



Overpricing Analysis in Brazilian Public Bidding Items

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

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

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


Guilherme Bezerra  [Universidade Federal de Minas Gerais | guilhermebezerra@dcc.ufmg.br]

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

 *Institute of Exact Sciences, Department of Computer Science, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Pampulha, Belo Horizonte, MG, 31270010, Brasil.*

Received: 08 November 2023 • **Accepted:** 15 January 2024 • **Published:** 18 January 2024

Abstract: Analyzing overpricing in public bidding items is essential for government agencies to detect signs of fraud in acquiring public goods and services. In this context, this paper presents two main contributions: a methodology for processing and standardizing bid item descriptions and a statistical approach for overpricing detection using the interquartile range. We evaluated a comparative analysis of three distinct grouping strategies, each emphasizing different facets of the item description standardization process. Furthermore, to gauge the efficacy of both proposed approaches, we leveraged a ground-truth dataset for a thorough evaluation containing quantitative and qualitative analyses. Overall, our findings suggest that the evaluated strategies are promising for identifying potential irregularities within public bidding processes.

Keywords: Overpricing, Public Bidding, Fraud, Methodology, Statistical, Purchase

1 Introduction

Bidding processes are procedures adopted by the public administration to secure fairness and the identification of optimal proposals when acquiring public goods or services [Green *et al.*, 1994; Brandão *et al.*, 2023]. During the bidding process, the bidder establishes a reference value that serves as a basis for determining the bidding category and, among other purposes, verifying whether the proposals submitted are advantageous. However, such price evaluation can be challenging, especially considering the diversity of items being bid and market price fluctuations.

To address such a problem, it is a common practice for government agencies to collect and analyze historical data from bids, including the prices quoted, to detect patterns and indications of overpricing or irregularities. Notable examples of such initiatives include the *Controladoria Geral da União* (CGU)¹ and the *Tribunal de Contas do Estado de Minas Gerais* (TCE/MG)² systems. Nevertheless, persistent challenges exist in standardizing textual inputs, unifying units of measurement for the traded items, and establishing clear thresholds for identifying instances of overpricing.

Such a problem can be approached using complex and simpler methodologies [Correa and Leal, 2018]. Complex approaches, such as machine learning models or time series analysis, offer the advantage of adaptability and predictive power [Matschak *et al.*, 2022]. Machine learning models can learn patterns in the data, making them suitable for detecting subtle irregularities [Lima *et al.*, 2020]. Time series analysis can capture temporal trends in pricing, which is

crucial for understanding how prices evolve [Xiao and Jiao, 2021]. However, these approaches often require a significant amount of labeled data. They may introduce complexity that could be challenging to interpret and implement in a real-world government setting.

On the other hand, simpler approaches, such as statistical models, offer a practical and transparent solution for addressing the challenge of overpricing detection in public bidding [Reis *et al.*, 2023]. In this work, we introduce a statistical approach that leverages easily interpretable statistical measures to establish a clear threshold for identifying overpricing and anomalies, providing straightforward guidelines for public administration agencies. Additionally, we propose a novel methodology for standardizing the item descriptions, ensuring consistency in item descriptions, and making price comparisons more accurate.

This work is an extended version of a previously published article [Silva *et al.*, 2023], in which we extended its content by using a ground-truth dataset to validate both proposed approaches through quantitative and qualitative analyses. Our main contributions are structured as follows:

- We introduce a comprehensive methodology to tackle the critical challenge of detecting overpricing in public bidding. This methodology encompasses two distinct but interrelated components. The first component focuses on pre-processing and standardizing item descriptions, addressing the inherent variability and inconsistencies in how items are described in bidding processes. The second component is centered around the statistical approach based on the Interquartile Range (IQR), enabling a systematic and interpretable method for overpricing detection;

¹<https://paineldepregos.planejamento.gov.br/>

²<https://bancodepreco.tce.mg.gov.br/>

- We conduct a comparative analysis of three different grouping strategies—*Original*, *FToken*, and *Words*—each emphasizing distinct aspects of the item description standardization process. Through such comparisons, we identify the strategy that offers the most promising results in accurately estimating overpricing;
- To assess the effectiveness of our proposed methodology, we leverage a ground-truth dataset, enabling a thorough evaluation through quantitative and qualitative analyses.

The current work is organized as follows. In Section 2, we explore related research that deals with similar challenges in text standardization and the identification of overpricing. In Sections 3 and 4, we provide insights into our dataset and the methodology we adopted for processing and standardizing it. In Section 5, we present the approach employed for overpricing detection, presenting our findings and offering a detailed analysis of case examples. In Section 6, we undertake a quantitative evaluation of both proposed approaches, namely, the statistical model and the text standardization method. Finally, in Section 7, we conclude our work and discuss future research directions.

2 Related Work

This work aims to detect overpricing in public purchasing and consequently identify fraud evidence. In this context, although public bidding documents adhere to a structured format, they often employ different terms and phrasings to describe identical items, posing a challenge for automated document processing, as reported in prior research [Silva et al., 2022; Oliveira et al., 2022]. To address this issue, specialized Natural Language Processing (NLP) techniques are usually employed for text data processing. The literature predominantly delves into two key areas: fraud detection and the textual analysis of public government documents.

2.1 Fraud Detection

Fraud detection in public bidding is a critical concern for government agencies and entities responsible for overseeing public bidding processes. Public bidding, which involves the acquisition of goods and services by government organizations, represents a substantial portion of public expenditure. It is, therefore, susceptible to various forms of fraud, including overpricing, collusion, and bid rigging [Costa et al., 2022; Brandão et al., 2023]. Detecting these fraudulent activities is essential to safeguard public funds and ensure fair and transparent bidding processes.

Researchers and organizations are actively working to enhance fraud detection and prevent fraudulent practices in the context of public bidding. For instance, Oliveira et al. [2022] describe a hierarchical decision-making approach incorporating data preprocessing to enhance textual structuring and standardization. Such a method categorizes data into three groups based on conformity and frequency rates, enabling the classification of purchased items based on transaction validity. The results reveal that combining readily available bidder data with extracting bidding item descriptions can significantly contribute to fraud detection.

Reis et al. [2023] conduct a comparative analysis of privacy policies to evaluate whether they align with the Brazilian General Data Protection Law [Brasil, 2018]. To do this, the researchers collected 82 privacy policies and segmented them into paragraphs, which were then labeled and subjected to statistical analyses and data visualizations, such as word clouds, to assess the text's content and readability. In conclusion, the article states that 40% of the segments contain vague statements to the public regarding how user data will be processed. Therefore, there is a need to adapt the text for the public and express data processing with clarity and transparency. As future work, the article suggests using machine learning models to extract information about data processing practices mentioned throughout the corpus.

