Journal of Information and Data Management, 2025, 16:1, doi: 10.5753/jidm.2025.4344

Evaluating Preprocessing and Textual Representation on Brazilian Public Bidding Document Classification

Michele A. Brandão [**Instituto Federal de Minas Gerais** | ***michele.brandao@ifmg.edu.br***]
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***]
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. Antônio Carlos, 6627, Pampulha, Belo Horizonte, MG, 31270-010, Brazil.

Received: 3 April 2024 • **Published:** 20 January 2025

In this paper, we tackle the task of classifying public bidding documents, which holds significant importance for both public and private entities seeking precise insights into bidding processes. Our study evaluates the impact of various preprocessing techniques and textual representation models, particularly word embeddings, on the accuracy of document classification. Overall, our results reveal while preprocessing techniques have minimal influence on classification outcomes, the choice of textual representation model significantly affects the representativeness of document classes. Moreover, we perform a qualitative analysis of misclassification cases, providing valuable insights into potential areas for improvement in document classification methodologies. Our findings underscore the importance of selecting appropriate textual representation models to enhance the accuracy and efficiency of document classification systems.

Keywords: Public Bids, Digital Government, Document Classification, Text Preprocessing, Text Representation

1 Introduction

The widespread adoption of open government data policies worldwide aims to enhance transparency and accountability within public institutions. In Brazil, the Access to Information Law (Law No. 12,527, of November 18, 2011)[1] stands as a pivotal democratic milestone, empowering society with greater involvement in government actions. However, managing this large volume of government data poses numerous challenges, including the diversity and complexity of data sources. The constant influx of new information highlights the need to use automated approaches to deal with such data.

Among such challenges, classifying public bidding documents holds significant importance for both government entities and private companies seeking accurate insights into bidding processes, which are fundamental in Brazil's public administration. Bidding processes involve soliciting bids from private companies to provide goods or services to the government and are governed by strict regulations outlined in laws such as the Public Bidding Law (Law No. 14,133/21).[2]

Bidding documents cover a range of materials, including public notices, minutes, and contracts, which must be accurately classified to ensure compliance with legal requirements and fair competition among bidders. However, manual analysis of large volumes of data can be time-consuming and subject to human error [Oliveira *et al*., 2022; P. Oliveira *et al*., 2023]. Faced with such challenges, adopting classification algorithms becomes increasingly important, streamlining the process while enhancing accuracy and scalability.

However, the efficacy of classification algorithms relies heavily on the proficiency of preprocessing and text representation. Preprocessing usually involves text standardization and vocabulary reduction, optimizing data representation, and mitigating sparsity. In turn, text representation converts text into a format suitable for classification algorithms, typically numeric vectors. Both steps help algorithms capture complex nuances and contextual information within documents, thus improving classification accuracy.

The unstructured nature of textual data can make the classification task even more challenging, with common words, technical jargon, and linguistic variations introducing ambiguities and different interpretations. Consequently, employing sophisticated approaches such as neural networks becomes imperative, as they can capture subtle nuances and complex word relationships, thereby enhancing the efficacy of document classification.

This paper evaluates the impact of different preprocessing techniques and text representation models, specifically word embeddings, on classifying public bidding documents using artificial neural networks. This work extends the paper presented on the 38th Brazilian Symposium on Databases [Brandão *et al*., 2023]. As a new contribution, we conduct a qualitative analysis to investigate misclassification cases resulting from the best experimental configuration. Our main contributions are summarized as follows.

1. We present a detailed methodology for classifying bidding documents, which can be easily adapted to other

[1] Access to Information Law: `http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm`

[2] Public Bidding Law: `https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm`

domains and classification tasks.
2. We propose a comprehensive evaluation of different preprocessing techniques and text representation models to classify public bidding documents.
3. We provide experimental results using real data, which allow us to assess the performance of different classification approaches in practical scenarios.
4. We perform a qualitative investigation into misclassification cases, providing insights into potential areas of improvement and further enhancing the classification.

The remainder of this paper is organized as follows. Section 2 provides an overview of related work, highlighting existing literature on document classification and preprocessing techniques. In Section 3, we delve into the methodology, detailing the steps involved in preprocessing, text representation, and classification of public bidding documents. Section 4 presents our experimental evaluation, including the experimental setup and results analysis. In Section 6, we discuss the findings and implications of our study, including the qualitative analysis of misclassification cases. Finally, Section 7 concludes the paper, summarizing the main contributions and suggesting avenues for future research in document classification and preprocessing techniques.

2 Related Work

Document classification in the legal domain poses significant challenges due to the intricate vocabulary and technical terminology found in legal texts, particularly in the Portuguese language, where available datasets are limited [Bambroo and Awasthi, 2021; Luz de Araujo *et al*., 2023]. In this section, we discuss related work in two main contexts: the classification of legal documents (Section 2.1) and the exploration of preprocessing and text representation techniques for document classification tasks (Section 2.2).

2.1 Legal Documents

Several initiatives in the literature aim to tackle the challenges of legal document classification [Martins and Silva, 2023]. For example, the VICTOR project [Luz de Araujo *et al*., 2020] offers a labeled dataset of documents from the Brazilian Supreme Court, supporting document classification by type and multi-label classification tasks. LiPSet [Silva *et al*., 2022, 2024], on the other hand, focuses on public bidding documents from 16 municipalities in Minas Gerais, Brazil, providing a structured dataset sourced from municipal transparency portals.

