



LiPSet: A Comprehensive Dataset of Labeled Portuguese Public Bidding Documents

Mariana O. Silva   [Universidade Federal de Minas Gerais | mariana.santos@dcc.ufmg.br]

Gabriel P. Oliveira  [Universidade Federal de Minas Gerais | gabrielpoliveira@dcc.ufmg.br]

Henrique Hott  [Universidade Federal de Minas Gerais | henriquehott@dcc.ufmg.br]

Larissa D. Gomide  [Universidade Federal de Minas Gerais | larissa.gomide@dcc.ufmg.br]

Bárbara M. A. Mendes  [Universidade Federal de Minas Gerais | barbaramit@ufmg.br]


Clara A. Bacha  [Universidade Federal de Minas Gerais | clarabacha@ufmg.br]

Lucas L. Costa  [Universidade Federal de Minas Gerais | lucas-lage@ufmg.br]

Michele A. Brandão  [Instituto Federal de Minas Gerais | michele.brandao@ifmg.edu.br]

Anisio Lacerda  [Universidade Federal de Minas Gerais | anisio@dcc.ufmg.br]

Gisele L. Pappa  [Universidade Federal de Minas Gerais | glpappa@dcc.ufmg.br]

 *Computer Science Department, Universidade Federal de Minas Gerais, Av. Presidente Antônio Carlos, 6627, Pampulha, Belo Horizonte, MG, 31270-901, Brazil.*

Received: 14 June 2023 • **Published:** 5 April 2024

Abstract Collecting, processing, and organizing governmental public documents pose significant challenges due to their diverse sources and formats, complicating data analysis. In this context, this work introduces LiPSet, a comprehensive dataset of labeled documents from Brazilian public bidding processes in Minas Gerais state. We provide an overview of the data collection process and present a methodology for data labeling that includes a meta-classifier to assist in the manual labeling process. Next, we perform an exploratory data analysis to summarize the key features and contributions of the LiPSet dataset. We also showcase a practical application of LiPSet by employing it as input data for classifying bidding documents. The results of the classification task exhibit promising performance, demonstrating the potential of LiPSet for training neural network models. Finally, we discuss various applications of LiPSet and highlight the primary challenges associated with its utilization.

Keywords: dataset, document classification, data labeling, open government data, public bidding

1 Introduction

The implementation of the Access to Information Law (Law No. 12,527, sanctioned on November 18, 2011),¹ in Brazil, has significantly increased citizens' access to public information from various governmental bodies. Although this information is available in different file formats, often lacking standardization, it is important for many applications [Mata *et al.*, 2019; Shimron *et al.*, 2022]. For example, Pereira [2022] investigates how open government data on Brazilian education are used by a given community, even for defining projects. On the other hand, Costa *et al.* [2022] use data from public bids to identify possible fraud in such bids.

Specifically, open data on public bids comprehend a range of documents, including notices, errata, minutes, contracts, adjudication, and homologation records. Each document has a specific format, providing information about the different stages of a bidding process, from publication to approval. However, these documents often lack standardization and are typically provided in different formats, posing challenges to effective utilization. Therefore, extracting meaningful information from these documents requires applying specific techniques, such as Natural Language Processing (NLP), which involves developing computational models to process and in-

terpret information expressed in natural language [Meera and Geerthik, 2022].

The collection, analysis, and processing of these public bidding documents present challenges for both humans and machines. Humans require tools to manage and analyze large volumes of documents to gain comprehensive insights. Meanwhile, machines must automate the collection, analysis, and processing tasks. The development of LiPSet by Silva *et al.* [2022] represented a step towards addressing these challenges. LiPSet is a dataset constructed by collecting, processing, and labeling public bidding documents from Minas Gerais state, specifically in Portuguese.

This work extends the paper that introduced LiPSet [Silva *et al.*, 2022] and was presented on the Dataset Showcase from the 37th Brazilian Symposium on Databases. As a new material, it presents an experimental setup based on Recurrent Neural Networks (RNN) that utilizes LiPSet to classify public bid documents. Overall, we observe that LiPSet can be used to train a classification model, allowing different experimental setups, and the results are promising.

Our main contributions are summarized as follows.

1. Introduce LiPSet and provide an overview of the data collection process, the methodology employed for data labeling, and the creation of a meta-classifier to assist in the manual labeling process.

¹About the Access to Information Law: <https://bit.ly/acao-a-informacao>

2. Conduct an exploratory data analysis to summarize key features and contributions of the LiPSet dataset, emphasizing its relevance and potential applications.
3. Showcase a real-world application of LiPSet by employing it as input data for classifying public bid documents and highlighting the potential of LiPSet for training neural network models.
4. Investigate additional applications of LiPSet beyond classification, shedding light on its versatility.

The remainder of this work is organized as follows. Section 2 provides an overview of related works. Section 3 outlines the methodology employed to construct the LiPSet dataset. Section 4 characterizes LiPSet based on its meta-classes. Section 5 presents a real-world application of LiPSet by employing it to classify public bid documents. Section 6 explores additional applications of this dataset. Section 7 discusses the main limitations and challenges associated with LiPSet. Finally, Section 8 details final considerations.

2 Related Work

Since the enactment of the Freedom of Information Act, several papers have been published targeting the use of open government data [Coelho *et al.*, 2022; Lima *et al.*, 2020; Lyra *et al.*, 2021; Nai *et al.*, 2022]. Public data are generally collected and organized through different strategies, for example, storing in graph-oriented databases [Erven *et al.*, 2017] or labeling the data [Lima *et al.*, 2020]. Public bidding data, in particular, has garnered attention in several studies, demonstrating promising results in its application. For example, Gabardo and Lopes [2014] use social network analysis techniques to verify the formation of cartels in civil construction companies in Paraná. Similarly, da Silva *et al.* [2020] discuss applying data mining techniques to support the army’s internal audit processes.