Complementary, Pereira et al. [2022] introduce a graph-based modeling approach coupled with centrality metrics, followed by classification algorithms to distinguish companies as fraudulent or legitimate. Their proposed approach achieved a precision rate of more than 71% and an accuracy of 68%. Luna and Figueiredo [2022] implement a similar methodology, focusing on computing metrics to detect potential fraud indicators without needing a classification algorithm. Their results indicate the complexity of the bidding process, revealing substantial variations among companies and public agencies in various aspects, such as the entry and exit degrees within the company network and the bid values.

Gabardo and Lopes [2014] detail a strategy for identifying cartels among construction companies, using social network analysis and complex networks to represent the relationship between companies participating in public bids. They found several groups of companies whose composition and actions suggest the formation of cartels. More recently, Lima et al. [2020] introduce a new dataset formed by public bidding available on the Brazilian Official Gazette (*Diário Oficial da União*), using 15,132,968 textual entries, of which 1,907 are annotated risky entries. The study employed bottleneck deep neural networks and Bi-LSTM models, which exhibited strong competitiveness compared to classical classifiers, achieving precision rates of 93.0% and 92.4%, respectively.

Regarding overpricing detection, Correa and Leal [2018] apply ontology-based text mining and clustering techniques to unveil overpricing instances within the products procured by the federal government of Brazil, focusing specifically on medications procured by the Ministry of Health. The authors use such techniques to create a consolidated price base for each medication, which can be used as a benchmark against which current prices are compared. Any substantial distortions in the prices practiced are flagged, allowing for identifying potential irregularities that warrant further examination and may reveal fraudulent activities.

In this work, we do not propose a strategy to detect fraud directly but to detect overpricing and anomalies that may indicate fraud in public bids. Thus, this work is similar to the others, as Luna and Figueiredo [2022]; Oliveira et al. [2022]; Pereira et al. [2022], by extracting information from data that can indicate fraud.

2.2 Government Documents Processing

Government agencies, at different levels, generate and handle extensive documentation related to public policies, regulations, bidding processes, and more. Such documents contain various formats, including legal texts, reports, contracts, and public bidding records. Managing and processing such data effectively ensures transparency, accountability, and efficient government operations [Oliveira *et al.*, 2023].

In this context, researchers have applied advanced techniques, such as NLP and document management systems, to improve government documents' analysis significantly. Pereira *et al.* [2021], for instance, investigate token diversity used to refer to the same services on government websites. They propose a taxonomy to organize terms to enhance standardization and data structure. Constantino *et al.* [2022], in turn, gather government documents, perform text segmentation, employ active learning to build a model, and ultimately achieve semantic classification. Such approaches contribute to the efficiency of textual data analysis, making the processing of government documents more effective.

Researchers have recently leveraged machine learning approaches to automate government document analysis. For instance, Coelho *et al.* [2022] and Brandão *et al.* [2023] address the document classification problem, explicitly focusing on the Brazilian legal and public bidding domains, respectively. With a different task, Hott *et al.* [2023] explore the potential of BERT-based models to generate clusters that effectively encapsulate the underlying topics within bidding data. These efforts represent the growing adoption of advanced computational techniques in government document analysis.

In turn, Monteiro *et al.* [2023] perform a comparative analysis of government Chat Bots from the perspective of Human-Computer Interaction studies. To do this, the author selects several usability metrics of the systems to assess how they communicate efficiently with their users and help solve health, transportation, and travel tasks, among others. Ultimately, the author highlights five main problems that impact the studied systems' usability and subsequent adoption.

Our literature review revealed the importance of detecting potential fraud indicators in public bidding. However, we identified a notable gap in the existing research: the lack of studies specifically addressing the challenges of standardizing input data and comparing statistical parameters between analogous items in depth. Therefore, the main contribution of this work is to propose approaches that effectively address such critical issues. Note that although Oliveira *et al.* [2022] undertake some level of pre-processing on textual bid data, our proposed methodology stands out for its comprehensiveness, which includes the categorization of similar items, as elaborated in Sections 4.2 and 4.3.

3 Dataset

This section introduces the dataset considered in the methodology developed to detect overpricing in public bidding. First, Section 3.1 describes some key aspects of the data, including its primary sources, the time span covered, and its scope. Next, Section 3.2 highlights the exploratory analysis

of bidding items, focusing on the considered attributes of the present work, such as the item description and its nature.

3.1 Data Description

The dataset used in this work contains public bidding data from 853 cities within the state of Minas Gerais, Brazil. In this context, the term *item* refers to both tangible goods (e.g. automobile) and intangible services (e.g. web development). Besides the bidding items, we also considered the goods and services that were waived from public bidding by Federal Law n° 14.133, of April 1, 2021.³ An illustrative instance of a waived item involves the bidding of medicines designated for the treatment of rare diseases, as delineated by the Health Minister (Article 75).

The dataset is composed of publicly available information derived from two primary sources. The public bidding data of cities is sourced from the *Sistema Informatizado de Contas dos Municípios (SICOM)*⁴, a system developed by the *Tribunal de Contas do Estado de Minas Gerais (TCE-MG)*⁵ which acts as a centralized repository for data from all transparency portals representing cities in Minas Gerais. Despite having two reliable data sources, an additional phase of data integration and aggregation is necessary to compute and analyze overpricing. The final dataset contains details about bidding or waived items, including their respective IDs, the year of the exercise, textual descriptions, units of measure, approved unit values, and the nature of expenses.

3.2 Exploratory Data Analysis

This section presents an exploratory analysis of the dataset of public bidding items described in the previous section. Such analyses serve a dual purpose: to foster a comprehensive understanding of the dataset and to assist in shaping the methodology for overpricing detection in bidding processes. The objectives of the analyses presented here include gathering insights into the annual count of bidding items, identifying predominant expense categories, pinpointing frequently occurring terms within item descriptions, and mapping the distribution of description lengths.

