Tender Documents Information Extraction

**Erick Correia Silva, Ivo Paixão de Medeiros,
Maria Viviane de Menezes, Adailton Ferreira de Araújo**

[1]Programa de Pós-Graduação em Computação (PCOMP)
Universidade Federal do Ceará — Campus Quixadá
Quixadá, Ceará, Brasil

erickbastos.cs@alu.ufc.br, vivianemenezes@ufc.br

ivopdm@gmail.com, adailton@inf.ufg.br

Abstract. *This paper presents the development of information extraction from tender documents, focusing on technology products. The system integrates Natural Language Processing and machine learning techniques to extract relevant information from the documents. The proposed solution aims to optimize the time and accuracy of tender document analysis by dealing with the complexity and diversity of data present in the notices. The experimental results demonstrate the effectiveness of identifying bidding items, highlighting their potential for practical application in public procurement processes.*

1. Introduction

Bidding is a procedure used by the Brazilian public administration to acquire products, construction works, services, and disposals [da União 2024]. Regulated by Law No. 14.133/2021 [da República 2021], the bidding process comprises seven phases: preparatory, publication of the **bidding notice**, submission of proposals and bids (when applicable), judgment, qualification, appeals, and approval [dos Santos Chaves 2015]. The preparatory phase involves the preparation of the necessary documents and studies. In the publication phase of the notice, the document containing rules and specifications is released, divided into sections such as the object of the bidding, conditions of participation, and judgment criteria. In the submission of proposals and bids, interested companies submit their proposals according to the notice. The judgment phase evaluates the proposals, followed by qualification, which verifies the technical and legal capacity of the participants. The appeals phase allows challenges against decisions, ensuring transparency and fairness. The approval phase ratifies the result, authorizing the formalization of the contract.

For companies selling technological equipment, competitive participation in bidding requires identifying sales opportunities in the notices, which are usually available as PDF files on federal, state, and municipal government websites, such as the one shown in Figure 1 and Figure 2. In companies, the traditional process begins with the collection of notices from different platforms. Then, an analyst must manually read the content, understand the object of the bidding, identify products, quantities, and other requirements, and then evaluate whether the organization has the technical capacity and whether participation is economically advantageous. This is a highly manual, time-consuming, and repetitive procedure. With the introduction of automation, these steps of reading and

data extraction are delegated to computational systems, allowing the analyst to focus on strategic decision-making and defining the best way to participate.

From a financial perspective, efficient participation in bidding represents, for technology suppliers, a strategic opportunity to access highly relevant public markets, given the significant volume of resources managed through these contracts. In 2023, for example, the federal government committed more than R$ 150 billion in public procurement, according to data from the Transparency Portal [da União 2024]. For public agencies, in turn, bidding aims to ensure the selection of the most advantageous proposal, considering the best cost-benefit ratio, guaranteeing quality, innovation, and efficiency in the use of public resources. In this context, the automation of reading and analyzing notices through NLP techniques not only optimizes the process of identifying opportunities for suppliers but can also contribute to greater competitiveness and improvement of the submitted proposals.

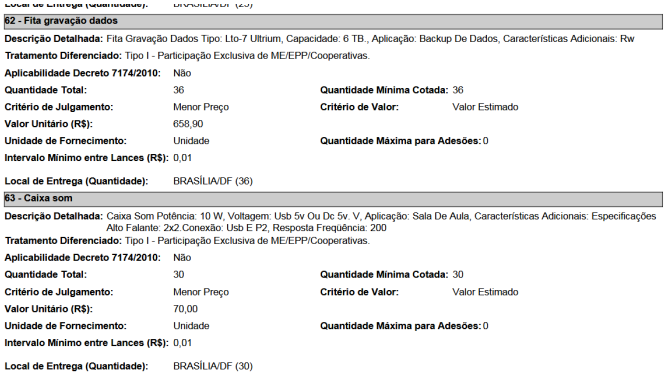
PREFEITURA MUNICIPAL DE MANHUAÇU
Lei Provincial nº. 2407 de 5/XI/1877 - Área: 628.43 Km² - Altitude: 612 metros
MANHUAÇU - MINAS GERAIS

