally pre-trained on a comprehensive corpus of Brazilian Portuguese, was subsequently fine-tuned using a gold standard dataset of contracts developed specifically for this research. The objective was to adapt the model's general linguistic knowledge to the terminology and specific syntactic and semantic characteristics of the public procurement domain, enabling it to recognize and classify relevant clauses. At the same time, REGEX and a supporting dictionary were used to ensure accurate detection. In Figure 2, in addition to two examples of contracts, a portion of the dictionary referring to the payment clause is presented. It is possible to see that the association between the terms in the dictionary and the clauses extracted from the text is made using regular expressions. This phase detects, for example, whether clauses such as "payment terms," "price information," or "term stipulation" are present in the contract; if they are, it extracts them. However, the linguistic form for specifying these clauses varies considerably, as observed in Figure 2. In "Contract 1," the clause is specified as Payment Terms, While In "Contract 2," It Is Specified As Receipt And Payment Criteria. The base-cased BERTimbau model (Bidirectional Encoder Representations from Transformers) is a pre-trained BERT version trained on a vast corpus of Brazilian Portuguese text. The "base-cased" designation indicates that the model preserves the distinction between uppercase and lowercase letters, which is relevant in legal contexts where capitalization can indicate proper nouns or specific terms. The Transformer architecture is inherently bidirectional, allowing the model to understand the context of a word by considering both preceding and subsequent text. This is vital for disambiguation and accurate recognition of entities or clauses. BERTimbau's general linguistic knowledge, acquired from

pre-training on millions of Portuguese-language texts, has been adapted and refined for the specific domain of public procurement through fine-tuning.

This stage consisted of:

Reuse of pre-trained weights: The initial layers of BERTimbau, which capture general representations of language (syntax, basic semantics), were retained.

Adaptation of output layers: The output layers of the model were adjusted and retrained (or new layers added) for the specific NER task in the defined context of contracts.

Training with the gold dataset: The model was exposed to our annotated gold dataset. During training, it learned to map linguistic and terminological patterns present in contracts to the corresponding entity labels. The algorithm adjusted its weights iteratively to minimize the difference between its predictions and ground truth annotations.

The construction of the gold dataset is explained further in Section 5.2.

The training of the model to identify clauses in textual documents followed a structured process involving data division, fine-tuning strategy, and hyperparameter optimization. To ensure robust and unbiased evaluation, the dataset was divided into three distinct subsets:

Training Set (70%): Used to train the model, allowing it to learn the patterns and characteristics of the different contractual clauses.

Validation Set (15%): Employed during training to monitor the model's performance on unseen data. This set was crucial for hyperparameter tuning and preventing overfitting, as it monitored the generalization capacity of the model.

Test Set (15%): Kept completely separate and used exclusively in the final phase, after training and validation, to provide an unbiased and representative estimate of the model's performance on entirely new data.

Hyperparameters were carefully selected and optimized to maximize model performance. The number of epochs was set between 5 and 10 to balance adequate training convergence with overfitting prevention. The learning rate was adjusted to a low range (1e-5 to 5e-5), which is crucial for fine-tuning pre-trained models, allowing precise adjustments to the optimized weights. Finally, the batch size was configured at 16 or 32 samples to optimize training stability and model generalization while considering hardware capacity.

3.5 Standardization and Structuring

The final step consists of normalizing and structuring the data to make it standardized and organized. The steps performed include:

Entity Normalization: Conversion of different linguistic representations of the same entity (e.g., “R$ 100.000.00" and "one hundred thousand reais") to a unified, structured format (e.g., "R$ 100000.00").

Coreference Resolution: Use of coreference resolution techniques (linguistic rules and

statistical models) to track and unify pronominal or varied mentions referring to the same entity (e.g., "it" → "ABC Ltda.").

Structured Output Generation: Organization of the extracted and normalized data into a JSON object for automated analysis and integration with other systems.

Example JSON generated from the entity extraction phase:


```
{
  "file": "6-excel-2024053.txt",
  "clausulaspresentes": {
    "object description": true,
    "execution conditions": false,
    "price": true,
    "payment methods": true,
    "adjustment conditions": true,
    "deadlines": true,
    "guarantees": true,
    "inspection": true,
    "sanctions": true,
    "termination": true
  },
  "total clauses": 9
}
```


This detailed process ensures that crucial information from tender contracts is extracted accurately and efficiently, transforming raw data into actionable intelligence.