Despite advances in research on open data, some challenges persist when dealing with data from different government spheres in Brazil [Coelho *et al.*, 2022; de Oliveira and Silveira, 2018]. Consequently, many studies aim to address these challenges and enhance the quality, accessibility, and reusability of government data. An example is QualiSuS, a relational database created from data extraction from the Portal DataSUS [Clarindo *et al.*, 2019]. It adopts standards such as disease identifiers and data available in CSV and JSON format. In the legal area, JusBD presents an unlabeled dataset for forensic audits in data from the judiciary branch [Mata *et al.*, 2019]. Finally, Araújo and Souza [2011] use a Web collector to obtain data from Brazilian politicians.

As presented in [Nai *et al.*, 2022], different publications that use public data aim to detect fraud. Such literature review also revealed that most publications adopt traditional machine learning methods, and a small group of papers use neural networks and network analysis. Nai *et al.* [2022] also claim that recent methodologies developed NLP and neural network techniques to detect public data fraud since most of them are in text format.

Dealing with extracting information from PDF documents is a non-trivial task, mainly due to the lack of standardization

[Mata *et al.*, 2019]. In this context, LiPSet makes a valuable contribution to analyzing government public documents as it consists of data extracted from Brazilian government documents in PDF format. NLP techniques were employed to extract and preprocess the textual content of these documents [Pedrosa *et al.*, 2021]. LiPSet also provides labeled data, facilitating its utilization in supervised machine learning algorithms. Furthermore, including Portuguese documents in LiPSet adds to the challenges of effectively applying NLP techniques [Carneiro *et al.*, 2017].

Indeed, this work presents a real-world application of LiPSet by classifying public bid documents. This classification task is similar to previous works such as [Coelho *et al.*, 2022] and [Pedrosa *et al.*, 2021], which utilize embeddings to represent documents. Additionally, it shares similarities with the work of [Carneiro *et al.*, 2017] by employing Portuguese text in the classification task. Notably, [Coelho *et al.*, 2022] present a study more closely related to this paper by focusing on classifying legal document texts.

3 LiPSet

This section introduces LiPSet, a comprehensive dataset comprising labeled public bidding documents from Minas Gerais, Brazil. Sections 3.1 and 3.2 describe the data collection and dataset construction processes. Section 3.3 presents the evaluation of the meta-classifier based on the keyword approach. Next, Section 3.4 provides insights into the storage and organization of the dataset. Finally, Section 3.5 highlights the public download location for accessing LiPSet and instructions on upgrading the dataset version.

3.1 Data Crawling

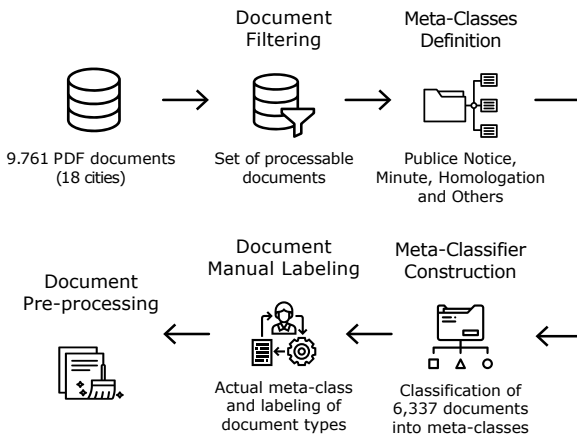
For collecting public bidding documents, the data sources used were transparency and/or bidding portals from 18 municipalities in Minas Gerais. The first stage of the data collection process is the survey of the links of the portals that must have their bids collected, and it defines which municipalities must have the bidding documents collected. Then, the portal of each municipality is analyzed, since most of them have no standard for making documents available. Considering the structure of the portals, a web crawler is developed to automate the process of accessing each available link on the portal pages and downloading the bidding documents.

Generally, transparency portals of various cities host links that provide access to individual bidding documents. In this context, a web crawler was employed to simulate “clicking” on each link, thereby facilitating the automated download of the desired files. The data collection process occurred in two distinct periods: July and December 2021. Due to privacy considerations associated with the Analytical Capacities Program in collaboration with the Public Ministry of Minas Gerais, we cannot provide the specific code used for accessing the transparency portals.

Table 1 presents the total number of files collected for each municipality and the corresponding collection month. The dataset consists of 9,761 documents, all in PDF format (Portable Document Format). While the portals of each

Table 1. Distribution of files collected in July and December 2021.

Month of the crawler	#	City	# Document
July	1	Arantina	937
	2	Coqueiral	1,528
	3	Cristais	1,737
	4	Ijaci	455
	5	Itamarati de Minas	1,111
	6	Olaria	42
	7	Passa-Vinte	412
	8	Pirapetinga	1,108
	9	Ribeirão Vermelho	686
	10	São Bento Abade	275
December	1	Bias Fortes	159
	2	Cana Verde	402
	3	Contagem	136
	4	Governador Valadares	93
	5	Palma	93
	6	Pedro Teixeira	179
	7	Rio Preto	279
	8	São Tomé	129
Total	18		9,761

**Figure 1.** Methodology used in the labeling of Brazilian governmental documents.

municipality also offer files in HTML, DOC, and CSV formats (Comma-separated values file), this study focuses exclusively on PDF documents for standardization during data processing. Notably, HTML and CSV files primarily contain information extracted directly from the visited web pages. Their inclusion in the dataset is mainly driven by the structural aspects of the respective web pages rather than their content. Therefore, these files were disregarded for this work, and the analysis is limited to PDF documents.

3.2 Methodology for Document Labeling

In this section, we present the methodology developed for building the LiPSet, which involves labeling public bidding documents. Figure 1 provides an overview of the applied methodology, and each step is described as follows.

Document Filtering. Following the data collection phase, the collected documents undergo a filtering process to separate them into non-processable (scanned/corrupted documents) and processable (documents that allow direct extraction). This separation is performed using the PDFPlumber Python library,² which cannot extract text from scanned, corrupted documents or images. Documents from which no text can be extracted are considered unprocessable and are

²PDFPlumber: <https://github.com/jsvine/pdfplumber>

Table 2. Meta-classes and associated keywords.