Another notable contribution is presented by Lima *et al*. [2020], who propose a new methodology for detecting fraud in public procurements using recurrent neural networks. For this purpose, the authors build a public procurement dataset from documents published in the Brazilian Official Gazette. In addition to the contribution of the new dataset, the proposed classification model achieves competitive results regarding precision, recall and F1 metrics compared to other state-of-the-art models, indicating the effectiveness of using deep learning models for such a task.

Moreover, Aguiar *et al*. [2021] investigate different text classification methods and different combinations of embeddings. Similarly, Coelho *et al*. [2022] tackle the classification of moral damage values in legal opinions, employing preprocessing techniques and word embeddings to train classification models. Their results indicate that models based on word embeddings outperform baselines that use TF-IDF to generate attributes.

2.2 Preprocessing and Text Representation

Numerous studies explore the effects of preprocessing and text representation techniques on classification tasks. For instance, Noguti *et al*. [2020] compare textual representation approaches for categorizing service descriptions by the Prosecution Office of Paraná, Brazil. In turn, Muniz Belém *et al*. [2023] investigate specialized preprocessing steps to enhance named entity detection and relationship extraction.

Considering both representation and preprocessing impacts, Albalawi *et al*. [2021] investigate the effects of preprocessing on health-related Arabic texts, using different preprocessing techniques and word embeddings. Their findings reveal that only four of the 26 preprocessing techniques significantly impact the performance of the evaluated classifier models. Furthermore, the use of textual normalization techniques specific to the language of the problem proved to be more effective. Models based on deep learning achieved superior results than traditional models, regardless of word embeddings and preprocessing configuration.

Similarly, Souza Júnior *et al*. [2022] assess different preprocessing methodologies for topic modeling in Brazilian Portuguese. They evaluate three document representation models, including two novel proposals based on the *CluWords* model adapted to Portuguese. While increasing preprocessing complexity has a positive, albeit minor, impact on TDF-IDF-based representation, the new proposals yield significantly improved coherence metrics. When combined with the preprocessing pipeline, these novel approaches achieve results approximately nine times better than the baseline model, representing the best performance reported in the literature for Brazilian Portuguese datasets.

Overall, the related work highlights the importance of considering text preprocessing for natural language models and the need to use specific techniques for textual representation of each language to perform a target task. Building upon this foundation, this work proposes a new approach that evaluates various preprocessing methods combined with word embedding models to classify public bidding documents in Brazilian Portuguese. By expanding on a previous work [Brandão *et al*., 2023], we incorporate a qualitative analysis of misclassifications and address the complexities of classifying long, unstructured texts in this domain.

3 Methodology

In this section, we present our methodology for classifying bidding documents. All documents are sourced from LiPSet [Silva *et al*., 2022, 2024], a dataset containing 9,083 manually labeled documents according to their type (e.g., notice,

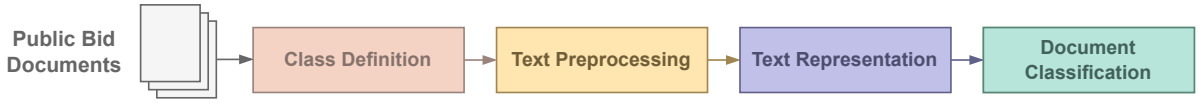

Figure 1. Methodology for public bidding document classification.

Table 1. Distribution of documents per meta-class and class.

Meta-class	Class	Documents	%
Minutes	price registration	2200	24.2%
	minutes of waiver	181	2.0%
	face-to-face auction	160	1.8%
	others	145	1.6%
Public Notice	public notice	3589	39.5%
Homo./Adj.	homologation/adjudication	408	4.5%
Others	others	1114	12.3%
	contract	451	5.0%
	notice	338	3.7%
	amendment	211	2.3%
	ratification	176	1.9%
	erratum	110	1.2%
TOTAL		**9083**	**100%**

minutes, notice). Since the actual class of most documents is present in the file title, the manual labeling process was performed by checking the titles of each document.

Figure 1 presents the four main steps of the methodology for classifying such bidding documents. From the raw dataset, we start by defining our target classes in Section 3.1. Then, Sections 3.2 and 3.3 detail the preprocessing and text representation stages performed over the documents' text. Specifically, we evaluate four preprocessing approaches and three distinct word embeddings in the classification task. Finally, Section 3.4 presents our LSTM classification model.

3.1 Class Definition

In its original version, LiPSet comprises 56 types of documents divided into four main meta-classes: minutes, public notice, adjudication/homologation, and others. However, some document types are very similar to each other (i.e., amendment and notice amendment), justifying the merging of such types into more representative classes. Therefore, we perform a new manual analysis of the documents by grouping similar documents. In short, such a process considers the following criteria: the meta-class, as it allows an appropriate separation of the documents; the number of documents collected, as it is not feasible to train a classifier with few documents representing it; and the importance of this document in a bidding process.

Overall, the new manual analysis resulted in 12 classes of bidding documents (Table 1). Following its hierarchical definition, each class still belongs to a meta-class. In short, our classes by meta-class are (i) Minutes: price registration, minutes of waiver, face-to-face auction, other minutes; (ii) Public notice: public notice; (iii) Homologation/Adjudication: homologation/adjudication; (iv) Others: contract, notice, amendment, erratum, ratification, others. Next, we provide a brief description of each class.

Price registration. Documents with prices, suppliers, supply conditions, and participating bodies, in accordance with the provisions of the notice and the winning bid proposals.

Minutes of waiver. Documents from a purchase method in which the use of bidding is not required (in accordance with Law No. 14,133/2021).[3]

Face-to-face auction. Documents related to a bidding modality realized in a public session, involving evaluating proposals and bids to determine the classification and qualification of the bidder offering the lowest price proposal.

Other minutes. Other types of minutes (e.g., cancellation and judgment minutes).