The final dataset contains 12,805,984 bidding or waived items at the municipal level and 1,361,523 at the state level. Such items were part of bidding processes conducted between 2014 and 2021 for cities and from 2009 to 2021 for the state level. Figure 1 displays the annual distribution of items throughout this period. Notably, the municipal data shows a drop in items for 2016 and 2020 compared to other years. This may be related to municipal elections occurring in these years. Although public bidding is not prohibited during election periods illegal during elections, electoral regulations impose restrictions on certain conduct by public officials.⁶

We also delve into the nature of bidding items to gain insights into the dataset. Table 1 highlights the top five most

³Lei n° 14.133/21: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14133.htm

⁴SICOM: <https://portalsicom1.tce.mg.gov.br/>

⁵<https://www.transparencia.mg.gov.br/compras-e-patrimonio/compras-e-contratos>

⁶Lei n° 9504/97: https://www.planalto.gov.br/ccivil_03/leis/l9504.htm

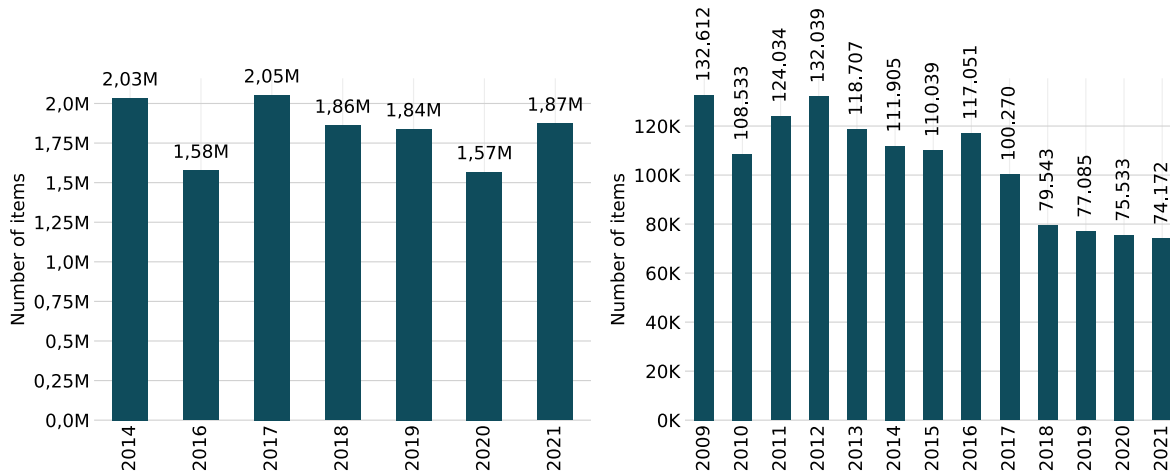


Figure 1. Number of items bidding/waived per year at both municipal (left) and state (right) levels.

Table 1. Top five most frequent expense categories in bidding items.

Municipal		
Category	Items	%
Consumables	5,346,063	41.7%
Unknown	4,327,655	33.8%
Equipment and permanent materials	777,307	6.1%
Other third-party services	683,899	5.3%
Waiver	498,663	3.9%

State		
Category	Items	%
Consumables	1,161,179	85.3%
Services	135,089	9.9%
Permanent materials	58,700	4.3%
Permanent/Consumable	3,410	0.3%
Construction	3,145	0.2%

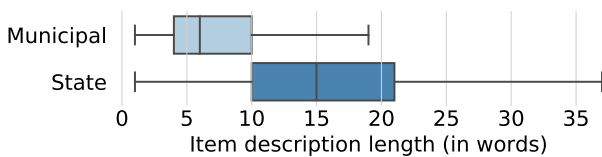


Figure 2. Distribution of item descriptions length in words for municipal and state levels.

frequent expense categories for items in the dataset, categorized by municipal or state level. In both cases, the most common nature category is *Consumables*, corresponding to 41,7% of municipal items and 85,3% of state ones. Despite minor variations in category naming, there are additional nature types such as *Permanent materials* and *Services*. Notably, more than 33,8% of municipal items lack a category (*Unknown*), which means that the nature of the expense could not be identified. This issue might impact the overpricing detection performance as it involves a relatively high number of instances with unknown expense categories.

Finally, the analysis of item descriptions reveals notable distinctions between municipal and state levels. Firstly, we notice a significant difference between the groups concerning description length or the number of words (Figure 2). Overall, municipal items feature shorter descriptions than state items (with a median of 6 words for municipal items

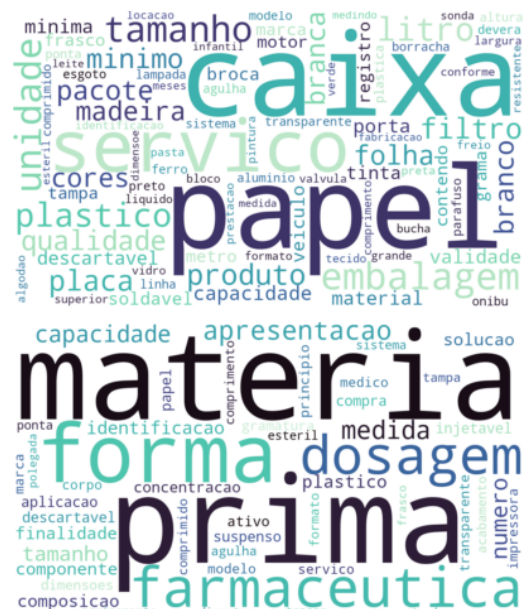


Figure 3. Word clouds of most frequent terms in municipal (left) and state (right) item descriptions.

and 15 words for state items). Furthermore, distinct patterns emerge for the most frequent terms when analyzing the word cloud for both item sets (Figure 3). In municipal items, the predominant terms are associated with consumables, including “caixa” (box), “papel” (paper), “folha” (sheet), “plástico” (plastic), whereas state items are primarily linked to health-related terms, such as “matéria prima” (raw material), “dosagem” (dose), “farmacêutica” (pharmaceutical), “composição” (composition), among others.

However, both item sets contain standard terms in their descriptions that do not significantly contribute to the analyses. These often include units of measurement such as “unidade” (unit) and “grande” (large). Furthermore, specific terms are expressed in two or more variations despite having the same meaning. Consequently, it is imperative to employ textual preprocessing techniques to enhance standardization across the item descriptions. With this pre-processed and standardized dataset, applications that utilize such information may produce more accurate and faithful outcomes aligned with their objectives. The following section describes our pro-

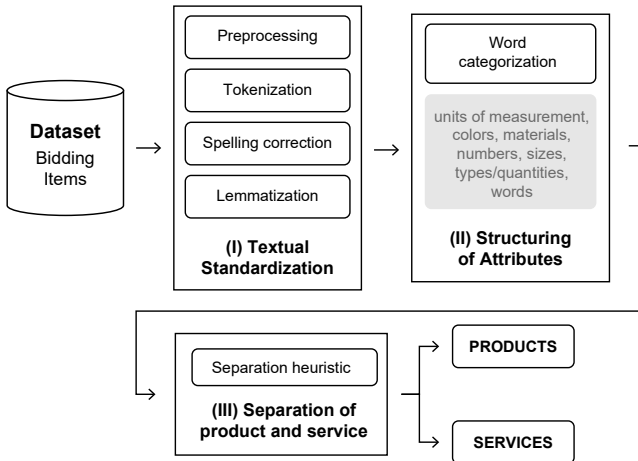


Figure 4. Overview of the item description standardization methodology.

4 Item Description Standardization

To standardize the descriptions of public bidding items, we introduce a textual processing and standardization methodology that contains three phases: (i) textual standardization, (ii) structuring of attributes, and (iii) separation of products and services. Figure 4 illustrates an overview of the methodology, and each step is detailed and exemplified in the subsequent sections.