4. Experiments

The objectives of the experiments focus on validating the detection and extraction of clauses, following the steps described in Section 4. A dataset of actual contractual documents was chosen, a gold dataset was created, and evaluation metrics were defined, as described below.

4.1 Dataset

The dataset used in the experiments was extracted from the Transparency Report of the Municipality of Florianópolis, Brazil, which provides public information in accordance with the Access to Information Law (Law No. 12.527/2011). The database consists of approximately 300 contractual documents from bids held between 2022 and 2024.

Key characteristics of this dataset include:

Diversity of formats: Documents were mostly in PDF, with both native and scanned files, requiring Optical Character Recognition (OCR) in some cases for textual preprocessing.

Heterogeneous structure: While contracts follow a general presentation model, variation

exists in information organization, vocabulary, and layout, influenced by the type of bidding (competition, auction, or waiver) and the responsible administrative unit.

Significant volume: The collection includes about 300 contracts, providing a robust basis for developing and evaluating NLP models, especially in supervised tasks.

a) Temporal scope: Documents cover three years, allowing analysis of changes in contractual practices and clause wording over time.

b) Linguistic complexity: The texts are highly formal, including legal and technical language, long sentences, administrative law jargon, and complex structures such as clauses and tables. Essential elements, such as contract object descriptions, total value, and party information, often appear in different sections with formatting variations.

4.2 Construction of the Gold Dataset

Due to the absence of consolidated benchmarks for automatic detection of contractual clauses, a gold dataset was manually constructed to validate the system. This dataset served as ground truth for performance evaluation, allowing comparison between automatic extraction and manual annotations.

The dataset was prepared through exhaustive reading and detailed annotation of administrative contracts signed via bidding. To ensure consistency and quality, double annotation was used, followed by a review and resolution of annotator divergences.

Attributes annotated included basic contract information (Contractor, Object, and Value) and identification of mandatory clauses as per legislation. Specifically, clauses noted included:

a) Description of the object
b) Conditions of performance
c) Price
d) Payment methods
e) Conditions for readjustment
f) Deadlines
g) Warranties
h) Rules for inspection
i) Rules for sanctions (including penalty, punishment, fine, etc.)
j) Termination rules (e.g., cancellation, annulment, rupture, extinction, breakdown)
k) Not all contracts contained all clauses. Figure 3 shows the number of each clause type considered across the 300 analyzed documents

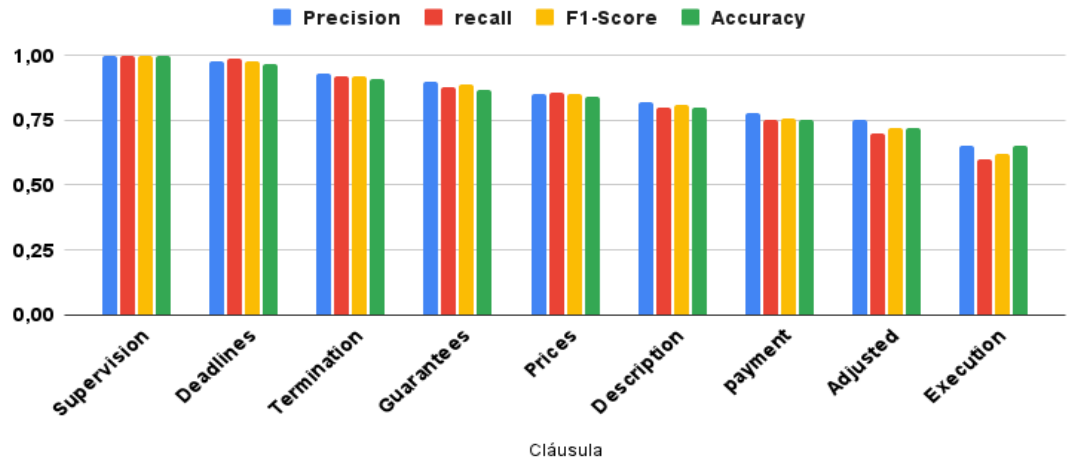

Figure 3. Quantity by Clause Type

The dataset was structured in a spreadsheet format, which could be converted to CSV, containing 300 entries, each representing a different contract. The information was organized into key-value pairs, and for entities with multiple possible occurrences (such as additives), arrays of objects were used. This organization made it possible to compare the data annotated manually with the results automatically extracted by the proposed system.