ANEXO II - MODELO DE CARTA PROPOSTA

PROCESSO LICITATÓRIO Nº 1692/2025
PREGÃO ELETRÔNICO Nº 19/2025

ITEM	CATMAT	DESCRIÇAO MINIMA	UNIDADE	QUANT
1	614635	APARELHO TELEFÔNICO CELULAR ** SMARTPHONE SISTEMA OPERACIONAL ANDROID 14 Câmera Selfie de 12MP, Tela de 6.8" 1-120Hz, Armazenamento 512GB, 12GB RAM, Resolução 1440 x 3120 (QHD+) Modelo da CPU Cortex Velocidade da CPU 3,4 GHz Câmera Quádrupla de 200MP + 50MP +12MP + 10MP; Selfie de 12MP Dual Pixel AF;	UNID	1

Figure 1. Excerpt from a bidding notice for technological products formatted in a table.



62 - Fita gravação dados

Descrição Detalhada: Fita Gravação Dados Tipo: Lto-7 Ultrium, Capacidade: 6 TB., Aplicação: Backup De Dados, Características Adicionais: Rw
Tratamento Diferenciado: Tipo I - Participação Exclusiva de ME/EPP/Cooperativas.
Aplicabilidade Decreto 7174/2010: Não
Quantidade Total: 36 **Quantidade Mínima Cotada:** 36
Critério de Julgamento: Menor Preço **Critério de Valor:** Valor Estimado
Valor Unitário (R$): 658,90
Unidade de Fornecimento: Unidade **Quantidade Máxima para Adesões:** 0
Intervalo Mínimo entre Lances (R$): 0,01
Local de Entrega (Quantidade): BRASÍLIA/DF (36)

63 - Caixa som

Descrição Detalhada: Caixa Som Potência: 10 W, Voltagem: Usb 5v Ou Dc 5v. V, Aplicação: Sala De Aula, Características Adicionais: Especificações Alto Falante: 2x2.Conexão: Usb E P2, Resposta Frequência: 200
Tratamento Diferenciado: Tipo I - Participação Exclusiva de ME/EPP/Cooperativas.
Aplicabilidade Decreto 7174/2010: Não
Quantidade Total: 30 **Quantidade Mínima Cotada:** 30
Critério de Julgamento: Menor Preço **Critério de Valor:** Valor Estimado
Valor Unitário (R$): 70,00
Unidade de Fornecimento: Unidade **Quantidade Máxima para Adesões:** 0
Intervalo Mínimo entre Lances (R$): 0,01
Local de Entrega (Quantidade): BRASÍLIA/DF (30)

Figure 2. Excerpt from a bidding notice for technological products formatted in text.

This study focuses on the automatic extraction of structured information from public bidding documents, particularly those related to technological procurement. Specifically, the proposed approach aims to identify and extract four key elements contained in the notices: lot, item, product, and quantity. These elements represent the core procurement entities necessary for downstream analysis and classification tasks, enabling a standardized and machine-readable representation of tender information. This article is

organized as follows: Section 2 presents the theoretical framework on natural language processing; Section 3 discusses related work; Section 4 describes the methodology; Section 5 presents the results obtained; and Section 6 brings the conclusions and suggestions for future work.

2. Natural Language Processing

Natural Language Processing is a subfarea of Artificial Intelligence (AI), focused on developing computational models that simulate human linguistic comprehension and reasoning. It plays an important role in enabling more natural and accessible human-machine interactions [Hazboun et al. 2021]. Despite its relevance, implementing NLP systems remains challenging due to the complexity of linguistic structures and semantic context [Wang et al. 2021].