Meta-class	Keywords
Minutes	minutes, public session
Public notice	invitation, notice
Adjudication/Homologation	adjudication, homologation
Others	schedule, addition, order of service, answer, extract, official diary, warning of, rectification, administrative contract

Algorithm 1: Heuristic Meta-classifier

```

Input: Documents of bidding processes in PDF and a set of keywords for each meta-class
Output: The predicted meta-class for each document
1 begin
2   for each PDF bidding document do
3     Extract the title and first-page content of the document;
4     Declare countWordsTitle variable; // Occurrence of keyword
      in title, by meta-class
5     Declare countWordsContent variable; // Occurrence of
      keyword in first-page content, by meta-class
6     for each meta-class do
7       Update countWordsTitle with the number of keywords
          that occurred in the title;
8       Update countWordsContent with the number of keywords
          that occurred in the first-page content;
9     end
10    if "Others" meta-class keywords exist then
11      | meta_class ← "Others"
12    end
13    if "Adjudication/Approval" meta-class keywords exist then
14      | meta_class ← "Adjudication/Approval"
15    end
16    Sort countWordsTitle in descending order;
17    Sort countWordsContent in ascending order;
18    if there is a keyword occurrence in the first-page content then
19      | meta_class ← meta-class associated
20    end
21    meta_class ← "Others"
22  end
23 end
24 return List of labeling documents by meta-class

```

excluded from further analysis. Processable documents proceed to the subsequent stages of the methodology.

Meta-Classes Definition. In this step, we identify the essential types of documents encountered in the bidding process. Based on empirical knowledge gained from document analysis, we propose four meta-classes: Minutes (all available minutes), Public Notice (public notice documents and invitations in the Invitation mode), Adjudication/Approval (adjudication and homologation documents, or documents that contain both types of information), and Others (other document types such as errata, annexes, contracts, and descriptive memorials). Section 4 provides a more detailed description of the meta-classes mentioned here.

Meta-Classifier Construction. Through continuous interactions with the bidding documents, we observed structural patterns and identified keywords that showed promise in distinguishing documents belonging to the proposed meta-classes. Table 2 presents the defined keywords used to identify each meta-class in the documents. Based on this framework, we developed a heuristic classification method, referred to as the meta-classifier, to facilitate the manual labeling process. Algorithm 1 outlines the main steps of the meta-classifier, which relies on keywords and analyzes their occurrence in the title and content of each bidding document. After

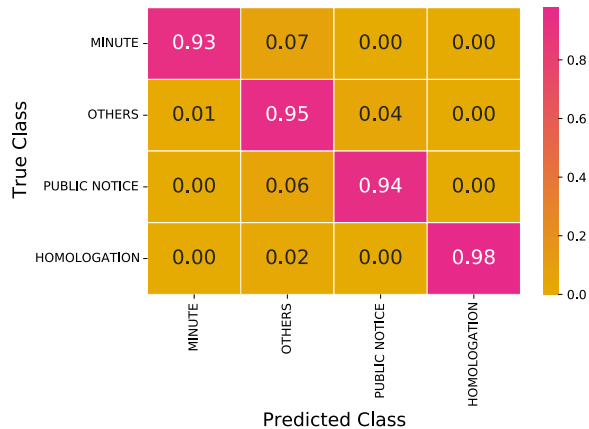


Figure 2. Confusion matrix of the meta-classification.

designing and refining the rules used in the meta-classifier’s construction, it was applied to all 6,337 documents obtained after the filtering step.

Document Manual Labeling. To ensure the accuracy of the dataset, seven members of the Analytical Capacities Program - MPMG/UFMG actively participated in the manual labeling process. All 6,337 documents were analyzed, and each document was manually labeled according to its corresponding meta-class and document type. The labeling process resulted in a total of 56 document types, including categories such as errata, notice, ratification, official journal publication, contract, and amendment. As most documents’ actual classes are present in the file titles, the manual labeling process involved examining the titles of each document. Verification of agreement between labelers was optional due to the reliability of the labeling process.

Document Pre-processing. Following the manual labeling, the text of each document was pre-processed using a set of functions. This pre-processing involved transforming the text to lowercase and removing proper names, emails, URLs, pronouns, adverbs, special characters, accents, stop-words, hours, number symbols, numbers, contracted and shortened words, single letters, and extra spaces. Proper nouns were removed based on exact matches to names present in a dictionary containing over 7,000 common Brazilian names.

3.3 Meta-classifier Evaluation

The evaluation of the meta-classifier based on the keyword approach yielded impressive results, surpassing expectations. Despite its simplicity, the meta-classifier demonstrated an accuracy of 94% and a Macro F1-score of 95%. Figure 2 presents the confusion matrix generated from the classification of all documents. The horizontal lines in the matrix represent the original class of the documents and should always sum up to 1, while the vertical lines represent the class assigned by the model. The diagonal of the matrix shows the intersection between the predicted and actual classes, indicating the proportion of correctly predicted documents.

The overall performance of the proposed model is promising for all meta-classes, with accuracy rates ranging from 93% to 98%. Notably, the meta-class with the highest error rate is “Minutes”, where 7% of the documents were mis-

Table 3. Data dictionary, containing an example entry.

Field	Type	Example
<i>file_id</i>	string	d2a0a04e5954c3095c1c1bbabc5a107
<i>original_name</i>	string	d2a0a04e5954c3095c1c1bbabc5a107.pdf
<i>n_pages</i>	int	1
<i>text_content</i>	array	[“PREFEITURA MUNICIPAL DE OLARIA - TERMO DE RETIFICAÇÃO - Processo Licitatório nº 055/2019 Pregão Presencial nº 014/2019, SOFREU ALTERAÇÕES na data de entrega de documentos de habilitação e proposta, devido o objeto da licitação estar escrito incorretamente, dessa forma, ONDE SE LÊ dia 22/05/2019, LEIA – SE dia 30/05/2019 as 09:00 (nove) horas ...”]
<i>table_content</i>	array	[]
<i>status</i>	string	SUCCESS
<i>city</i>	string	olaria
<i>text_preprocessed</i>	string	termo retificacao processo licitatorio pregao presencial sofreu alteracoes data entrega documentos habilitacao proposta devido objeto licitacao estar escrito incorretamente forma le dia leia ...
<i>meta_class</i>	string	OTHERS
<i>type_document</i>	string	erratum

classified as belonging to the “Others” meta-class. On the other hand, the “Homologation” and “Others” meta-classes achieved hit rates of 98% and 95%, respectively. These results demonstrate the meta-classifier’s effectiveness in accurately classifying documents within these meta-classes. However, a more complex classifier may be required to improve classification performance for the remaining two meta-classes (“Public Notice” and “Minutes”).