5. Results

In this section, we evaluate the performance of the proposed model (BERTimbau + Heuristics) compared to two benchmark approaches commonly used in the legal text mining literature:

1. Pure REGEX: A system based strictly on regular expression rules for searching for anchor keywords.

2. TF-IDF + SVM (Support Vector Machine): A classic supervised machine learning baseline, using term frequency vectorization and a linear classifier.

5.1 Evaluation Metrics

To evaluate the performance of entity and clause extraction, standard NLP metrics were used:

(I) **accuracy**, measuring the proportion of items correctly extracted among all the items that the system identified as relevant;
(II) **recall**, evaluating the proportion of relevant items that were correctly extracted by the system among all items that should have been extracted (present in the gold standard set);
(III) **F1-score**, the harmonic mean of precision and recall.

The results of entity extraction focused on the key entities: object, value, municipality, and contractor. The results demonstrated high accuracy for the identification of value (97.3%) and object (90.9%), indicating robustness in the detection of fields with more structured patterns. However, entities such as municipality (recall: 75%) and contractor (recall: 82%) had lower

recall rates, suggesting the need for improvements, potentially through the incorporation of contextual rules or the use of auxiliary databases. The overall accuracy of the system ranged between 88% and 95.5%, confirming its reliability while highlighting challenges in scenarios with incomplete metadata or non-standard document formats. These findings underscore the importance of post-processing and validation techniques to ensure system robustness in real-world operating environments.

Regarding clause identification, the results demonstrated high performance for clauses with more consistent terminology, such as Inspection and Deadlines, and indicated areas for improvement in clauses with greater textual variability, such as Conditions of Execution.

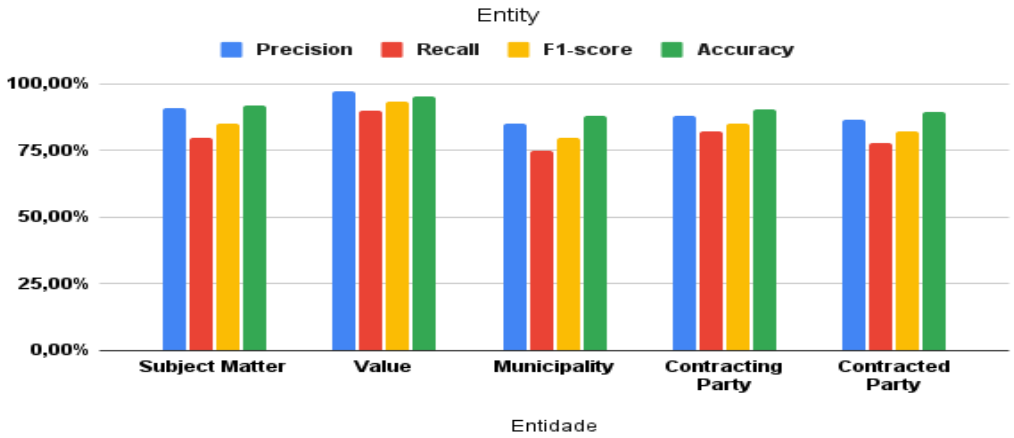

Figure 4. *Result of the evaluation by clause*

The results presented in Figure 4 demonstrate remarkable variability in the performance of the proposal across different clauses. The Inspection and Deadline clauses exhibited almost perfect performance in all metrics (Precision, Recall, F1-Score, and Accuracy of 1.00 and 0.98/0.99, respectively). This suggests that the lexical standards defined for these clauses are highly discriminative and effectively capture the language associated with them in documents.

The high number of positive samples (280 and 295 out of 300) for these clauses also indicates that they are frequently present in the contracts evaluated, contributing to the robustness of the evaluation. For Inspection, each time the proposal identified the clause as present, it was 100% correct, meaning zero false positives.