Public notice. Documents with all the criteria for judging a bid and its other information in complete form. It generally contains all the rules for the Public Administration to purchase products or contract services.

Homologation/adjudication. Homologation is related to the approval of the bidding procedure by the administrative authority. Next, adjudication is the formal act in which the Public Administration officially awards the contract to the winning bidder and invites them to sign the agreement.

Contract. It represents the agreement between governmental bodies and individuals, wherein both parties commit to forming a bond and outlining reciprocal obligations.

Notice. It serves as a promotional tool for the notice, providing essential details about the bidding process, including information on the object, modality, evaluation criteria, date, time, and venue of the public session, among others.

Amendment. Documents that add information to other already published documents (e.g., a document that adds new conditions to public notices).

Erratum. Documents with corrections to already published documents.

Ratfication. Similar to homologation, it is the act in which the higher authority validates waiver and unenforceability.

Others. Other types of documents that do not belong to any other class (e.g., orders and manuals).

3.2 Text Preprocessing

Preprocessing is a crucial initial phase in adequately representing documents, ensuring that the input data for classification remains relevant and thorough. It typically involves standardizing the text by removing accents and converting it to lowercase, among other techniques. Additionally, reducing the vocabulary by identifying and eliminating irrelevant terms, such as stopwords, can make the data less sparse and easier for computational processing. Effectively using these

[3]Law No. 14,133/2021: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm

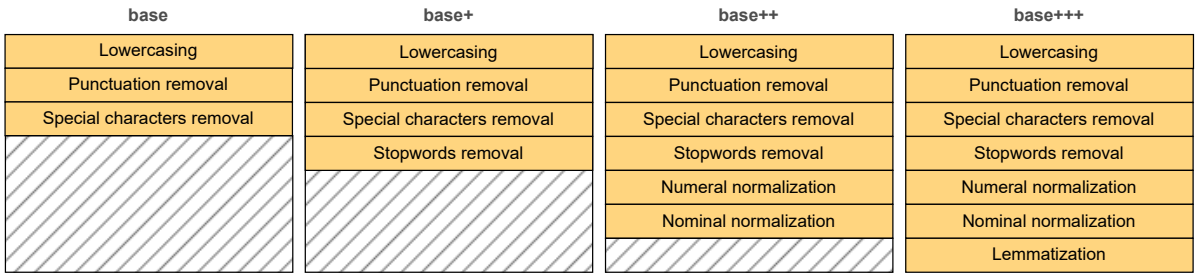

Figure 2. Overview of four preprocessing approaches.

preprocessing techniques can significantly improve the accuracy of automatically classifying bid documents.

This study evaluates four distinct preprocessing strategies, as detailed in Figure 2. Each strategy comprises a set of preprocessing techniques selected based on previous studies [Noguti *et al*., 2020] and preliminary studies of public bidding documents. The first strategy, labeled as `base`, consists of three operations: *lowercasing*, *punctuation removal* and *special character removal*. The subsequent three strategies build upon the `base`, with `base+` including an additional *stopwords removal* step, `base++` including *numeral and nominal normalization* steps, and `base+++` also including *lemmatization*. Each technique is described as follows.

Lowercasing. Converts all characters in the text to lowercase. Lowercasing is useful for reducing text variability, ensuring that identical words are not treated differently due to capitalization variations. It standardizes text data, minimizing redundancy, especially where words can appear in different cases (e.g., “ata” and “Ata”).

Punctuation removal. Removes all punctuation marks from the text, including commas, periods, colons, semicolons, and others. Such a technique simplifies textual data and decreases the number of unique words by discarding punctuation marks with no meaningful context.

Special characters removal. Removes non-alphanumeric characters from the text, such as hashtags, at signs, dollar signs, and other symbols that are not letters or numbers. It also simplifies text and reduces the number of unique words.

Stopwords removal. Removes common words from the text, such as articles (e.g., “the”, “a”), prepositions (e.g., “in”, “de”, “for”) and conjunctions (e.g., “and”, “or”). This technique aims to reduce noise in the data and improve the accuracy of subsequent analysis tasks by removing words that do not convey meaningful meaning or context. Here, we use a list of Brazilian Portuguese stopwords made available by the NLTK library.[4] Additionally, we remove city names from each document to prevent location information from overloading the classification model and degrading performance.

Numeral normalization. Converts all numerals in the text to a standard format. This may involve replacing digits with corresponding words (e.g., “7” becomes “seven”) or replacing all numeric values with a generic symbol (e.g., “1,000” becomes if “NUM”). This technique aims to reduce the variability of text data and simplify subsequent analysis tasks by treating all numeric values consistently. Here, following [Noguti *et al*., 2020], we replace all numerals with zero.

Nominal normalization. Converts proper nouns in the text to a standard format. The goal is to mitigate variations in naming conventions, such as abbreviations, spelling errors, or discrepancies. Normalizing these names can increase classification accuracy and ensure that relevant information is correctly identified. We employ a dictionary containing common Brazilian proper names, mapping all names to the term *proper_name*, following the Noguti *et al*. [2020] approach.

Lemmatization. Reduces words in the text to their base or dictionary form, known as a lemma. This process identifies the root form of a word and maps all inflected forms of that word to the same lemma (e.g., “walk”, “walked”, “walking” map to “walk”). The purpose of lemmatization is to reduce data variability and simplify subsequent parsing tasks by treating all inflected forms of a word as a single entity. Here, we use the spaCy library (for the Portuguese language)[5] to perform text lemmatization.