4.1 Textual Standardization

Within the textual standardization phase, the item description undergoes four essential operations: (i) preprocessing, which involves the removal of special characters, punctuation, and stopwords; (ii) tokenization, (iii) spelling correction, which verifies spelling accuracy and substitutes misspelled words with appropriate corrections, and (iv) lemmatization, which simplifies words to their fundamental forms, known as lemmas, by eliminating inflections, gender variations, numbers, and tense variations. Such operations are fundamental to ensure uniformity in the description of bidding products and services, allowing a more precise comparison of prices.

To illustrate the textual standardization process, consider the item description: “OLEO 20W40 MOTOR A GASOLINA 500ML” (or “OIL 20W40 GASOLINE ENGINE 500ML”). After applying the four operations, the resulting output is a list of pre-processed tokens: “oleo”, “20”, “w”, “40”, “motor”, “gasolina”, “500” e “mililitro” (or “oil”, “20”, “w”, “40”, “engine”, “gasoline”, “500” and “milliliter”). Note that two significant transformations occurred in such a description: (i) the token “a” present in the original description is removed for being considered a stopword, and (ii) the token “ml” was replaced by “mililitro” (milliliter) during the spelling correction.

4.2 Structuring of Attributes

Following the text standardization, the attribute structuring step is fundamental for accurately identifying the characteristics described in the bidding items. In this phase, each token within the item description is assigned to a specific label.

Algorithm 1: Heuristic for distinguishing between products and services.

Input: item description d_i , expense nature nd_i , unit of measure um_i , $expense_nature$, $unit_measure$, $unigrams$ and $bigrams$

Output: classification of the item as **product** or **service**

```

1 begin
2   if  $nd_i \in expense\_nature$  then
3     return service
4   if  $um_i \in unit\_measure$  then
5     return service
6   foreach Word  $p$  in  $d_i$  do
7     if  $p \in unigrams$  then
8       return service
9   foreach bigram  $b$  in  $bigrams$  do
10    if  $b \in d_i$  then
11      return service
12  return product

```

Seven labels are considered in this process: *Unit measures*, *Colors*, *Materials*, *Numbers*, *Sizes*, *Type/Quantity* e *Words*. This attribute structuring enables a more precise description of the bidding item, ultimately leading to a more equitable comparison of prices.

The category *Unit measures* comprises terms such as “liter”, “gram”, and “kilogram”, among others. The label *Colors* covers terms related to colors, such as “blue”, “red”, “light”, and “dark”. The category *Materials* includes terms that describe the materials used in the item’s manufacturing, such as “steel”, “metal”, “wood”, and “plastic”. The label *Numbers*, in turn, encompasses all numeric tokens present in the item description, while *Sizes* covers terms that refer to size variations, such as “small”, “big”, “unique”, among others. The category *Types/Quantities* describes the item’s presentation format and quantity, including “package”, “tablet” and “unit”. Lastly, the *Words* category includes all terms that do not fall into any of the previous categories.

As an example, consider the same description as the previous step, i.e., “OLEO 20W40 MOTOR A GASOLINA 500ML” (or “OIL 20W40 GASOLINE ENGINE 500ML”). In total, three categories were identified: *Unit Measures*, including the tokens “w” and “mililitro”; *Number*, including the tokens “20”, “40” e “500”; and *Words*, including the tokens “oleo”, “motor” e “gasolina”.

4.3 Separation of Products and Services

After the phases of textual standardization and structuring of attributes, the next step in our methodology involves the separation of bidding items into two categories: products and services. Although this information should ideally be indicated in the input data, it is not consistently provided. Such a separation is essential because clustering without distinguishing between products and services can result in an unbalanced estimation of overpricing. Therefore, we developed a heuristic based on keywords to perform this classification automatically. Algorithm 1 describes the proposed heuristic, which takes an item description as input data along with its metadata and four predefined lists: $expense_nature$, $unit_measure$, $un-$

Table 2. Examples of descriptions and keywords for each predefined list.

<i>expense_nature</i>	labor rental, other third-party services, consulting services, construction and installations, services, works
<i>unit_measure</i>	construction, daily, supply, rental, maintenance, work, service provision, procedure, service, show, transportation
<i>unigrams</i>	service, services, provision, rental, contracting, maintenance, construction, installation, consulting, advisory, supply
<i>bigrams</i>	school transport, labor force, artistic show, musical show

igrams, and *bigrams*.⁷

The four lists are generated based on frequent keywords extracted from item descriptions. Table 2 provides a selection of examples from each list. The list *expense_nature* contains descriptions associated with the nature of expenses related to services, while the other lists include keywords relevant to the *Unit Measure* and *Words* categories. Consequently, for each item, the heuristic verifies if its expense nature corresponds to any entry in the *expense_nature* list, or if its unit measure matches any entry in the *unit_measure* list, or if any token of the category *Words* is contained in the lists *unigrams* and *bigrams*.

If any of such conditions are met, the item is categorized as a **service**; otherwise, it is labeled as a **product**. To illustrate the application of the heuristic, consider the same example as presented in previous steps, i.e., “OLEO 20W40 MOTOR A GASOLINA 500ML” (or “OIL 20W40 GASOLINE ENGINE 500ML”). In this case, the nature of the expense for this item is *unknown*, and the unit measure is *unit*, discarding the conditions of lines 2 and 4. Furthermore, no description token is found in the *unigrams* and *bigrams* lists. Therefore, the item is classified as a **product**.

After applying the complete processing and standardization methodology, 368,644 (2,88%) items were ignored due to missing or inconsistent descriptions. The remaining 12,437,340 items were processed and categorized as either products or services. Out of these, 10,558,948 items (84.9%) were classified as **product**, while 1,878,392 items (15.1%) were classified as **service**.

5 Overpricing Detection

Detecting overpricing in public bidding can be modeled as an outlier detection problem. In this context, the goal is to identify values that significantly deviate from the typical behavior of the dataset. Commonly, statistical techniques are employed for outlier detection, which can include methods like regression analysis, analysis of variance, or distance-based approaches. In this work, we introduce a statistical approach that relies on the interquartile range, detailed in section 5.1. Furthermore, we present a comparative analysis in section 5.2, as well as examples of overpricing and anomalies to illustrate the practical application of the proposed methodology and its limitations, as discussed in Section 5.3.

⁷The complete contents of these four lists are available at bit.ly/description_lists.

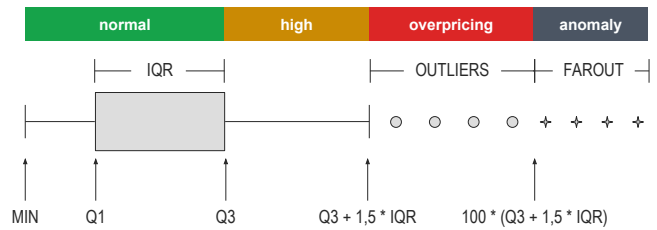


Figure 5. Statistical approach overview.