3.4 Data Storage and Organization

To store the data related to each document, JSON format (JavaScript Object Notation) files were used, as they can be easily converted into dictionaries. This choice offers several advantages, including the ability to store different data types, providing flexibility in the stored information. Table 3 presents the fields contained in each JSON file, along with their corresponding data types and example entries.

Regarding the information stored in each field, a standardized hexadecimal identification code is stored in the “file_id” field, unique to each document, with the original name of the document in the source database stored in the “original_name” field, and the number of pages entered in the “n_pages” field. The fields “text_content” and “table_content” each have an array of the texts and tables in the document’s original file. Both pieces of information were extracted using the PDFPlumber library. Finally, the “status”, “city” and “text_preprocessed” fields store the document’s status (processable or not), the city of origin, and the preprocessed text.

To ensure the dataset’s usability for future applications, each JSON file includes the fields “meta_class” and “type_document”. These fields store the meta-class and type of document obtained from the manual labeling process described in Section 3.2. For instance, this information can be valuable in text and document classification tasks.

3.5 Usability and Update

LiPSet is publicly available in a repository on Zenodo.³ For each municipality in the dataset, a file containing information on all public bidding documents is made available, including the processed text, the meta-class, and the document

³LiPSet: <https://doi.org/10.5281/zenodo.6974237>

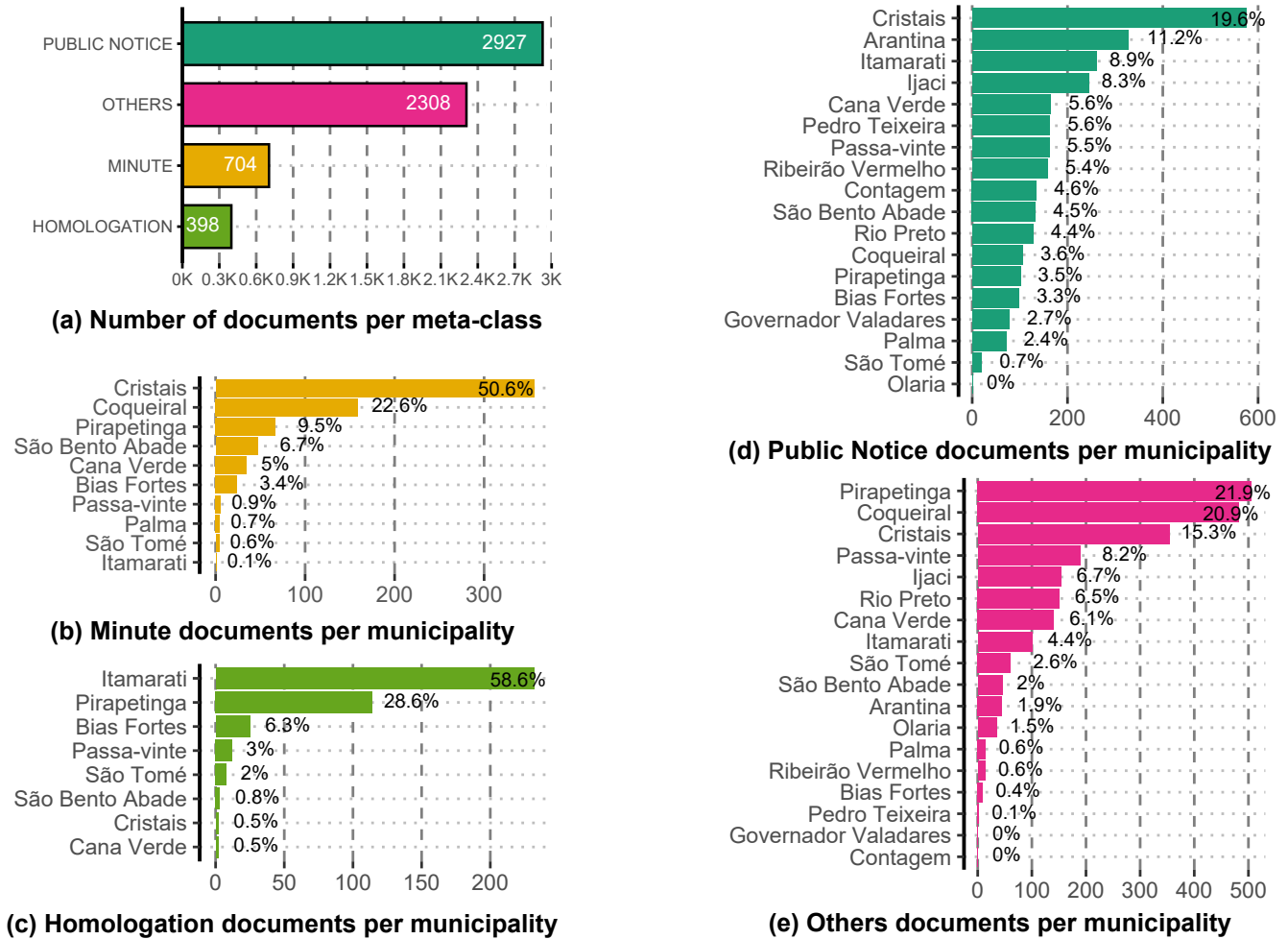


Figure 3. (a) Number of documents per meta-class. (b–e) Number of documents for each meta-classes by municipality.

type. This file is available in two versions (CSV and JSON), which can be used depending on the application considered. For example, CSV files may be more useful in complex analysis in Python or R, while JSON files are more useful in web applications. The common Brazilian names dictionary is also available in the Zenodo repository.