3.3 Text Representation

The next methodology step involves text representation, transforming textual data into a format conducive to machine learning algorithms. One widely adopted approach is word embeddings [Wang *et al*., 2019]. Word embeddings encode words into compact vector representations, capturing semantic relationships between them. This enables algorithms to understand the meaning and context of words in a document. We evaluate three distinct word embedding models, described as follows.

GloVe [Pennington *et al*., 2014]. Captures the relationships between words based on their co-occurrence probabilities across the entire corpus. It learns to map words into a continuous vector space, where the distance between vectors represents the semantic similarity between words. GloVe embeddings excel in capturing both semantic and syntactic relationships, making them suitable for a wide range of natural language processing tasks.

Word2Vec [Poetsch *et al*., 2019]. Operates on the principle of predicting the context of a word given its neighboring words (Skip-gram) or predicting a word based on its context (CBOW). Unlike GloVe, Word2Vec employs a neural network architecture to learn word embeddings, enabling it to capture nuanced semantic relationships between words. This

[4] NLTK: `https://www.nltk.org/howto/portuguese_en.html#stopwords`

[5] spaCy: `https://spacy.io/models/pt`

approach results in dense and contextually rich word representations, making Word2Vec a versatile choice for various natural language processing tasks.

Wang2Vec [Church, 2017]. Builds upon the architecture of Word2Vec, but novel sampling strategy and its focus on handling word boundaries in languages like Chinese more effectively. By sharing weights across different network parts, Wang2Vec reduces the computational overhead of training large-scale word embeddings.

3.4 Document Classification

In the domain of document classification, there exists a spectrum of algorithms, ranging from classical methods such as Naive Bayes and Decision Trees to more sophisticated neural network-based approaches Coelho *et al*. [2022]; Noguti *et al*. [2020]. While classical classifiers are often more straightforward, they may struggle to capture textual data's complex nuances and contextual information. Conversely, neural networks, particularly Long Short-Term Memory (LSTM) networks, excel in handling complex tasks such as classifying bidding documents due to their ability to effectively model sequential data and capture long-term dependencies.

LSTM networks, a subclass of recurrent neural networks, have demonstrated effectiveness in natural language processing tasks. Similar to other neural networks, LSTMs can incorporate multiple hidden layers. As data passes through these layers, relevant information is retained while irrelevant details are discarded in each cell. Consequently, LSTMs retain and prioritize important and pertinent information, disregarding irrelevant elements.

However, LSTM networks' efficacy depends on the quality of preprocessing and text representation techniques, as neural networks are sensitive to input data quality. Therefore, evaluating different preprocessing and text representation strategies is critical when using LSTM networks to classify bidding documents. Such evaluation helps identify the most effective techniques for the task at hand, thereby maximizing classification performance.

4 Experimental Setup

This section presents the technical details of the experimental setup and evaluation. Despite the imbalance present between the classes in our dataset (see Table 1), we do not perform any balancing strategy to keep the experiment more aligned with reality. For instance, public notices are more prevalent than amendments, as each public bid must have a public notice, whereas amendments are issued only when required.

In addition to the different preprocessing and text representation approaches, we evaluate two stratified training-test split strategies: (i) stratification by class and (ii) stratification by class and city. We choose such strategies to evaluate how stratification can impact bidding document classification performance. We also apply cross-validation in both strategies to ensure a robust assessment of the performance of the different setups. Next, we detail each strategy.

1. **Stratification by class.** Considers that classes of bidding documents may have different frequencies in the database, and the split is done in such a way that the proportion of each class is maintained in each fold;
2. **Stratification by class and city.** Considers that bidding documents from the same city may present similar characteristics, and, therefore, the split is made so that the proportion of each class and city is maintained in each fold.

Table 2. Summary of the experimental setup.

Category	Options
Preprocessing	`base` | `base+` | `base++` | `base+++`
Word Embedding	GloVe | Word2Vec | Wang2Vec
Stratification	class | class and city
Cross-validation	5-fold
Classifier	LSTM

Therefore, we perform 24 experiments, one for each combination of the two experimental configurations, four preprocessing approaches, and three word embedding models. Each experiment was run with a 5-fold cross-validation. We do not use more folds due to the high processing time. Table 2 shows the summary of the experimental setup.

Word embedding models. The word embedding models used in this work were sourced from NILC-Embeddings,[6] a repository dedicated to storing and sharing word embeddings for the Portuguese language. This repository contains an extensive collection of vectors derived from a diverse range of sources, capturing linguistic nuances across both Brazilian and European Portuguese. As previously stated, we employed three distinct models for our experimentation: GloVe, Word2Vec, and Wang2Vec. Each model generates word vectors in varying dimensions, with options including CBOW and Skip-Gram variations. In this work, we use GloVe, Word2Vec, and Wang2Vec with 600 dimensions, leveraging the Skip-Gram approach for our classification tasks.

LSTM Configuration. To build the classifier, we choose an LSTM network architecture with three recurrence layers. In addition, we add a dropout layer with 20% probability to avoid overfitting. We train the model using the Adam optimization technique, with an initial learning rate of 0.001 and a decay rate of 1e-6. We set the number of training epochs to 8 and the training batch size to 64.

Evaluation Metrics. To evaluate the experiments, we consider two metrics: F1-Macro and F1-Weighted. The former is the harmonic mean of the F1 scores for each class and is useful for evaluating the model's ability to handle imbalanced classes, whereas the latter is the harmonic mean of the F1 scores weighted by the number of samples in each class and is best suited for evaluating the overall accuracy of the model across all classes. Such metrics are calculated for each model evaluated in the public bidding dataset.

5 Experimental Results

This section presents the results obtained for the 24 experiments performed with LSTM, using a 5-fold cross-validation.