5.1 Statistical Approach

The interquartile range (IQR) is a robust statistical measure that effectively captures data dispersion concerning its median. This is achieved by calculating the difference between the third and first quartiles of the dataset, resulting in a range that encompasses the central values of the data. Any data point falling outside this range is considered a potential outlier. To establish an upper limit for the price of a specific product or service, we can multiply the IQR by a scaling factor, which can be adjusted according to the sensitivity of the model. Here, this scale factor is set to 1.5. Furthermore, we define not only an upper limit but also a more stringent threshold to distinguish outliers from anomalies by applying a multiplication factor of 100 to the upper limit.

Figure 5 illustrates the definition of four distinct price levels: *normal*, *high*, *overprice*, and *anomaly*. Such levels are defined by empirical data analysis and insights from outlier detection literature. The *normal* level refers to prices that fall below or within the IQR range. The *high* level contains prices that exceed the IQR range but still remain within the upper bound. The level *overpricing* level applies to prices that surpass the upper bound limit. Finally, the *anomaly* concerns prices that go beyond the maximum limit.

It is important to emphasize that the proposed methodology should not be regarded as an absolute solution for detecting overpricing but rather as a means to enhance the efficiency and transparency of public bidding. In other words, when one comes across suspicious prices labeled as *overprice* or *anomaly*, a more in-depth analysis is essential to ascertain whether these prices are justifiable or genuinely indicate fraudulent activities.

5.2 Comparative Analysis

The initial step involves aggregating the data based on the item description to apply the proposed statistical approach for a particular good or service. Different grouping techniques can be employed in this case, including grouping by the original description, the processed description, or structured and grouped by one or more categories derived from the attribute structuring phase. Note that the grouping method chosen significantly affects the quality and reliability of the resulting outcomes. Inadequate aggregation may lead to less dependable price estimates, potentially undermining the ability to detect overpricing or signs of fraud.

To evaluate the effectiveness of the statistical approach and the textual processing and standardization methodology, three distinct experiments are conducted, each varying the method of grouping products and services. In the baseline experiment, the original descriptions of the items are considered without any text processing or standardization. In

Table 3. Comparison of the results for each clustering.

	# groups		# suspicious groups		% overpricing		% anomaly	
	product	service	product	service	product	service	product	service
<i>Original</i>	4,244,671	1,179,081	2,600	548	0.072%	0.036%	0.000%	0.001%
<i>Words</i>	2,069,335	796,475	20,003	4,857	0.579%	0.719%	0.001%	0.004%
<i>FToken</i>	3,365	16,710	2,274	2,011	4.171%	3.544%	0.019%	0.189%

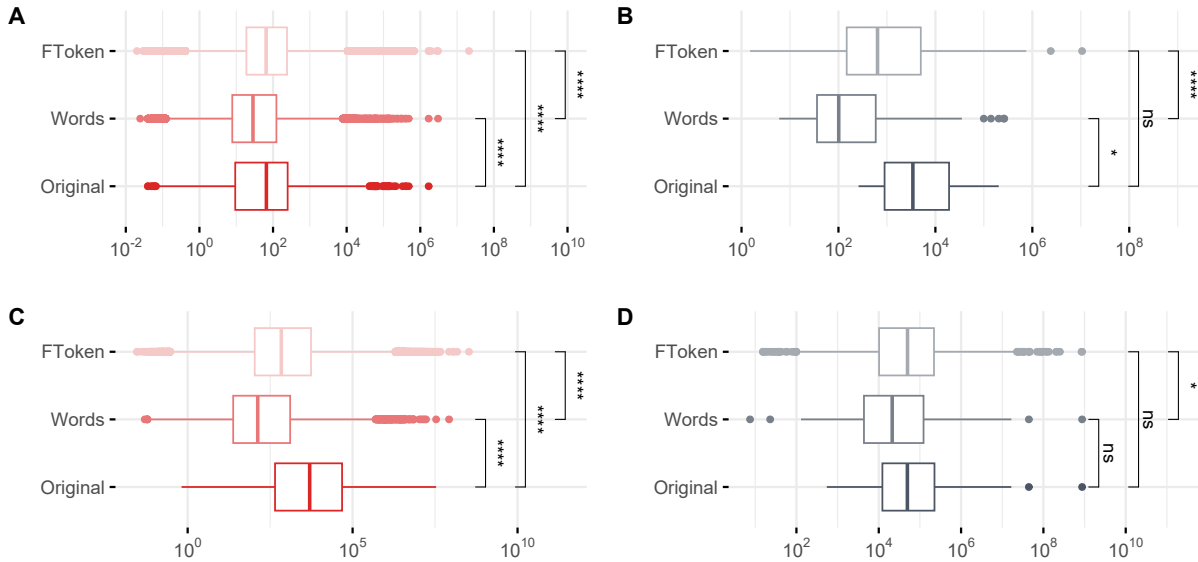


Figure 6. Price distribution for (A-B) products and (C-D) services at overpricing and anomaly levels, respectively. *p*-value levels are symbolized as (1) ns and *: $p > 0.01$, (2) ***: $p < 0.001$, (3) ****: $p < 0.0001$.

the second experiment, the grouping approach combined all the terms from the *Words* category, while in the third experiment, only the first token from this category is considered. In all three scenarios, the district (known as “Comarca” in Portuguese, representing a regional judicial division), the year, and the month of the bidding process are taken into account. Such additional factors, i.e., the location and time period, can have a significant impact on the prices of products and services and are therefore included in the analysis.

The findings presented in Table 3 reveal that the grouping strategy based on the original item descriptions (*Original*) leads to a larger number of groups for both products and services. This result is expected since the original descriptions are typically more detailed and noisy, resulting in various descriptions for the same item. Consequently, this approach generates more specific and smaller groups, leading to a lower rate of overpricing and anomalies compared to other grouping methods. In contrast, the *Words* grouping strategy presents a balance with an intermediate number of groups. However, grouping by just the first token of such a category (*FToken*) results in the fewest number of groups, suggesting a significant loss of information and a tendency to group different items into the same group.

Although the *Words* strategy produced an intermediate number of groups, it significantly outperformed the other two approaches in terms of suspected cases. This suggests that the *Words* strategy can create more representative groups, allowing it to identify a higher number of suspected cases of overpricing and anomalies. Despite having fewer groups compared to the approach based on the original item descriptions (*Original*), the *Words* strategy demonstrated greater ef-

iciency in detecting suspicious cases, boasting a substantially higher rate of both overpricing and anomalies.