To update LiPSet by incorporating new documents from various cities and states, researchers and contributors can follow a systematic process to ensure the seamless integration of geographically diverse data. The initial step involves collecting public bidding documents through the transparency portals of each municipality or state. Note that such a step is optional if the desired documents are already accessible through other reliable sources. Next, researchers can integrate the acquired data into LiPSet by following each step of the document labeling methodology outlined in Section 3.2. Our methodology provides a structured and consistent framework for labeling documents, ensuring uniformity and compatibility within the dataset.

4 Dataset Characterization

This section provides a characterization of the documents included in LiPSet based on their meta-class and municipality (Section 4.1) as well as the distribution of documents according to the number of pages (Section 4.2).

4.1 Distribution of Documents by Meta-class and Municipality

Of the 9,761 documents from 18 municipalities, 2,223 were not classified due to being unprocessable, identified when the “status” field is marked as *FAILED*. Additionally, 1,201 documents are of the PDF type but do not belong to any meta-class because they are images, blueprints, or attached photographs. Therefore, LiPSet has 6,337 public bidding documents classified into one of the four meta-classes: Adjudication/Homologation, Minutes, Public Notice, and Others. Figure 3a presents the distribution of the number of documents by meta-class, revealing that approximately 83% of the documents belong to the Public Notice and Others meta-classes.

Figures 3b to 3e depict the distribution of documents for each meta-class across different municipalities. The 704 documents labeled with the Minutes meta-class are distributed among ten cities, with approximately 51% of them originating from the municipality of Cristais (Figure 3b). Similar patterns can be observed for the Homologation meta-class in Figure 3c. Out of the 398 labeled documents, approximately 87% are attributed to Itamarati and Pirapetinga. Conversely, all 18 municipalities encompass documents from the Public Notice and Others meta-classes, as illustrated in Figures 3d and 3e. Cristais, Arantina, and Itamarati are responsible for nearly 40% of the Public Notices, while Pirapetinga, Coqueiral, and Cristais account for over 50% of Others.

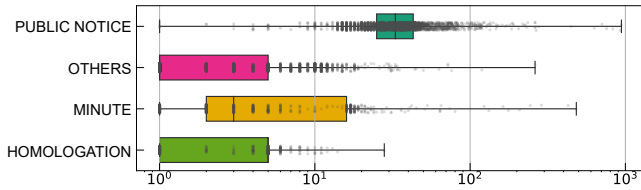


Figure 4. Distribution of the number of pages on a logarithmic scale by meta-class.

4.2 Characteristics of the Meta-classes in the Public Bids Documents

Figure 4 shows the distribution of the number of pages in the documents across different meta-classes. Regarding the Public Notice meta-class, a wide dispersion is observed in the values of the number of pages, indicated by numerous outliers. The document length ranges from 1 page, the smallest value, to 943 pages, the largest value. The most frequently occurring number of pages in Public Notices is 27. For the Others meta-class,⁴ the number of pages varies between 1 page, the most prevalent value, and 262 pages, the maximum outlier. In the Homologation meta-class, no outliers are present, and the document length ranges from 1 to 28 pages, with the most common value being 5. Finally, the Minutes meta-class exhibits a significant dispersion in the number of pages, as evidenced by the presence of outliers. The document spans 1 to 483 pages, with 1 being the most prevalent value.

Analyzing the number of pages can help determine the relevant page range to consider for applications using LiPSet. For instance, it may be sufficient to focus on the text’s vocabulary within the first few pages to classify bidding documents, given that most documents do not exceed 10 pages.

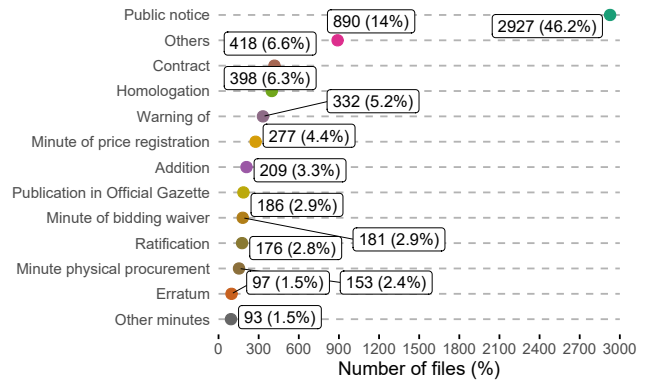
5 Classifying Public Bid Documents with LiPSet

LiPSet can be used to train a classification model and then classify new documents without the effort of new manual labeling. In addition, with suitable adaptations, the classification model can be applied to classify documents from the bidding processes of other Brazilian states. This is possible because the bidding processes are similar in different spheres of government, resulting in comparable documents. In this section, we demonstrate a practical application of LiPSet for classifying public bid documents through the following steps: definition of classes and document labeling (Section 5.1), description of the experimental setup and classification model definition (Section 5.2), and presentation of the results of the classification task (Section 5.3).

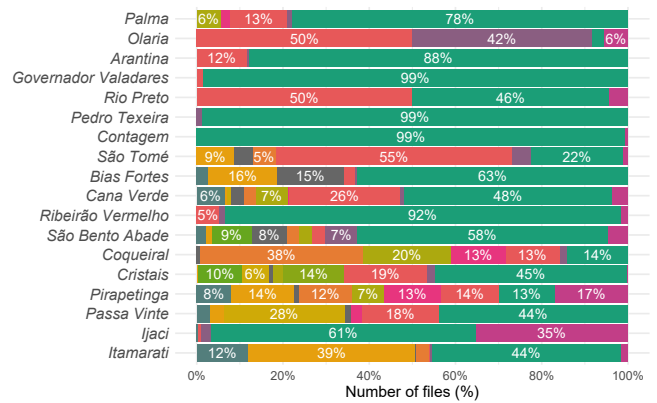
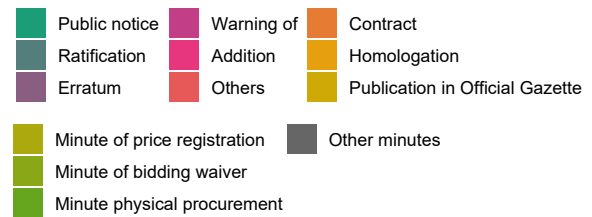
5.1 Definition of Classes in Public Bids Documents

As specified in Section 3.2, all 6,337 documents manually received a label according to their type. In total, 56 types of documents were identified and after several analyses, 13 possible classes were defined according to the doc-

⁴Other document types such as errata, annexes, contracts, and descriptive memorials.