Advances in NLP are largely driven by Machine Learning techniques [Kang et al. 2020], which allow for effective modeling of language patterns. Tasks such as *segmentation* [Luca 2025], which organizes text into meaningful units; *summarization* [Luca 2025], which condenses content while preserving essential information; and *text classification*, which assigns semantic labels to text segments, are now performed in a remarkably efficient way. Supervised methods based on Deep Learning, such as CNNs [da Silva et al. 2022] and RNNs [da Silva et al. 2022], have demonstrated strong performance in tasks like sentiment analysis, news categorization, and question answering, especially in processing large volumes of unstructured data [Lavanya and Sasikala 2021]. More recently, large language models (LLMs)—pretrained on web-scale corpora and adapted via techniques such as in-context learning and instruction tuning—have achieved state-of-the-art results across these tasks, enabling zero- and few-shot generalization and robust performance in domain-specific corpora, including legal and administrative documents [OpenAI 2024]

2.1. Information Extraction (IE)

Information Extraction (IE) is a technique in NLP, aiming to convert unstructured textual data into structured representations by automatically identifying entities, relationships and events. IE is fundamental for applications such as knowledge graph construction, inference mechanisms and question answering systems. Typical tasks associated with IE include Named Entity Recognition (NER), Relation Extraction (RE), and Event Extraction (EE) [Xu et al. 2024]. Traditional IE models employ specialized architectures for each subtask, relying on specific annotation schemes. However, these models face significant limitations, such as dependence on large volumes of manually annotated data, high annotation costs, and low generalization capabilities in data-scarce contexts or with unseen labels. [Lou et al. 2023]

3. Related Work

The article by [Luca 2025] provides an in-depth analysis of the application of Natural Language Processing (NLP) techniques in the automation of document analysis in various domains, such as business, health, law, academic research, and finance. Using fundamental methods such as tokenization, lemmatization, stemming, stopword removal, named

entity recognition (NER), text classification, sentiment analysis, topic modeling, and summarization (extractive and abstractive), the work highlights how NLP is capable of transforming unstructured text into organized information, promoting efficiency, scalability, and accuracy in document flows. The study emphasizes the impact of deep learning models, especially those based on transformers, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), which offer significant advances in semantic understanding and natural language generation. The benefits of pre-trained language models, such as T5, RoBERTa, and DistilBERT, and their advantages in terms of performance and resource savings via fine-tuning are also discussed. On the other hand, the article recognizes the main challenges faced by NLP in document analysis, such as semantic ambiguity, linguistic variation, adaptation to specific domains, privacy and security of sensitive data, in addition to the high computational demand for training and real-time inference. Finally, the author points out promising directions for future research, including the development of multimodal models, compression and quantization techniques for environments with limited resources, greater transparency and interpretability of language models, and collaborative approaches between humans and AI systems, consolidating NLP as a strategic tool in digital transformation and document intelligence on an institutional scale. Our article distinguishes itself from the work of [Luca 2025] by focusing specifically on bidding documents (public procurement notices) and, unlike [Luca 2025], (i) adopting a *page-level* segmentation strategy suited to heterogeneous Brazilian tender PDFs and (ii) employing GPT-4 for few-shot *summarization* over page chunks to extract procurement-salient fields (lots,items, product, quantity) [OpenAI 2024].

The study by [ANDRADE and BAPTISTA 2022] proposes the automation of the analysis and auditing of **procurement documents** in PDF format through a supervised learning model capable of identifying key information in **public bidding notices**. Using data from the Government Portal of the State of Acre and the CRISP-DM methodology [Schröer et al. 2021], the authors tested various algorithms, with BERTimbau yielding the best results for sentence classification. The work is similar to ours in its use of BERTimbau for extracting information from bidding notices. However, our approach differs by incorporating other techniques for segmentation and using GPT-4o-mini for summarization.

Ito and Nakagawa [Ito and Nakagawa 2024] proposed the Tender Document Analyzer, a web-based system that combines supervised learning techniques (BERT) with enhancement through large language models (LLM). This approach enables efficient item extraction and identification of important phrases in tender documents. The authors demonstrated significant improvements over exclusively LLM-based methods, both in extraction performance and usability for bidders, particularly for less experienced users. This approach is very close to ours; the differences are that we apply it to Brazilian Portuguese language, we use different segmentation approach, and our approach extracts the information in fact, not only points out its position in document.