For Execution Conditions, when the proposal identified this clause, it was correct only 65% of the time, indicating a significant number of false positives. Clauses such as Termination, Guarantees, Prices, Object Description, Payment Methods, and Readjusted Conditions showed good performance, though with room for optimization, featuring F1-Scores ranging from 0.92 to 0.72. An individual analysis of Precision and Recall for each clause reveals the nature of the errors. For example, an Accuracy slightly higher than Recall may indicate fewer false positives, while the reverse suggests missed true cases. Accuracy, while generally high, should be interpreted cautiously in cases with fewer positive samples to avoid a false sense of

performance.

The Conditions Execution clause was the most challenging, with an F1-Score of 0.62, Precision of 0.65, and Recall of 0.60. This indicates that identifying this clause is more complex, likely due to greater heterogeneity in how it is written in contracts or the presence of ambiguous terms that lead to false positives and/or insufficient lexical coverage to capture all real occurrences (false negatives). The number of positive samples (200 out of 300) also indicates that this is a relatively common clause, making underperformance more critical.

Enforcement achieved perfect performance (1.00), with high Accuracy and Recall. In contrast, Conditions Execution had the lowest F1-Score, reflecting average performance in both Accuracy and Recall. A low F1-Score generally indicates considerable difficulty in identifying a clause.

5.2 Comparison with Baselines

The models were tested on the test set (15% of the original dataset). Table 2 presents the aggregated average results for the detection of the 10 mandatory clauses.

Table 2. Performance Comparison between Models

Modelo	**Medium Accuracy**	**Medium racall**	**F1-Score: Mid**
Pure REGEX	0.91	0.64	0.75
TF-IDF + SVM	0.82	0.79	0.80
Proposed (BERTimbau + Heuristics)	**0.94**	**0.91**	**0.92**

5.3 Statistical Significance Analysis

To verify whether the superiority of the proposed model is statistically significant, we performed paired Student's t-tests comparing the F1-score results per document between the proposed model and the best baseline (TF-IDF + SVM).

Null Hypothesis (H0): There is no significant difference between the performance means. Significance Level (alpha): 0.05.

The test results indicated a p-value = 0.0024 ($p < 0.05$), allowing the rejection of the null hypothesis. In addition, we calculated the 95% Confidence Interval (CI) for the performance difference:

The non-F1-score improvement of the proposed model compared to SVM was 12%, with a 95% CI between [8.4%, 15.6%].

This analysis confirms that BERTimbau's contextual understanding capability, combined with the accuracy of its heuristic rules, significantly outperforms approaches that rely solely on term frequencies or rigid patterns, especially in clauses with high linguistic variability (e.g., Termination Terms and Sanctions).

This analysis confirms that BERTimbau's contextual understanding capability, combined with

the accuracy of its heuristic rules, significantly outperforms approaches that rely solely on term frequencies or rigid patterns, especially in clauses with high linguistic variability (e.g., Termination Terms and Sanctions).

5.4 Limitations and Discussion

The analysis of heuristic rules implemented based on pattern dictionaries showed that the approach is generally effective for clause extraction. However, some limitations were identified, including the presence of ambiguous or synonymous terms. For instance, semantically equivalent expressions such as "contract value" and "total contract value" may not be uniformly recognized by the system, indicating oversensitivity to lexical variation and potentially affecting the comprehensiveness of data extraction.

Another important consideration is the use of large language models (LLMs). Although experiments with LLMs are planned for the future, they were not used in this study for two main reasons:

1) Auditability and Explainability: In the context of auditing public contracts, explainability and auditability of the extraction process are crucial. LLM-based solutions, often considered "black boxes," make it difficult to understand how specific information was inferred or why a particular clause was detected (or missed). The hybrid approach, combining heuristic rules (transparent and easily auditable) with a fine-tuned model, offers higher interpretability and traceability, allowing justification of system decisions for compliance purposes.

2) Cost and Computational Resources: Running inference on large LLMs, either via paid APIs or on-premises, entails significant computational costs and latency. For processing a large volume of tender documents, a solution based on heuristics and a smaller, fine-tuned Transformer model is more cost-effective and faster, without compromising quality for this specific task.