Indeed, the results clearly indicate that the *Words* grouping strategy yields a higher rate of anomalies and overpricing compared to the strategy based on the original descriptions (*Original*) but lower than the approach that uses only the first token (*FToken*). This suggests that using all the words within the *Words* category can effectively refine item descriptions, thereby reducing the occurrence of anomalies and overpricing compared to the *FToken* strategy. While there is some loss of information compared to the *FToken*, the *Words* strategy presents a practical balance, making it a valuable intermediate option, particularly when dealing with noisy or less detailed original descriptions.

After analyzing the different grouping strategies, we further evaluated the price distributions within each group to identify substantial differences among them. Figure 6 shows the price distribution for (A-B) products and (C-D) services within the overpricing and anomaly categories, respectively. The results indicate significant variations in price distributions depending on the grouping strategy, regardless of the item’s nature. Notably, the *Words* method shows lower average prices in comparison to the other grouping strategies. Such a result can be attributed to the possibility of information loss and the subsequent reduction of data variance that this grouping strategy might introduce.

Regarding the anomaly level, the price distribution is similar in most of the cases, indicating that the presence of anomalies is not significantly affected by the grouping technique employed. However, it is essential to emphasize that for anomalies, the statistical and textual standardization method-

Table 4. Top three products and services classified as overpricing. The terms used in the *W* and *FT* strategies are underlined.

	Product			Service		
	Original Description	Price (R\$)	Threshold (R\$)	Original Description	Price (R\$)	Threshold (R\$)
<i>O</i>	TNT GREEN TYPE FABRIC	485,000.0	59,780.4	ASPHALT REPAIR SERVICES...	871,685,340.0	8,010,707.4
	HOSPITAL EQUIPMENT	392,000.0	41,600.0	EVENTS	44,348,000.0	217,000.0
	DIESEL OIL S10 4892018 447337	1,688,830.8	1,378,620.0	OTHER THIRD-PARTY SERVICES...	34,000,000.0	10,500,000.0
<i>W</i>	<u>PARTS</u>	3,036,463.9	199,000.0	<u>REPAIR OF ASPHALT REPAIR SERVICES...</u>	871,685,340.0	8,010,707.4
	<u>TNT GREEN TYPE FABRIC</u>	485,000.0	59,780.4	<u>EVENTS</u>	44,348,000.0	217,000.0
	<u>HOSPITAL EQUIPMENT</u>	392,000.0	41,600.0	33903923 <u>OTHER THIRD-PARTY SERVICES...</u>	34,000,000.0	10,500,000.0
<i>FT</i>	BOOK LITERARY WORK FOR SCHOOL KIT COMPOSITION 3...	202,696.3	49.6	<u>ASPHALT REPAIR SERVICES...</u>	871,685,340.0	1,514,915.3
	GRANT OF REAL RIGHT OF USE REFERENCE 01 (ONE) LAND...	21,000,000.0	2,250,000.0	<u>HIRING OF COMPANY TO CONTINUE THE WORK...</u>	841,365,170.0	32,224.0
	<u>LABORATORY TESTS GENERIC ITEM-378218</u>	10,643,942.0	152.5	<u>HIRING OF COMPANY</u>	336,330,975.6	4,591,747.5

O = Original *W* = Words *FT* = FToken

Table 5. Top three products and services classified as anomalies. The terms used in the *W* and *FT* strategies are underlined.

	Product			Service		
	Original Description	Price (R\$)	Threshold (R\$)	Original Description	Price (R\$)	Threshold (R\$)
<i>O</i>	LITERARY WORK BOOK FOR SCHOOL KIT COMPOSITION 3...	202,696.3	4,957.5	ASPHALT REPAIR SERVICES...	871,685,340.0	801,070,738.5
	LITERARY WORK BOOK FOR SCHOOL KIT COMPOSITION 3 TO 5...	8,722.0	2,667.0	EVENTS	44,348,000.0	21,700,001.5
	BRONZE SCREW FOR THE CROWN	1,349.8	944.5	HIRING OF COMPANY	16,626,939.6	100.0
<i>W</i>	<u>LITERARY WORK BOOK FOR SCHOOL KIT COMPOSITION 2...</u>	263,712.1	3,577.5	<u>ASPHALT REPAIR SERVICES...</u>	871,685,340.0	801,070,738.5
	<u>LITERARY WORK BOOK FOR SCHOOL KIT COMPOSITION 3...</u>	256,230.4	3,577.5	<u>EVENTS</u>	44,348,000.0	21,700,001.5
	<u>UNLEADED GASOLINE</u>	99,500.0	396.0	<u>HIRING OF COMPANY</u>	16,626,939.6	100.0
<i>FT</i>	<u>LABORATORY TESTS GENERIC ITEM-378218</u>	10,643,942.0	15,250.0	<u>HIRING OF COMPANY TO CONTINUE THE WORK...</u>	841,365,170.0	3,222,400.0
	CRANKSHAFT 366	2,398,025.0	94,868.5	<u>ASPHALT REPAIR SERVICES...</u>	871,685,340.0	151,491,532.5
	PARTICLE ACCELERATOR TUBE, REF. 11303910, FOR ACCELERATOR...	747,957.6	8,302.5	<u>SERVICE - IMPLEMENTATION OF THE PLATFORM...</u>	214,692,000.0	20,050.0

O = Original *W* = Words *FT* = FToken

ologies effectively identify and address these outliers appropriately, ensuring the reliability of the results obtained. Therefore, the overall result analyses suggest that both proposed approaches efficiently identify evidence of overpricing and anomalies in both product and service items. Furthermore, our findings imply that the grouping approach based on *Words* may be a viable option for detecting possible irregularities in public bidding, as it effectively balances the detection of suspicious cases and the number of groups.

5.3 Examples of overpricing and anomaly

For each grouping strategy, we provide examples of the top three products and services classified as overpricing and anomalies. These represent the three items with the most significant discrepancies between the registered price and the overpricing/anomaly threshold, as explained in Section 5.1). Tables 4 and 5 present in detail such results. Note that items classified as overpricing exhibit values significantly higher than the threshold, potentially indicating irregular pricing. However, it is important to emphasize that grouping results may differ among the different approaches, underscoring the importance of a thorough evaluation of these results.

In the case of items classified as anomalies, there is a noticeable and substantial difference between the registered price and the anomaly threshold across all instances. This

suggests the potential presence of typing errors in the item description or pricing of the product or service. Moreover, it is common to find unclear descriptions, as exemplified by entries such as “EVENTS” or “HIRING OF COMPANY”, or written in an unusual way, as in the case of “BOOK LITERARY WORK FOR SCHOOL KIT COMPOSITION 3...”, which may be related to issues during the item registration process. Such results highlight the importance of preprocessing and standardizing item descriptions, as well as the necessity for a thorough evaluation of identified anomalies.