[6]NILC-Embeddings: `http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc`

Table 3. Comparison of the 24 experimental configurations in classifying bidding documents using LSTM. The best result for each stratification and metric is underlined.

		Stratification by class		Stratification by class and city	
Preprocessing	**Word Embedding**	**F1-Macro**	**F1-Weighted**	**F1-Macro**	**F1-Weighted**
base	Word2Vec	0.863 ± 0.203	0.955 ± 0.068	0.893 ± 0.064	0.971 ± 0.015
	Wang2Vec	0.954 ± 0.003	0.984 ± 0.001	0.942 ± 0.044	0.983 ± 0.010
	GloVe	0.950 ± 0.017	0.985 ± 0.004	0.908 ± 0.128	0.973 ± 0.031
base+	Word2Vec	0.957 ± 0.007	0.986 ± 0.002	0.953 ± 0.003	0.985 ± 0.002
	Wang2Vec	0.960 ± 0.002	0.986 ± 0.001	0.964 ± 0.002	0.987 ± 0.002
	GloVe	0.971 ± 0.012	0.989 ± 0.004	0.969 ± 0.005	0.989 ± 0.003
base++	Word2Vec	0.937 ± 0.016	0.981 ± 0.003	0.932 ± 0.009	0.977 ± 0.003
	Wang2Vec	0.943 ± 0.016	0.981 ± 0.005	0.929 ± 0.045	0.976 ± 0.016
	GloVe	0.960 ± 0.016	0.986 ± 0.005	0.954 ± 0.009	0.985 ± 0.004
base+++	Word2Vec	0.925 ± 0.037	0.979 ± 0.009	0.914 ± 0.041	0.976 ± 0.012
	Wang2Vec	0.928 ± 0.024	0.977 ± 0.010	0.946 ± 0.022	0.983 ± 0.004
	GloVe	0.939 ± 0.025	0.981 ± 0.009	0.963 ± 0.005	0.987 ± 0.001

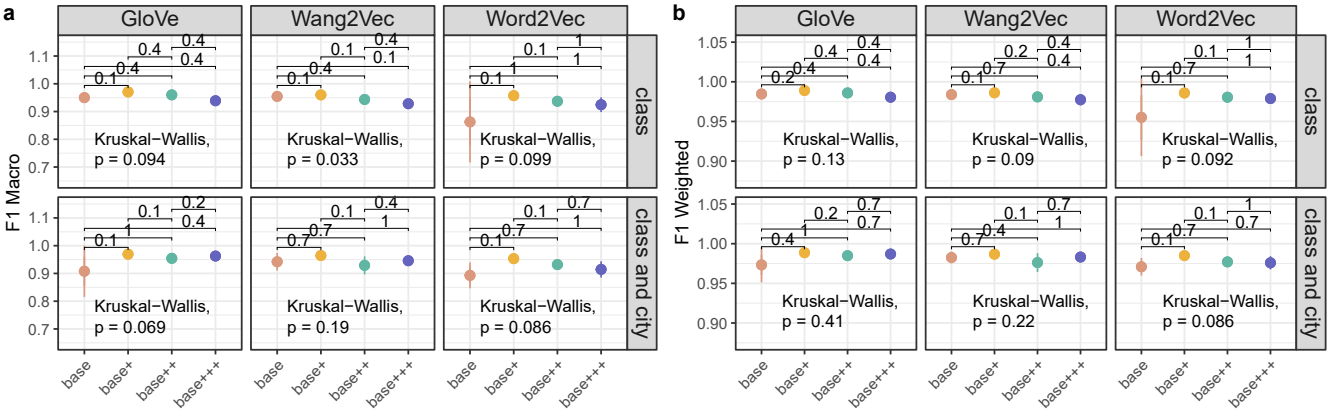

Figure 3. Result of the classification of experimental configurations according to (a) F1 Macro and (b) F1 Weighted. Statistical tests (Kruskal-Wallis and Wilcoxon paired) were used to calculate the significant difference between preprocessing approaches. The values between the midpoints represent the p-value resulting from the paired Wilcoxon test.

In particular, Table 3 shows that the results are quite similar for the different experimental combinations evaluated. The best results for F1-Macro (0.971) and F1-Weighted (0.989) were obtained for the experimental configuration with stratification by class, `base+` preprocessing, and text representation using the GloVe model. The worst results were obtained for the experimental configuration with stratification by class, `base` preprocessing, and text representation using the Word2Vec model, achieving an F1-Macro of 0.863.

At first, no clear predominance is observed among the various preprocessing techniques, as they yield varying results for each word embedding model. However, a notable differentiation is evident between the representation models. The Word2Vec model consistently yields the worst performance across all preprocessing configurations, showing lower F1-Macro and F1-Weighted scores compared to GloVe and Wang2Vec. Conversely, GloVe performs well across most preprocessing variations, demonstrating robustness in capturing semantic and syntactic relationships within the bidding documents. Such findings highlight the importance of choosing an appropriate word embedding model.

To assess whether there is a significant difference between the experimental configurations, Figures 3a and 3b present, respectively, the F1-Macro and F1-Weighted with Kruskal-Wallis and Wilcoxon paired test for each experimental setup. Both tests are non-parametric and are used to compare independent samples. The Wilcoxon paired test allows comparing just two samples, whereas the Kruskal-Wallis test allows comparing three or more samples [Kim, 2014]. The analysis of the p-value (p) of the Kruskal-Wallis test in Figures 3a and 3b reveals that it is not possible to reject the null hypothesis that the medians of the F1-Macro and F1-Weighted of each experiment are the same, as the p-value is greater than 0.05 (i.e., probability greater than 5%). Therefore, there is evidence that the difference observed between the experiments may be due to chance, meaning that there is no significant difference between the experiments.