(a) Distribution of files per class



(b) Distribution of files per class and cities

Figure 5. Distribution of files per (a) class, and (b) class and cities.

ument type. In short, such classes are determined based on the four meta-classes: Minutes (subdivided into four classes: bidding waiver, physical procurement, price registration, others), Public Notice, Adjudication/Homologation, and Others (subdivided into seven other classes: erratum, warning of, ratification, notice, contract, publication in the official gazette, addition, and others). Figure 5 presents the distribution of documents across the classes, providing an overview of the majority and minority classes across all crawled documents. Public Notice accounts for 46.2% of the documents, while the class of Other Minutes represents only 1.5%.

In a complementary way, Figure 5 shows the distribution of documents through the 13 classes for the crawled cities. The majority class is Public Notice in most cities. Examples of these cities are: Palma with 78% of the documents, Arantina (88%), Governador Valadares (99%), Pedro Teixeira (99%), Contagem (99%), Bias Fortes (63%), Ribeirão Vermelho (92%), São Bento Abade (58%) and Ijaci (61%). Other is also a majority class present in most of the cities, such as Palma (13%), Olaria (50%), Arantina (12%), Rio Preto (50%), São Tomé (55%), Cana Verde (26%), Coqueiral (13%), Cristais (19%), Pirapetinga (14%) and Passa-Vinte

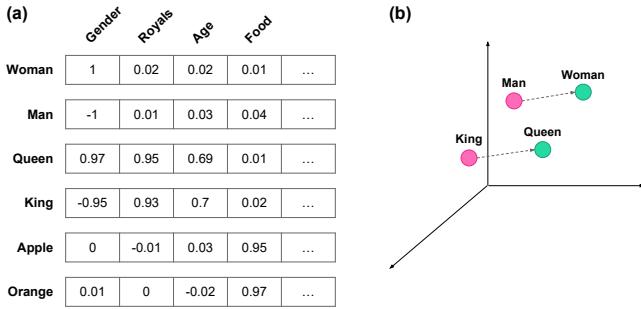


Figure 6. Word Embeddings: (a) Example of vector representation; and (b) Example of representation in vector space.

(18%). In addition, the other classes are the minority because few documents represent them. These classes are ratification, erratum, warning of, addition, contract, homologation, publication in the official gazette, minute of price registration, minute of bidding waiver, minute physical procurement, and other minutes.

It is important to note that, for the classification of Brazilian governmental documents, the documents from the municipality of Palma were excluded as they are often a compilation of different document categories. Thus, the experimentation considers only 17 cities.

5.2 Experimental Setup

For the classification of Brazilian public bids using LiPSet, we established an experimental setup that considers the text’s structural aspects, including the order of appearance of terms within the text. These characteristics can be effectively leveraged by implementing a Recurrent Neural Network (RNN), which is distinct from traditional networks in its ability to retain information over time during its execution. Among the various RNN architectures available in the literature, we use LSTM (Long Short-Term Memory), a specialized RNN that retains information for longer periods [Houdt *et al.*, 2020]. In an LSTM, input data is processed through multiple hidden layers, with the network retaining relevant information and discarding insignificant data.

To implement this model for text classification, we exploit the LSTM’s capability to preserve temporal information, or in the case of textual data, the sequential order of words. Among the different forms of representation explored in literature, word embeddings have emerged as a popular choice. Word embeddings represent words as vectors, wherein words with similar meanings exhibit similar vector representations, as illustrated in Figure 6a. Consequently, words used in similar contexts are positioned closely to one another in the vector space, as depicted in Figure 6b.

In this work, we employed a pre-trained embedding developed and publicly released by a research team from the Public Ministry of the State of Paraná (MPPR) [Noguti *et al.*, 2020]. The MPPR team’s research focused on analyzing various text classification algorithms to devise a solution to identify the domain area of public petitions received by the agency. This problem shares similarities with our work, such as the non-standardization of the texts under study and the objective of assigning the appropriate class corresponding to the document type associated with a bid. Hence, the embedding developed by the MPPR study proves highly valuable

Table 4. Experimental results with LiPSet.

Number of Documents	F1-Score	F1-Weighted
5.265 documents of 10 cities	0.953	0.965
6.245 documents of 17 cities	0.953	0.975

for our research.

To conduct the experiments, we stratified the data, ensuring that the distribution of document classes was preserved. Specifically, the training set consisted of 70% of the documents, the validation set encompassed 20%, and the testing set was allocated 10% of the documents.

5.3 Results

To evaluate the LSTM model’s performance, we conducted two experiments, considering ten cities (specifically, data from July, as obtained by the crawler) and all 17 crawled cities. The classification results for these two experiments are summarized in Table 4. The evaluation metrics employed are F1-Score and F1-Weighted. The results obtained for classifying public bid documents using LiPSet are highly promising, with an F1-Score of 0.953. These findings indicate the effectiveness of the LSTM model in accurately classifying the documents. Notably, increasing the number of documents did not significantly improve the F1-Score, as evidenced by the comparable performance between the experiments involving ten cities and all 17 cities.

The successful application of LiPSet in classifying public bid documents demonstrates its potential in this domain. Furthermore, given the large number of documents and cities available, various combinations can be explored to analyze the classification model further. For instance, it is possible to evaluate the model’s generalization by excluding the documents from a particular city from the training set and using them solely for testing purposes. Such analyses can provide deeper insights into the model’s performance and ability to handle diverse document sets.

6 Other Applications

In addition to its primary application in document classification, the LiPSet dataset holds great potential for various other contexts. The labeled nature of the dataset enhances its versatility, enabling the exploration of different applications. Some potential applications are described as follows.