Threats to validity. Approaches dealing with text processing face numerous challenges due to the lack of standardization, which can significantly impact the quality of results. An identified concern in the proposed heuristic is that, in some instances, discrepancies may arise from issues within the separation heuristic. For example, consider the item “LABORATORY TESTS GENERIC ITEM”. The description suggests it is a laboratory **service**, yet the separation heuristic classifies it as a **product**. This misclassification could lead to an inaccurate estimation of the anomaly threshold, resulting in false positives. However, classification errors are an expected part of a heuristic approach, and they should be taken into account when interpreting the results.

Another identified threat is the presence of significant discrepancies in magnitude between the threshold and item

Table 6. The five types of fuel in the ANP dataset with some basic data descriptions.

Products	# Cities	Period	Avg. Price (R\$)	Max Price (R\$)
DIESEL OIL	67	01/2013 to 09/2023	3.603	8.890
DIESEL OIL (S10)	67	01/2013 to 09/2023	3.728	8.990
COMMON GASOLINE	67	01/2013 to 09/2023	4.300	8.499
GASOLINE WITH ADDITIVES	58	10/2020 to 09/2023	6.176	8.899
HYDROUS ETHANOL	67	01/2013 to 09/2023	3.034	7.450

prices, along with confusing and unconventional patterns in item descriptions. Such irregularities can potentially impact the accuracy of overpricing detection and may be attributed to registration issues and the chosen grouping approach. Hence, for an actual application of overpricing analysis, it is essential to address these cases separately by either removing erroneous items or refining the grouping approach.

6 Experimental Evaluation

In this section, we evaluate both proposed approaches (i.e., the statistical approach and the textual standardization methodology) using a ground-truth dataset. We describe the dataset considered in Section 6.1, as well as the preprocessing and integration process performed to prepare the data for our experimental evaluation. Next, in Sections 6.2 and 6.3, we assess the performance of each approach using quantitative and qualitative analyses, respectively.

6.1 Data

The ground-truth dataset used for evaluation originates from the *Agência Nacional do Petróleo, Gás Natural e Biocombustíveis do Brasil* (ANP)⁸. The ANP is the regulatory authority for activities within the oil, natural gas, and biofuels industries in Brazil. Such data source contains electronic spreadsheets that contain historical price research series on a weekly and monthly basis, categorized by geographic coverage. It provides information covering various types of fuels, including C gasoline, hydrated ethanol, B diesel oil, and B S-10 diesel oil.

In our specific focus on the State of Minas Gerais, the ANP provides fuel price data from 67 cities. Within this dataset, the maximum selling price for each city is provided. Such maximum price serves as a benchmark to establish the presence of overpricing in public fuel bidding. Table 6 describes five distinct fuel types we utilized for our evaluation, alongside fundamental data descriptions. Each dataset entry includes information about the product (fuel type), city, year, month, average price, and maximum price.

Data Preprocessing. The ANP’s collection of electronic spreadsheets is organized yearly, with one file available for each pair of years starting from 2013. The data was filtered to exclusively retain cities located in Minas Gerais, given that our bidding data exclusively pertains to this state. Finally,

Table 7. Number of products resulting from data integration for each grouping strategy.

Grouping	Unfiltered	Filtered
<i>Original</i>	9,975,109	5,795 (0.06%)
<i>Words</i>	9,975,109	6,023 (0.06%)
<i>FToken</i>	9,975,109	3,091 (0.03%)

we formatted the period column to divide it into separate year and month columns, thus facilitating further analysis.

Data Integration. To evaluate the effectiveness of our proposed approaches, we merged our public bidding dataset with the ground-truth one, using as key attributes the period, city, and the same product description on each grouping (*Original*, *FToken*, and *Words*). Our data integration involved several steps to ensure a relevant match between the datasets. First, we exclusively select products from the public bidding dataset, as the ANP dataset focuses solely on fuels, a subset of products. Next, within the public bidding dataset, we specifically filter products containing key terms from the ANP dataset, which primarily included “Diesel”, “Gasoline” and “Ethanol”.

Finally, we employ fuzzy string matching to establish correspondence between the products in the public purchase dataset and the fuels listed in the ANP dataset. To ensure a robust match, we set a stringent similarity threshold of at least 90%. We use the Python library *RapidFuzz*⁹ for the fuzzy string matching process. The matching procedure resulted in a subset of public bidding data and their corresponding real maximum reference prices from the ANP dataset. Table 7 shows the number of products resulting from data integration for each grouping strategy.

6.2 Quantitative Analysis

To assess the performance of both proposed approaches, we conducted a comparative analysis between the maximum reference prices provided by the ANP and the overpricing thresholds resulting from the three distinct groupings. Here, our evaluation focuses on the overpricing threshold, as the ANP dataset does not provide references for anomalies. Furthermore, to ensure robust and reliable results, we applied a filter to exclude groups with a size smaller than ten. We consider metrics commonly employed in evaluating prediction or estimation models, including R^2 Score, Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).

Table 8 shows the evaluation metrics results for each type of grouping. Overall, all grouping strategies performed well in terms of overpricing estimation, underscoring the effectiveness of the IQR-based statistical approach. Still, the

⁸ANP: <https://www.gov.br/anp/pt-br/assuntos/precos-e-defesa-da-concorrenca/precos/precos-revenda-e-de-distribuicao-combustiveis/serie-historica-do-levantamento-de-precos>

⁹RapidFuzz: <https://maxbachmann.github.io/RapidFuzz/>

Table 8. Evaluation metrics results for each type of grouping.

Grouping	R ² Score	MAE	MAPE
<i>Original</i>	80%	0.17	4%
<i>Words</i>	95%	0.20	4%
<i>FToken</i>	93%	0.31	6%

Words strategy presents the highest R² Score and a MAPE equal to the *Original*, indicating a slightly superior fit to the ANP data. Such a result suggests that our proposed *Words* strategy, which minimizes information loss while efficiently estimating overpricing, may be a promising approach for identifying potential irregularities in public bidding.

We also performed a visual analysis to compare the overpricing thresholds generated by the three grouping strategies with the threshold derived from the ANP dataset. In Figure 7A, we compare the distributions of both thresholds for each grouping strategy. We apply the Wilcoxon statistical test to verify whether there are significant differences in the threshold means among the different grouping strategies. The results indicate that there is little or no statistical difference between the means of the overpricing thresholds, regardless of the grouping strategy, reinforcing the consistency and robustness of the proposed statistical approach.

Figure 7B, on the other hand, illustrates the correlation between the overpricing thresholds derived from the IQR approach and those provided by the ANP. Here, we apply the Spearman correlation coefficient to evaluate the strength and direction of the association between these two sets of thresholds. Overall, all three grouping strategies show a strong positive correlation, indicating that our approach consistently estimates overpricing at a similar level to the ANP data. Regarding the performance of different grouping strategies, the *FToken* strategy stands out with the highest correlation coefficient, followed by the *Words* strategy.