An exception is the comparison between the preprocessing approaches using the Wang2Vec model with stratification by class for the F1-Macro evaluation metric, with a p-value of 0.033, that is, slightly less than 0.05. However, the F1-Weighted metric for this experimental configuration has a p-value of 0.09 (i.e., above the threshold). Therefore, we also consider that there is no significant difference between experiments with different preprocessing approaches in the experimental configuration of both stratifications.

Regarding the paired Wilcoxon test results, all p-values are greater than 0.05, indicating the absence of a significant difference between the experimental configurations when compared two by two. This result reinforces the Kruskal-Wallis test results, suggesting that the different combinations of preprocessing and text representation evaluated do not signifi-

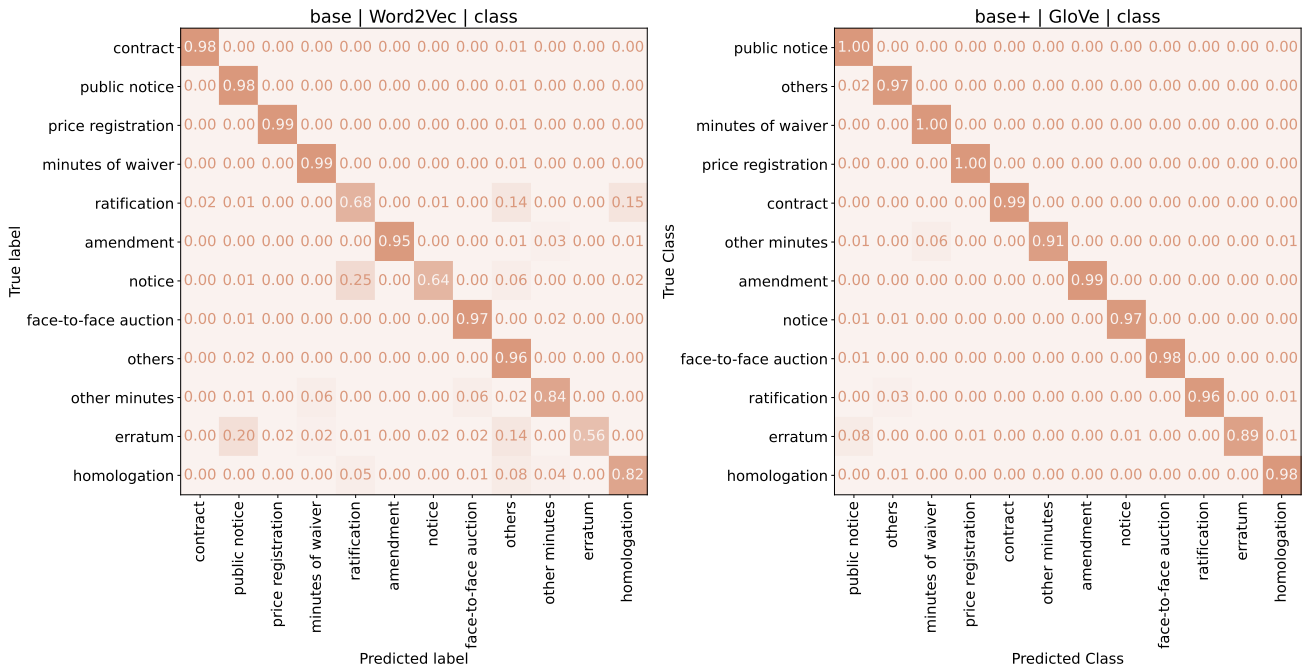

Figure 4. Confusion matrix of the worst and best experimental configuration.

cantly affect the performance of the LSTM network in the bidding document classification task.

From our knowledge of the problem, there are three possible reasons for explaining such results: (i) the nature of bidding documents, which are long texts in Portuguese with little standardization; (ii) the high number of classes that may not be well represented by the collected documents; and (iii) the usage of LSTM as our classification model, which, despite storing information from sequences of long texts, has limitations regarding the size of the input that the neural network uses to predict the next output [Zhang *et al*., 2018].

To deepen the understanding of the results, Figure 4 presents two confusion matrices considering the 12 classes defined for the classification of bidding documents. Figure 4a refers to the experiment with the lowest result for the F1-Macro metric, i.e., experiment with stratification only by class, `base` preprocessing, and text representation using Word2Vec. The results show that the most difficult class to classify is *erratum*, in which 20% of the documents are predicted as *public notice* and 14% as *others*. When comparing with the confusion matrix in Figure 4b, resulting from the experiment with the highest result for the F1-Macro metric (stratification by class, `base+` preprocessing and text representation with GloVe), *erratum* is still one of the classes with most errors, being mainly confused with *public notice* (8%).

Finally, the confusion matrix analysis indicates a problem in the representation of the *erratum* and *public notice* classes, as these are the classes in which the classifier makes the most mistakes in the different experiments. We do not present the confusion matrices for all 24 experiments due to space limitations and because they are similar. Therefore, it is crucial to better analyze how the texts are represented in the different classes to better represent the bidding documents, which is in line with the findings of Souza Júnior *et al*. [2022].

6 Qualitative Analysis

To delve into the misclassification cases resulting from the best experimental configuration (`base+`, GloVe and stratification by class), we perform a qualitative analysis focusing on cases where the F1-Macro score falls below 95% (Figure 4). Specifically, we investigate the classes *erratum* and *other minutes*, which achieved F1-Macros of 89% and 91%, respectively. Both classes show the highest confusion rates, indicating a possible overlap in their textual features.