4. Methodology

In this section, we introduce the strategy and experiments adopted for tender documents information extraction development, following the CRISP-DM cycle: Dataset, Model, Evaluation, and Results[Loeza-Mejía 2024].

4.1. Dataset

This subsection presents the three datasets used for the text relevance classification task. These datasets were designed to evaluate the model's ability to distinguish relevant textual segments within bidding documents.

4.1.1. Text relevance classification datasets (v1, v2, and v3)

Three datasets were constructed with text fragments representing the sections, subsections, or full pages extracted from the bidding documents, each fragment being labeled as "relevant" or "not relevant". Examples of relevant sections include "1 - Bidding Items", "TECHNICAL SPECIFICATIONS OF THE OBJECT", and "QUANTITIES OF THE EQUIPMENT", while non-relevant content appears in sections such as "Introduction", "LOCATION, DEADLINES AND SERVICE WARRANTY", and "CONDITIONS FOR THE PROVISION OF SERVICES".

The segmentation and labeling processes varied across dataset versions:

- **v1 and v2 (section-based segmentation):** Text was automatically segmented by identifying the beginning of sections and subsections, followed by manual labeling of each fragment. Dataset v1 comprises 570 samples, whereas v2 contains 8,000 samples, each including the document name, the extracted content, and the corresponding label.
- **v3 (page-level segmentation):** This dataset was constructed using full-page text extractions instead of section-based fragments. Although the same binary labeling scheme was retained, the segmentation criterion was modified: entire pages were analyzed and labeled as relevant if they contained any quantitative information about the procured products, otherwise as not relevant. The v3 dataset comprises 170 bidding documents from 2022 to 2024, totaling 687 relevant and 2,285 non-relevant pages.

4.1.2. Dataset for Evaluation

The evaluation dataset was manually constructed from 202 bidding documents, selected from different procurement processes. These documents correspond exclusively to the public notice phase, excluding later contract-related materials. In total, the dataset comprises 2,792 annotated products. The files describing the procured items varied in format—such as full notices, terms of reference, or item lists. For each product, the following fields were annotated: **Lot Number**, **Item Number**, **Product Name**, and **Quantity**. This dataset provides a representative basis for assessing the model's extraction performance across different document structures and writing styles.

Main challenges included: (i) lack of textual standardization across documents (tables, lists, or frames), (ii) absence or inconsistency of lot identifiers, (iii) duplicated items due to ME/EPP reserved quotas, and (iv) aggregated quantities that combine open and reserved lots.

4.2. Model

For each attachment in the notice, the system extracts items via three complementary procedures: (i) parsing HTML tables; (ii) analyzing unstructured text; and (iii) integrating metadata gathered through web scraping/crawling from external sources in an upstream data engineering stage (e.g., the RHS Licitações portal, and Conlicitacao).[RHS Licitações 2025][ConLicitação 2025]. Finally, the outputs from each stage are consolidated by a deduplication step: items obtained by methods are compared and duplicates are merged into a single canonical record. The flow of this process is illustrated in Figure 3.

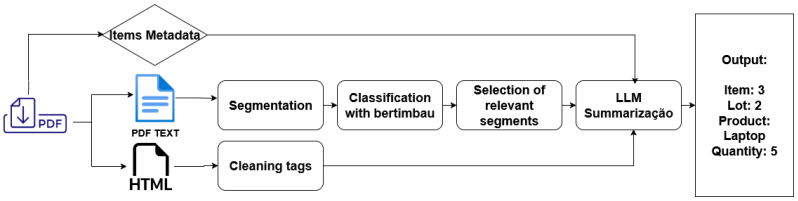


Figure 3. PDF File Extraction Flow

Extraction via HTML Tables: The code uses a function that locates and extracts items from structured HTML tables. In this flow, the document is converted to HTML format, and then unnecessary tags, that is, those that do not refer to the information tables, are removed. After this, the flow identifies the table tags in the files, allowing the LLM (Large Language Model) to summarize the information, excluding all HTML tags that are not `<table></table>`. This extraction technique is applied to extract item tables from documents.