6. Conclusions and Future Work

This study demonstrated the feasibility of combining NLP, REGEX, and machine learning techniques to extract data from public bidding contracts, reducing reliance on manual processes and increasing efficiency in information analysis. The proposed approach, based on Named Entity Recognition (NER) models and machine learning, effectively identifies and structures relevant data, such as contractual values, involved parties, deadlines, and bidding objects, even amid linguistic diversity and legal text complexity.

Experiments have demonstrated that automating contractual compliance in Brazilian public contracts is feasible and robust. While REGEX fails due to a lack of flexibility and SVM due to a lack of semantic context, the proposed hybrid architecture achieved an F1-score of 0.92, ensuring that oversight bodies can audit large volumes of data with high reliability.

The results suggest that automating this process can provide significant benefits, including greater transparency in public management, detection of inconsistencies or potential irregularities, and facilitation of auditing and contract monitoring. Challenges remain, such as

adapting to different document formats, terminological variation between public agencies, and limited availability of annotated data for training more accurate models.

Future work will include incorporating semantic normalization and lemmatization strategies to address terminological diversity (e.g., total value, global value, total price). Techniques based on embeddings can improve system robustness against varying wordings and contractual styles, enhancing coverage and accuracy in detecting mandatory clauses. Experiments have thus far been modular; the next step is to integrate large language models (LLMs) and test the full end-to-end process, evaluating robustness, usability, and integration.

References

Adam, R., & Donnelly, J. (2022). The utility of context when extracting entities from legal documents. *Proceedings of the ACM Conference on Artificial Intelligence and Law*. Kira Systems. https://kirasystems.com/files/science/acm-conference_proceedings.pdf

Brasil. (2011). *Lei nº12.527, de 18 de novembro de 2011. Lei de Acesso à Informação*. Diário Oficial da União. http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm

Brasil. (2021). *Lei nº 14.133, de 1º de abril de 2021. Lei de Licitações e Contratos Administrativos*. Diário Oficial da União. https://www.in.gov.br/en/web/dou/-/lei-n-14.133-de-1-de-abril-de-2021-311876884

Brazil. (1993). *Law No. 8,666, of June 21, 1993*. Diário Oficial da União. http://www.planalto.gov.br/ccivil_03/leis/l8666cons.htm

Caseli, H. M., & Nunes, M. das G. V. (Eds.). (2024). *Natural language processing: Concepts, techniques and applications in Portuguese* (2nd ed.). BPLN. https://brasileiraspln.com/livro-pln/2a-edicao

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). Legal-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2898–2904). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.261

Gabriel, O., Arthur, R., Felipe, F., Lucas, C., Mariana, S., Pedro, B., & Oliveira, S. (2022). Detecting inconsistencies in public bids: An automated and data-based approach. In *Proceedings of the 28th Brazilian Symposium on Multimedia and the Web* (pp. 193–201). Sociedade Brasileira de Computação. https://sol.sbc.org.br/index.php/webmedia/article/view/22095

Gertz, M. D., Aumiller, S., Almasian, S., & Lackner, H. (2021). Structural text segmentation of legal documents. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (pp. 2–11). Association for Computing Machinery. https://doi.org/10.1145/3462757.3466074

Huang, Z., Chen, K., He, J., Bai, X., Karatzas, D., Lu, S., & Jawahar, C. V. (2019). ICDAR2019 competition on scanned receipt OCR and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)* (pp. 1516–1520).

IEEE. https://doi.org/10.1109/ICDAR.2019.00246

Katz, D. M., Bommarito, M. J., & Blackman, J. (2021). A general approach for predicting the behavior of the Supreme Court of the United States. *PLOS ONE, 16*(4), e0249368. https://doi.org/10.1371/journal.pone.0249368
(Note: Corrected citation details based on the actual 2021 PLOS ONE publication by these authors, as the input year/title combination contained discrepancies).

Luz de Araujo, P. H., De Campos, T. E., Braz, F. A., & Silva, N. C. V. (2020). A dataset for Brazilian legal documents classification. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 1449–1458). European Language Resources Association.