As LSTM networks significantly rely on semantic and contextual information to achieve accurate classification, we investigate whether common terms between classes contribute to misclassification. To do so, we plot word clouds to visualize the most frequent terms within each document class, as shown in Figure 5. Here, we remove terms with fewer than three characters and numbers to ensure that the word clouds focus on meaningful textual features.

Overall, most document classes present similar terms, such as *municipal* (“municipal”), *proposal* (“proposta”), *bidder* (“licitante”), *public notice* (“edital”), and *city hall* (“prefeitura”). Such terms appear frequently across various classes, suggesting common themes prevalent in bidding documents. Such consistency underscores the challenge of distinguishing between classes based solely on individual terms and emphasizes the importance of capturing broader contextual cues for accurate classification.

Regarding the most confused class, *erratum*, the word clouds reveal notable overlaps with *public notice* and *notice* classes. Both classes are the ones most frequently confused with *erratum* by the classifier, particularly *public notice* (8%). This is probably because most errata must be corrections of public notices, as they share similar terms referencing municipal proceedings, city hall announcements, public notices, and procedural details. Thus, such similarity between errata and public notices may contribute to the classi-

Figure 5. Word clouds of each document class.

Table 4. Top and bottom 10 similarities among pairs of classes based on Jaccard similarity.

Top 10		
Class 1	**Class 2**	**Similarity**
Amendment	Contract	0.43
Erratum	Public Notice	0.36
Public Notice	Notice	0.35
Homologation	Ratification	0.35
Erratum	Notice	0.31
Erratum	Contract	0.30
Price registration	Contract	0.29
Minutes of waiver	Face-to-face auction	0.26
Homologation	Others	0.25
Erratum	Others	0.24
Bottom 10		
Class 1	**Class 2**	**Similarity**
Face-to-face auction	Contract	0.11
Price registration	Ratification	0.10
Price registration	Face-to-face auction	0.10
Face-to-face auction	Amendment	0.10
Public Notice	Ratification	0.09
Price registration	Minutes of waiver	0.09
Other minutes	Ratification	0.09
Other minutes	Amendment	0.09
Other minutes	Contract	0.09
Ratification	Face-to-face auction	0.09

fier's difficulty distinguishing between these classes.

Another significant confusion is between the *other minutes* and *minutes of waiver* classes (6%). These two classes show notable similarities in their word clouds, indicating overlapping content. The most frequent common terms include *value* ("valor"), *ltda*, *supplier* ("fornecedor"), indicating that the documents classified as other minutes may share characteristics with minutes of waiver documents. This similarity in terms may imply inadequate labeling for minute-related classes, contributing to ambiguity in classification and subsequent misclassification.

We also compute the Jaccard similarity between pairs of classes to enhance the qualitative analysis provided by the word clouds. Such a measure quantifies the overlap between classes by comparing the presence of the 100 most frequent terms in each class. Table 4 lists the top and bottom 10 similarities observed among pairs of classes. Notably, one of the most similar pairs of classes is *erratum* and *public notice*, indicating a significant overlap in their textual features.

However, while most pairs of similar classes correspond to those frequently confused by the classifier, there are some exceptions. For instance, the most similar pair of classes, *amendment* and *contract*, did not show confusion in the classifier's predictions, even with the worst experimental configuration. This suggests that similarity between classes does not always directly correlate with misclassification. Other factors, such as class imbalance or the classifier's ability to discern subtle differences between classes, may influence classification accuracy independently of class similarity.

In summary, the qualitative analysis of misclassification cases sheds light on the challenges faced in the document classification task. Despite the high overall performance of the LSTM model, certain classes show significant overlaps in their textual features, leading to confusion during classification. The word clouds and Jaccard similarity analysis provide valuable insights into the nature of these overlaps, highlighting areas where class definitions may need refinement or where the model's feature representation can be improved. These findings underscore the importance of continuous evaluation and refinement of classification models in complex textual domains such as bidding documents.

7 Conclusion

In this paper, we evaluated the impact of different preprocessing techniques and textual representation models on public bidding document classification accuracy. We proposed a methodology for classifying bidding documents, containing the class definition, text preprocessing, and representation steps. Specifically, we explored four distinct preprocessing approaches and employed three varied textual representation models. Such techniques were evaluated using a classifier based on an LSTM neural network.

Using a real-world dataset, we evaluated the impact of two

stratified training-test split strategies alongside various preprocessing and text representation approaches. Our findings indicate that while preprocessing techniques have minimal influence on classification outcomes, selecting a textual representation model notably impacts the representativeness of document classes. These results underscore the importance of choosing appropriate textual representation models to ensure the accurate classification of public bidding documents.

We also performed a qualitative analysis of misclassification cases, shedding light on the challenges faced in the document classification task. In summary, our findings revealed that similarity between classes does not consistently align with misclassification occurrences. Factors such as class imbalance or the classifier's capacity to discern subtle differences between classes may independently impact classification accuracy, irrespective of class similarity.

Limitations and Future Work. While our study provides valuable insights into classifying public bidding documents, some limitations warrant consideration. First, our analysis focused primarily on the impact of preprocessing techniques and textual representation models on classification accuracy. However, other factors, such as feature selection methods, hyperparameter tuning, and the choice of classification algorithms, could also influence classification performance and deserve further investigation.

Second, our study was constrained by the availability and quality of the dataset. While we used real-world data for our experiments, the dataset's size and scope may have limited the generalizability of our findings. Future work endeavors could explore larger and more diverse datasets to validate our results across different contexts and domains. Third, the experimental configurations presented in this work do not allow for evaluating whether the proposed classification model can be generalized to new bidding documents. Therefore, as future work, we plan to conduct an experimental setup to evaluate this generalization better.