PDF Text Extraction: The PDF text extraction process begins with text segmentation, which was evaluated using three different approaches—two of which are presented in [Silva et al. 2024]. In this study, we introduce a *page-level* segmentation method. After this step, the text segments are classified as *relevant* or *non-relevant* using a classifier that identifies and categorizes passages containing information about **lot**, **item**, **product**, and **quantity**. These classified segments are then processed by a Large Language Model (LLM), which summarizes the extracted information.

Item Metadata: At this stage, the system incorporates metadata obtained from external procurement sources [RHS Licitações 2025] [ConLicitação 2025] through web scraping and crawling, carried out in a previous data-engineering phase. Among the collected attributes, the *item* field stands out, as it is populated in roughly 60% of the analyzed records.When this field is available, its contents are directly extracted and provided to the LLM, which processes them to identify the items associated with each notice.

In the study discussed in [Silva et al. 2024], the performance of the extraction flow was evaluated using different segmentation approaches (by sentence and by section). Additionally, two techniques were explored for identifying key information in the text: one based on a *k-Nearest Neighbors* (k-NN) approach and another employing the *BERTimbau* language model. In the present work, we propose several improvements, including a new segmentation approach based on page-level division of the PDF content. We also updated the summarization model—from GPT-3.5 used in the previous work to GPT-4o-mini in

this study—and incorporated strategies to handle documents formatted with tables and metadata-based items.

The extractive summarizer is designed to analyze text containing product information and extract specific data such as product name, quantity, item number, lot number. This information is returned in a structured JSON format. The following Figure 4 shows how the prompt used is configured.

```
prompt: >
  Você está recebendo um texto que pode ter informações de produtos solicitados para uma compra. Os produtos são {lista_de_produtos}.
    Você deve identificar e extrair quais produtos são solicitados e extrair as informações de quantidade, valor, número do item, número do lote e restrição de
    cada produto dentro do texto.Caso você não identifique informações de quantidade, valor, número do item, número do lote ou restrição, você deve responder
    com o texto "Não há" para cada um. Se não encontrar nenhum produto, responda apenas com "0".

  Formato esperado do retorno:
  - Uma lista de dicionários onde cada dicionário representa um produto.
  - Cada dicionário deve conter as chaves "produto", "quantidade", "valor", "numero_do_item", "numero_do_lote" e "restricao".
  - "produto" deve conter o nome do produto.
  - "quantidade" deve conter a quantidade do produto ou "Não há" se a quantidade não for identificada.
  - "valor" deve conter o valor do produto ou "Não há" se o valor não for identificado.
  - "numero_do_item" deve conter o número do item ou "Não há" se não for identificado.
  - "numero_do_lote" deve conter o número do lote ou "0" se não for identificado.
  - "restricao" deve conter a restrição ou "Não há" se não for identificada.

  Exemplo de estrutura JSON esperada (não copie diretamente):
  [
     {{"produto": "Nome do Produto 1", "quantidade": "Quantidade do Produto 1", "valor": "Valor do Produto 1",
      "numero_do_item": "Número do Item 1", "numero_do_lote": "Número do Lote 1", "restricao": "Restrição do item"}},

     {{"produto": "Nome do Produto 2", "quantidade": "Quantidade do Produto 2", "valor": "Valor do Produto 2",
     "numero_do_item": "Número do Item 2", "numero_do_lote": "Número do Lote 2", "restricao": "Restrição do item 2"}},

     {{"produto": "Nome do Produto 3", "quantidade": "Não há", "valor": "Não há", "numero_do_item": "Não há", "
     numero_do_lote": "0", "restricao": "Não há"}}
  ]

  É crucial apresentar cada produto com seu nome completo e respectiva quantidade, valor, número do item, número do lote e restrição ou apenas "0",
  conforme detalhado no texto, com precisão na identificação e na organização das informações. Não adicione especificação do produto.
```

Figure 4. Prompt used to summarize information

4.3. Evaluation

To evaluate the cases, we report the results in multiple metrics, including F1-score, precision, recall, and additional performance measures for relevance classification.

4.3.1. Summarization evaluation metrics

In this context, **Precision** measures the proportion of correctly extracted fields among all extractions made by the model. For example, when the system identifies several *products*, precision indicates how many of those correspond to true products present in the notice.