In-depth analysis of specific types of documents. With knowledge of the different types of documents in the municipalities, it becomes possible to develop specialized workflows for extracting relevant information from each document type. This allows for more accurate and efficient extraction processes, as each document type has its own distinct data and structure. By tailoring the extraction process to specific document types, organizations can obtain more precise insights and streamline their information gathering.

Public expenditure analysis. The information extracted through specialized workflows, as demonstrated in this paper, can be utilized to build comprehensive historical databases of public agencies in an automated and efficient

manner. These databases can serve as valuable resources for analyzing public expenditure patterns and tracking price changes in products and services.

Fraud detection. The specialized workflows developed for information extraction from LiPSet can be instrumental in implementing audit trails and generating fraud alerts in public tenders, following the framework proposed in [Costa *et al.*, 2022]. In addition, Velasco *et al.* [2021] propose a methodology that uses data mining algorithms to detect corruption patterns, helping to identify fraud. Anowar and Sadaoui [2019] developed a fraud classifier that distinguishes between legitimate and non-legitimate bidders, using supervised machine learning algorithms. Some studies have also explored methodologies based on neural networks for detecting fraud [Pereira and Murai, 2021; Abidi *et al.*, 2021]. These works use data mining or machine learning techniques, and have methodologies strongly dependent on data. Therefore, a document classifier trained on LiPSet could help extract new bidding data and improve the aforementioned methodologies.

7 Challenges and Limitations

LiPSet, despite its usefulness, does have certain limitations that can be addressed in future research. These limitations are primarily associated with the challenges of working with diverse documents often made available without standardization on municipal transparency portals. The main challenges and limitations are listed as follows.

Processing documents in PDF format only. LiPSet currently includes data from documents exclusively in PDF format. Scanned documents and those containing only images are not included in the dataset. Since some municipalities may have many scanned documents, depending on the research focus, it may be necessary to consider them for specific studies. Future efforts can explore techniques to handle scanned documents and extract information effectively.

Lack of standardization. LiPSet contains highly diverse documents with minimal or no standardization. This lack of standardization can impact the performance of applications using the dataset as input. To address this limitation, it is suggested to group documents not only by meta-class but also based on their similarity. Implementing an approach that calculates document similarity can facilitate grouping and enhance the dataset’s quality.

Unbalanced data. Due to the wide range of document types and their variations, detailed categorization was challenging for many documents. As a result, the meta-class “Others” contains most documents. This class imbalance can pose challenges in terms of data representativeness and can impact classification results. Therefore, researchers aiming to consider all bidding or bidding-related documents should carefully analyze the files within the “Others” meta-class to

understand them better. Additionally, there is also an imbalance in the distribution of documents at the class level, which directly affects the performance of document classification. Future work should address these imbalances to ensure a more balanced dataset representation.

Limited number of municipalities. LiPSet currently includes documents from only 18 municipalities in Minas Gerais, as it was the focus of the research that used this dataset. Although the dataset coverage is relatively small compared to the number of Brazilian municipalities, the methodology presented in this work for building and labeling the dataset can be applied to expand LiPSet or construct new datasets using documents from other municipalities. Efforts to include a more diverse range of municipalities will enhance the dataset’s applicability and broaden its potential applications.

8 Conclusion

In this work, we introduced LiPSet, a dataset of Brazilian governmental documents, specifically public bidding documents. LiPSet was built by collecting and processing documents from the Transparency Portal of 18 municipalities in Minas Gerais. The dataset was carefully labeled to enable its application in various contexts. Our characterization of LiPSet revealed an imbalance between the number of documents across different meta-classes and municipalities. We also compared the results of a heuristic meta-classifier with manual labeling, demonstrating promising performance with a minimum hit rate of 77%.

We showcased a real application of LiPSet by focusing on the classification of public bid documents. We defined 13 classes based on document types and labeled the documents accordingly. Subsequently, we designed an experimental setup and employed a recurrent neural network (RNN) model for classification. The results demonstrated that LiPSet can effectively train a model to achieve acceptable classification performance, with an F1-Score of 0.953. Moreover, we discussed the potential applications of LiPSet beyond document classification, identified the challenges and limitations associated with the dataset, and highlighted opportunities for future research.

As future work, we plan to expand LiPSet by incorporating documents from additional municipalities in Minas Gerais to broaden its coverage. We also aim to employ similarity algorithms to group documents, facilitating their utilization. Additionally, we plan to explore alternative experimental setups for classifying bid documents using LiPSet.

Acknowledgements

The authors thank Amanda F. Paula, Iago A. D. Vaz, and Arthur P. G. Reis, the collaborators on this work.

Funding

This work was funded by the Prosecution Service of State of Minas Gerais (in Portuguese, *Ministério Público do Estado de Mi-*

nas Gerais, or simply MPMG) through its Analytical Capabilities Project and by CNPq, CAPES, FAPEMIG and the partnership project between AWS and CNPq.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in a repository on Zenodo.⁵