Finally, considering the current state-of-the-art in classifying text documents, particularly with the advancements brought by models like BERT, it is imperative to explore the impact of transitioning from LSTM to BERT-based models. This investigation could offer valuable insights into the performance improvements and potential enhancements achieved by leveraging more advanced and sophisticated models in classifying public bidding documents, paving the way for future advancements in this domain.

Acknowledgements

The authors thank Henrique Hott, who collaborated on the previous version of this work.

Funding

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

References

Aguiar, A., Silveira, R., Pinheiro, V., Furtado, V., and Neto, J. A. (2021). Text classification in legal documents extracted from lawsuits in brazilian courts. In Britto, A. and Valdivia Delgado, K., editors, *Brazilian Conference on Intelligent Systems*, pages 586–600. Springer International Publishing.

Albalawi, Y., Buckley, J., and Nikolov, N. S. (2021). Investigating the impact of pre-processing techniques and pre-trained word embeddings in detecting arabic health information on social media. *J. Big Data*, 8(1):95. DOI: 10.1186/s40537-021-00488-w.

Bambroo, P. and Awasthi, A. (2021). Legaldb: long distilbert for legal document classification. In *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pages 1–4. IEEE.

Brandão, M., Silva, M., Oliveira, G., Hott, H., Lacerda, A., and Pappa, G. (2023). Impacto do pré-processamento e representação textual na classificação de documentos de licitações. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 102–114, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2023.231658.

Church, K. W. (2017). Word2vec. *Natural Language Engineering*, 23(1):155–162.

Coelho, G. M., 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 *ICEIS (1)*, pages 355–363.

Kim, H.-Y. (2014). Statistical notes for clinical researchers: Nonparametric statistical methods: 2. nonparametric methods for comparing three or more groups and repeated measures. *Restorative Dentistry & Endodontics*, 39(4):329–332.

Lima, M., Silva, R., Lopes de Souza Mendes, F., R. de Carvalho, L., Araujo, A., 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*, pages 1580–1588, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.findings-emnlp.143.

Luz de Araujo, P. H., de Almeida, A. P. G. S., Braz, F. A., da Silva, N. C., de Barros Vidal, F., and de Campos, T. E. (2023). Sequence-aware multimodal page classification of brazilian legal documents. *Int. J. Document Anal. Recognit.*, 26(1):33–49.

Luz de Araujo, P. H., de Campos, T. E., Ataides Braz, F., and Correia da Silva, N. (2020). VICTOR: a dataset for Brazilian legal documents classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1449–1458, Marseille, France. European Language Resources Association.

Martins, V. S. and Silva, C. D. (2023). Text classification in

law area: a systematic review. *Journal of Information and Data Management*, 13(6). DOI: 10.5753/jidm.2022.2547.

Muniz Belém, F., Valiense, C., França, C., Carvalho, M., Ganem, M., Teixeira, G., Jallais, G., H. F. Laender, A., and A. Gonçalves, M. (2023). Contextual reinforcement, entity delimitation and generative data augmentation for entity recognition and relation extraction in official documents. *Journal of Information and Data Management*, 14(1). DOI: 10.5753/jidm.2023.3180.

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, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020*, pages 1–8. IEEE. DOI: 10.1109/IJCNN48605.2020.9207211.

Oliveira, G. P., Reis, A. P. G., Mendes, B. M. A., Bacha, C. A., Costa, L. L., Canguçu, G. L., Silva, M. O., Caetano, V., Brandão, M. A., Lacerda, A., and Pappa, G. L. (2022). Ferramentas open-source de qualidade de dados para licitações públicas: Uma análise comparativa. In *SBBD*, pages 116–127. SBC.

P. Oliveira, G., M. A. Mendes, B., A. Bacha, C., L. Costa, L., D. Gomide, L., O. Silva, M., A. Brandão, M., Lacerda, A., and L. Pappa, G. (2023). Assessing data quality inconsistencies in brazilian governmental data. *Journal of Information and Data Management*, 14(1). DOI: 10.5753/jidm.2023.3220.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. ACL. DOI: 10.3115/v1/d14-1162.

Poetsch, M., Correa, U. B., and de Freitas, L. A. (2019). A word embedding analysis towards ontology enrichment. *Res. Comput. Sci.*, 148(11):153–164.

Silva, M. O., Oliveira, G. P., Hott, H., Gomide, L. D., Mendes, B. M. A., Bacha, C. A., Costa, L. L., Brandão, M. A., Lacerda, A., and Pappa, G. L. (2024). Lipset: A comprehensive dataset of labeled portuguese public bidding documents. *Journal of Information and Data Management*. to appear in.

Silva, M. O. *et al*. (2022). LiPSet: Um conjunto de Dados com Documentos Rotulados de Licitações Públicas. In *SBBD DSW*, pages 13–24, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/dsw.2022.224925.

Souza Júnior, A. P., Cecilio, P., Viegas, F., Cunha, W., de Albergaria, E. T., and da Rocha, L. C. D. (2022). Evaluating topic modeling pre-processing pipelines for portuguese texts. In *WebMedia*, pages 191–201. ACM.

Wang, S., Zhou, W., and Jiang, C. (2019). A survey of word embeddings based on deep learning. *Computing*, 102:717–740. DOI: 10.1007/s00607-019-00768-7.

Zhang, J., Li, Y., Tian, J., and Li, T. (2018). Lstm-cnn hybrid model for text classification. In *2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 1675–1680. IEEE.