Recall quantifies the model's ability to find all relevant instances that actually exist in the document. For example, in the extraction of *quantities*, recall measures how many of the true quantities described in the tender were successfully detected.

The **F1-score** represents the harmonic mean between precision and recall, providing a balanced view of accuracy and completeness. A higher F1-score indicates that the system achieves a good equilibrium—correctly identifying *lot*, *item*, *product*, and *quantity* fields while minimizing both false detections and omissions.

4.3.2. Relevance text classifier metrics

Additionally, we evaluated the relevance text classifier using the following metrics:

- **Time Spent** measures the average time it takes the model to classify a text snippet as relevant or not. This metric provides an indication of the classifier's efficiency.

- **Mean Precision** is the average precision calculated from each document's precision, indicating the proportion of correctly identified relevant snippets
- **Mean Recall** is the average recall across all documents, quantifying the model's ability to identify all relevant snippets, even if some are missed.
- **Mean F1 Score** is the average F1 score of all documents, providing a balanced measure of precision and recall in identifying relevant snippets.
- **Accuracy** reflects the percentage of correctly classified text snippets relative to the total text snippets, providing an overall view of the classifier's reliability.

These metrics together provide a comprehensive evaluation of the extraction pipeline, revealing not only how accurate the model's predictions are but also how complete its coverage is across all procurement-relevant fields.

4.4. Results

This section presents the results obtained from both the summarization process and the relevance text classification task. The evaluation of the model's performance is divided into two parts: the summarization results, which assess the extraction of specific fields from procurement documents, and the relevance text classification results, which measure the ability of the model to identify relevant portions of text.

4.4.1. Summarization results

Table 1 presents the numerical results obtained for each field. The highest F1 score was achieved for the `product` field, while lower results were observed for `item`, `qtd`, and `lot`, indicating variations in the model's ability to capture different types of information within heterogeneous tender documents.

	Precision (%)	**Recall (%)**	**F1 Score**
product	75.40	56.15	64.36
item	70.59	50.07	58.58
qtd	63.01	41.57	50.09
lot	66.16	44.96	53.53

Table 1. Evaluation Metrics Table

Across all categories, the results show that precision values remained consistently higher than recall, indicating that the model tends to be conservative when identifying valid entities. The `product` field achieved the best quantitative performance, while `qtd` and `lot` exhibited the lowest recall rates, suggesting a reduced sensitivity to numeric and structured patterns.

4.4.2. Relevance text classifier results

The evaluation of the relevance text classifier, which classifies segments of text as relevant or non-relevant, yielded the following results:

Time Spent: 690 milliseconds

Mean F1-Score: 0.785

Recall Médio: 0.84

Mean Precision: 0.83

Accuracy: 0.84

These metrics show that the relevance classifier performs effectively, with relatively high precision and recall values. The Mean F1-Score of 0.785 reflects a good balance between precision and recall, indicating that the model successfully identifies relevant segments while minimizing false positives and false negatives. The accuracy value of 0.84 demonstrates the classifier's reliability in distinguishing between relevant and non-relevant segments of text.

5. Conclusion and Discussion

The evaluation results demonstrate that the proposed extraction model performs more effectively for the `product` field, with a precision of 75.40% and recall of 56.15%, yielding an F1 score of 64.36%. This indicates that the model correctly identifies most product names but still fails to detect a considerable portion of relevant instances.

For the `item` category, precision reached 70.59% and recall 50.07%, with an F1 score of 58.58%, suggesting greater difficulty in distinguishing item descriptions. The `qtd` field presented the lowest performance (F1 = 50.09%), revealing challenges in recognizing numeric patterns for quantities. Meanwhile, the `lot` category achieved intermediate results (F1 = 53.53%), with acceptable precision but reduced recall, showing a notable loss of relevant information.