References

- Abidi, W. U. H., Daoud, M. S., Ihnaini, B., Khan, M. A., Alyas, T., Fatima, A., and Ahmad, M. (2021). Real-Time Shill Bidding Fraud Detection Empowered With Fussed Machine Learning. *IEEE Access*, 9:113612–113621. DOI: 10.1109/ACCESS.2021.3098628.
- Anowar, F. and Sadaoui, S. (2019). Multi-class Ensemble Learning of Imbalanced Bidding Fraud Data. In *Advances in Artificial Intelligence - 32nd Canadian Conference on Artificial Intelligence, Canadian AI*, volume 11489 of *Lecture Notes in Computer Science*, pages 352–358. Springer. DOI: 10.1007/978-3-030-18305-9_29.
- Araújo, L. R. and Souza, J. F. (2011). Aumentando a transparência do governo por meio da transformação de dados governamentais abertos em dados ligados. *Revista Eletrônica de Sistemas de Informação*, 10(1).
- Carneiro, M. G., Cupertino, T. H., Zhao, L., and Rosa, J. L. (2017). Semi-supervised semantic role labeling for Brazilian Portuguese. *Journal of Information and Data Management*, 8(2):117–117.
- Clarindo, J. P., Fontes, W., and Coutinho, F. (2019). QualisUS: um dataset sobre dados da saúde pública no brasil. In *Anais do II Dataset Showcase Workshop, DSW*, pages 418–428. SBC.
- Coelho, G. M. C., Ramos, A. C., de Sousa, J., Cavaliere, M., de Lima, M. J., Mangeth, A., Frajhof, I. Z., Cury, C., and Casanova, M. A. (2022). Text Classification in the Brazilian Legal Domain. In *Proceedings of the 24th International Conference on Enterprise Information Systems, ICEIS*, pages 355–363. SCITEPRESS. DOI: 10.5220/0011062000003179.
- Costa, L., Reis, A., Bacha, C. A., Oliveira, G. P., Silva, M. O., Teixeira, M. C., Brandão, M. A., Lacerda, A., and Pappa, G. (2022). Alertas de fraude em licitações: Uma abordagem baseada em redes sociais. In *Anais do XI Brazilian Workshop on Social Network Analysis and Mining, BraSNAM*, pages 37–48. SBC. DOI: 10.5753/brasnam.2022.223175.
- da Silva, L. C., Junior, R. d. V. C., de Araújo Lopes, H., and dos Santos, M. (2020). Utilização de técnicas de Mineração de Dados para detectar possíveis relacionamentos entre empresas participantes de licitações nas Forças Armadas. *Acanto em Revista*, 7(7):85–85.
- de Oliveira, E. F. and Silveira, M. S. (2018). Open government data in brazil a systematic review of its uses and issues. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age, DG.O*, pages 60:1–60:9. ACM. DOI: 10.1145/3209281.3209339.
- Erven, G. C. G. V., Holanda, M., and Carvalho, R. N. (2017). Detecting Evidence of Fraud in the Brazilian Government Using Graph Databases. In *Recent Advances in Information Systems and Technologies - Volume 2, WorldCIST*, volume 570 of *Advances in Intelligent Systems and Computing*, pages 464–473. Springer. DOI: 10.1007/978-3-319-56538-5_47.
- Gabardo, A. C. and Lopes, H. S. (2014). Using Social Network Analysis to Unveil Cartels in Public Bids. In *2014 European Network Intelligence Conference, ENIC*, pages 17–21. IEEE Computer Society. DOI: 10.1109/ENIC.2014.11.
- Houdt, G. V., Mosquera, C., and Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, 53(8):5929–5955. DOI: 10.1007/S10462-020-09838-1.
- Lima, M. C., Silva, R., de Souza Mendes, F. L., de Carvalho, L. R., Araújo, A. P. F., and de Barros Vidal, F. (2020). Inferring about fraudulent collusion risk on Brazilian public works contracts in official texts using a Bi-LSTM approach. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1580–1588. Association for Computational Linguistics. DOI: 10.18653/V1/2020.FINDINGS-EMNLP.143.
- Lyra, M. S., Curado, A., Damásio, B., Bação, F., and Pinheiro, F. L. (2021). Characterization of the firm–firm public procurement co-bidding network from the State of Ceará (Brazil) municipalities. *Applied Network Science*, 6(1):1–10.
- Mata, W. R. R. d., Boechat, D. S., and Brandão, M. A. (2019). JusBD: Um Banco de Dados para Obtenção de Informações do Poder Judiciário. In *Anais do II Dataset Showcase Workshop, DSW*, pages 398–407. SBC.
- Meera, S. and Geerthik, S. (2022). Natural Language Processing. *Artificial Intelligent Techniques for Wireless Communication and Networking*, pages 139–153. DOI: <https://doi.org/10.1002/9781119821809.ch10>.
- Nai, R., Sulis, E., and Meo, R. (2022). Public Procurement Fraud Detection and Artificial Intelligence Techniques: a Literature Review. In *Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management, EKAW-C*, volume 3256 of *CEUR Workshop Proceedings*. CEUR-WS.org.
- Noguti, M. Y., Vellasques, E., and Oliveira, L. S. (2020). Legal Document Classification: An Application to Law Area Prediction of Petitions to Public Prosecution Service. In *2020 International Joint Conference on Neural Networks, IEEE IJCNN*, pages 1–8. IEEE. DOI: 10.1109/IJCNN48605.2020.9207211.
- Pedrosa, J. A. O., Oliveira, D. M., Meira Jr., W., and P. Ribeiro, A. L. (2021). Automated classification of cardiology diagnoses based on textual medical reports. *Jour-*

⁵LiPSet: <https://doi.org/10.5281/zenodo.6974237>

- Journal of Information and Data Management*, 12(1). DOI: 10.5753/jidm.2021.1940.
- Pereira, L. S. (2022). Caracterização da comunidade que utiliza dados abertos governamentais sobre a educação brasileira. Master's thesis, Universidade Federal de Campina Grande, Campina Grande, Brasil.
- Pereira, R. and Murai, F. (2021). Quão efetivas são Redes Neurais baseadas em Grafos na Detecção de Fraude para Dados em Rede? In *Anais do X Brazilian Workshop on Social Network Analysis and Mining, BraSNAM*, pages 205–210. SBC. DOI: 10.5753/brasnam.2021.16141.
- Shimron, E., Tamir, J. I., Wang, K., and Lustig, M. (2022). Implicit data crimes: Machine learning bias arising from misuse of public data. *Proceedings of the National Academy of Sciences*, 119(13):e2117203119. DOI: 10.1073/pnas.2117203119.
- Silva, M. O., Paula, A. F., Oliveira, G. P., Vaz, I. A., Hott, H., Gomide, L. D., Reis, A. P., Mendes, B. M., Bacha, C. A., Costa, L. L., *et al.* (2022). Lipset: Um conjunto de dados com documentos rotulados de licitações públicas. In *Anais do IV Dataset Showcase Workshop, DSW*, pages 13–24. SBC. DOI: 10.5753/dsw.2022.224925.
- Velasco, R. B., Carpanese, I., Interian, R., Neto, O. C. G. P., and Ribeiro, C. C. (2021). A decision support system for fraud detection in public procurement. *International Transactions in Operational Research*, 28(1):27–47. DOI: 10.1111/itor.12811.