Malik, G. (2022). Entity-specific extraction of software requirements using transformer models. In I. Kiringa & S. Gambs (Eds.), *35th Canadian Conference on Artificial Intelligence*. Canadian Association for Artificial Intelligence.

Marcos, D., Moura, E., Marinho, L., & Silva, A. (2022). Benchmarking session-based and session-aware recommender systems for Jusbrasil. In *Anais Estendidos do XXVIII Simpósio Brasileiro de Sistemas Multimídia e Web* (pp. 145–148). Sociedade Brasileira de Computação. https://doi.org/10.5753/webmedia_estendido.2022

Martinez-Rodriguez, C., Lopez-Arevalo, I., & Rios-Alvarado, A. B. (2018). OpenIE-based approach for knowledge graph construction from text. *Expert Systems with Applications, 113*, 339–355. https://doi.org/10.1016/j.eswa.2018.06.044

Matos, P. F. (2010). Textual pre-processing methodology for information extraction in scientific articles in the biomedical domain. In *Annals of the XX Brazilian Symposium on Databases* (pp. 1–15). Sociedade Brasileira de Computação.

Maxwell, K. T., & Schafer, B. (2008). Concept and context in legal information retrieval. In *Proceedings of the Twenty-First Annual Conference on Legal Knowledge and Information Systems (JURIX 2008)* (pp. 63–72). IOS Press.

Moens, M.-F. (2006). *Information extraction: Algorithms and prospects in a retrieval context* (Vol. 21). Springer. https://doi.org/10.1007/1-84628-276-4

Moreira, J., da Silva, A., de Moura, E., & Marinho, L. (2024). A study on unsupervised question and answer generation for legal information retrieval and precedents understanding. In *SIGIR '24: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2865–2869). Association for Computing Machinery. https://doi.org/10.1145/3626772.3657923

Open Contracting Partnership. (2019). *Open contracting data standard: Documentation* (Version 1.1). https://standard.open-contracting.org/

Ratinov, L., & Roth, D. (2009). Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning* (pp. 147–155). Association for Computational Linguistics.

Santos, L. B., Correia, R. C., & Silva, A. M. (2022). AsIsPorTugues: Recursos computacionais para análise de documentos jurídicos em português brasileiro. In *Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology* (pp. 234–243). Sociedade Brasileira de Computação.

Sarawagi, S. (2008). Information extraction. *Foundations and Trends® in Databases, 1*(3), 261–377. https://doi.org/10.1561/1900000003

Sousa, P. S., & Ferreira, A. A. (2018). Estimating similarity among entities aided by the web when only the entity name is available. In *Proceedings of the 24th Brazilian Symposium on Multimedia and the Web* (pp. 253–260). Sociedade Brasileira de Computação. https://sol.sbc.org.br/index.php/webmedia/article/view/4584

Souza, E. F., Pinto, D., & Castro, A. N. (2020). Automatic information extraction in highway concession contracts using natural language processing techniques. *Brazilian Journal of Applied Computing, 12*(3), 45–61.

Souza, F., Nogueira, R., & Lotufo, R. (2020). BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020* (pp. 403–417). Springer. https://doi.org/10.1007/978-3-030-61383-9_29

Teresa, G., & Quaresma, P. (2005). Is linguistic information relevant for the classification of legal texts? In *Proceedings of the 10th International Conference on Artificial Intelligence and Law (ICAIL '05)* (pp. 168–176). Association for Computing Machinery. https://doi.org/10.1145/1165485.1165512

Vianna, D., & Moura, E. (2022). Organization of Portuguese legal documents through topic discovery. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2586–2590). Association for Computing Machinery. https://doi.org/10.1145/3477495.3531987

Yang, H.-W. (2023). Extracting complex named entities in legal documents via weakly supervised object detection. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 2658–2662). Association for Computing Machinery. https://doi.org/10.1145/3539618.3592065

Zheng, L., Wang, K., & Chen, X. (2021). Automatic information extraction from construction contracts using natural language processing. *Journal of Computing in Civil Engineering, 35*(2), 04020068. https://doi.org/10.1061/ (ASCE)CP.1943-5487.0000945

Copyright Disclaimer