In summary, the model achieves satisfactory precision levels but presents low recall, demonstrating a tendency to miss relevant information. The F1 scores suggest a balanced yet improvable performance. Future work should focus on enhancing contextual understanding and pattern recognition, particularly for numeric and tabular data. Incorporating domain-specific post-processing and hybrid approaches that combine rule-based filtering with large language models may further improve coverage and accuracy in automated extraction tasks for public procurement documents.

References

[ANDRADE and BAPTISTA 2022] ANDRADE, S. and BAPTISTA, C. d. S. (2022). Uso de processamento de linguagem natural e aprendizagem de máquina para a extração de informação em editais de licitações não-estruturados. In *Universidade Federal de Campina Grande*. UFCG.

[ConLicitação 2025] ConLicitação (2025). Conlicitação. Acessado em 05-10-2025.

[da República 2021] da República, P. (2021). Lei de licitações e contratos administrativos. last accessed 18 jul. 2024.

[da Silva et al. 2022] da Silva, F., Guimarães, G., Marcacini, R., Queiroz, A., Borges, V., Faleiros, T., and Garcia, L. (2022). Named entity recognition approaches applied to legal document segmentation. In *Anais do X Symposium on Knowledge Discovery, Mining and Learning*, pages 210–217. SBC.

[da União 2024] da União, C.-G. (2024). Portal da transparência. last accessed 18 jul. 2024.

[dos Santos Chaves 2015] dos Santos Chaves, E. (2015). Aspectos importantes da fase interna da licitação: uma análise sobre o conjunto de elementos necessários e suficientes para a caracterização do objeto do processo licitatório. *Revista Controle: Doutrinas e artigos*, 13(1):149–170.

[Hazboun et al. 2021] Hazboun, F., Owda, M., and Owda, A. (2021). A natural language interface to relational databases using an online analytic processing hypercube. *AI*, 2(4):720–737.

[Ito and Nakagawa 2024] Ito, T. and Nakagawa, S. (2024). Tender document analyzer with the combination of supervised learning and llm-based improver. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion)*, New York, NY, USA. ACM.

[Kang et al. 2020] Kang, Y., Cai, Z., Tan, C., Huang, Q., and Liu, H. (2020). Natural language processing (nlp) in management research: A literature review. *Journal of Management Analytics*, 7(2):139–172.

[Lavanya and Sasikala 2021] Lavanya, P. and Sasikala, E. (2021). Deep learning techniques on text classification using natural language processing (nlp) in social healthcare network: A comprehensive survey. In *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, pages 603–609.

[Loeza-Mejía 2024] Loeza-Mejía (2024). Comparative study of kdd and crisp-dm methodologies. In *Proceedings of Ninth International Congress on Information and Communication Technology: ICICT 2024, London, Volume 3*, volume 1013, page 317. Springer Nature.

[Lou et al. 2023] Lou, J., Lu, Y., Dai, D., Jia, W., Lin, H., Han, X., Sun, L., and Wu, H. (2023). Universal information extraction as unified semantic matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13318–13326.

[Luca 2025] Luca, C. (2025). Natural language processing (nlp) for document analysis.

[OpenAI 2024] OpenAI (2024). Gpt-4 technical report.

[RHS Licitações 2025] RHS Licitações (2025). Rhs licitações. Acessado em 05-10-2025.

[Schröer et al. 2021] Schröer, C., Kruse, F., and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–534.

[Silva et al. 2024] Silva, E., Medeiros, I., Menezes, M., and Kamikawachi, D. (2024). Segmentation and summarization for extracting information about information technology equipment from government procurement notice. In *Anais do XII Symposium on Knowledge Discovery, Mining and Learning*, pages 145–152. SBC, Porto Alegre, RS, Brasil.

[Wang et al. 2021] Wang, B., Yin, W., Lin, X., and Xiong, C. (2021). Learning to synthesize data for semantic parsing. *arXiv preprint*. `https://arxiv.org/abs/2104.05827`.

[Xu et al. 2024] Xu, D., Chen, W., Peng, W., Zhang, C., Xu, T., Zhao, X., Wu, X., Zheng, Y., Wang, Y., and Chen, E. (2024). Large language models for generative information extraction: a survey. *Frontiers of Computer Science*, 18(6):186357.