In summary, our findings highlight interesting trade-offs between the different grouping strategies. The *FToken* strategy demonstrates a robust systematic relationship with the ANP's thresholds, as evident from its high correlation coefficient. However, it exhibits bias in predicting real price differences, resulting in a high MAPE. Such a bias could be attributed to the inherent loss of detailed information in this strategy, causing over- or underestimation of actual values by a relatively high percentage. While it effectively captures overarching trends, refinements are needed to enhance the precision of its predictions and bring them closer to actual values.

Conversely, despite being the noisier strategy, the *Original* excels in terms of MAPE as it closely approximates actual prices in percentage error terms. Although it does not consistently follow the same patterns as the ANP dataset (low correlation) or explain much of the variability in the reference thresholds (low R² Score), its predictions are, on average, entirely accurate in terms of percentage error. Such accuracy suggests that the *Original* strategy captures local nuances in the data, contributing to precise individual price estimates. However, this focus on localized details may cause deviations from the broader trends and patterns in the ANP data, resulting in lower correlation and R² Score.

Finally, the *Words* strategy accurately estimates overpricing

levels with a strong alignment to the ANP's data. Considering a wide variety of terms within item descriptions successfully captures the granular details and broader patterns in the data. Its balanced approach strikes a good compromise between capturing local details and overarching trends. Such balance likely contributes to its high correlation, R² Score, and low MAPE, demonstrating its excellence in accurately identifying potential irregularities in public bidding.

6.3 Qualitative Analysis

In addition to the quantitative analysis in the previous section, we conducted a qualitative assessment of the proposed approaches. To do so, we investigate the textual descriptions within each grouping strategy corresponding to each ANP product. Unlike the previous analysis, we included all groups without applying a filter based on size, focusing on the textual content of descriptions for a comprehensive evaluation of the methods. Table 9 shows the total number of unique descriptions within each grouping strategy for every ANP product, along with illustrative examples.

According to the textual analysis, the *Original* strategy yielded the most significant number of descriptions for each of the five ANP products. While this strategy successfully captures the granular nuances in item descriptions, it tends to group related items less effectively. Different textual descriptions of the same product are treated as distinct, which can be challenging when estimating overpricing thresholds. For example, regarding the product *DIESEL OIL (S10)*, at least four descriptions are written in slightly different ways referring to the same product.

Conversely, the *FToken* strategy is the most restrictive, focusing exclusively on the first token from the *Words* category. Such a restriction leads to a significantly smaller number of distinct descriptions for each product. This strategy is only associated with three of the five ANP products. While its simplicity results in fewer unique descriptions, it also restricts the strategy's applicability. In cases where the descriptions significantly deviate from the first token, this strategy may not effectively group related items.

Finally, the *Words* strategy balances specificity and generality. As it undergoes textual standardization, it adeptly aligns related items more effectively than the *Original* strategy. The distinct descriptions for each product fall between the extremes of the *Original* and *FToken* approaches, indicating that this strategy combines local details and broader trends. Therefore, the *Words* strategy may be the most promising option in overpricing detection, as it standardizes descriptions and groups related items more effectively.

7 Conclusion

This work proposed two approaches to detect overpricing in public bidding for products and services: a methodology for processing and standardizing bidding items and a statistical approach based on the interquartile range for identifying overpricing. The textual standardization methodology includes a heuristic to classify items into either products or services, as well as description standardization through re-

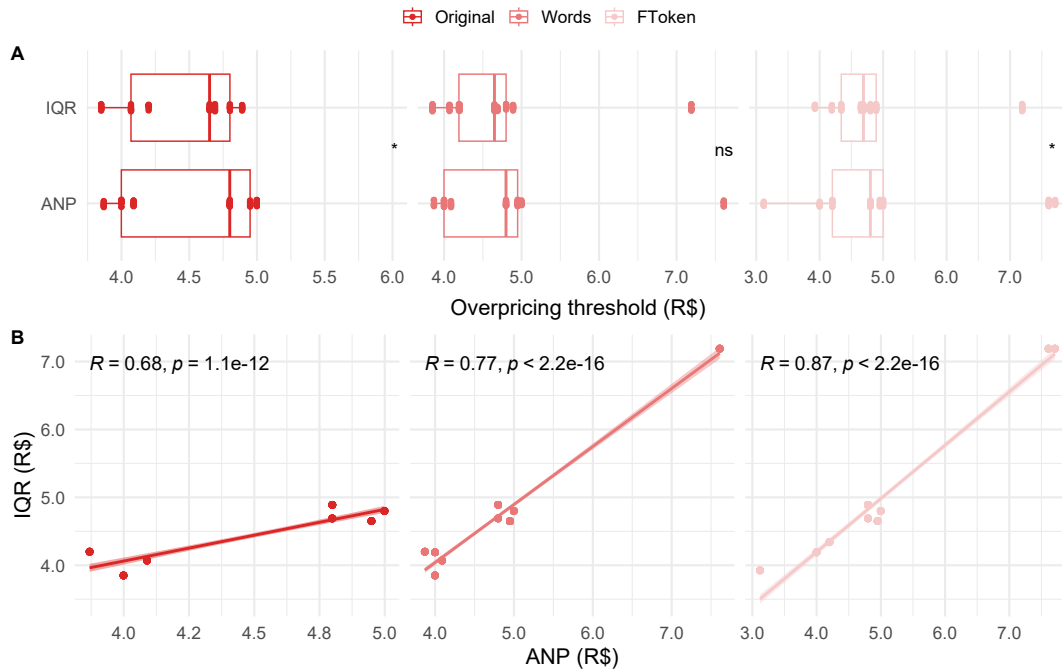


Figure 7. Visual comparative analysis between the overpricing thresholds for all three grouping strategies. (A) Overpricing threshold distribution comparison, using the Wilcoxon statistical test. (B) Correlation plot between both overpricing thresholds, using the Spearman correlation coefficient.

Table 9. Number of distinct descriptions in each grouping strategy for each ANP product, along with some examples.

Product	Original		Words		FToken	
	#	Examples	#	Examples	#	Examples
DIESEL OIL	153	“DIESEL”, “DIESEL OIL”, “COMMON DIESEL OIL”, “FUEL - DIESEL OIL”, “AUTOMOTIVE DIESEL OIL”	75	“diesel”, “diesel oil”, “diesel engine oil”	1	“diesel”
DIESEL OIL (S10)	259	“DIESEL OIL S10”, “COMMON S-10 DIESEL OIL”, “DIESEL OIL S-10”, “DIESEL FUEL OIL S-10”	127	“diesel s10”, “diesel oil s10”, “diesel oil s10 added”, “biodiesel oil s10”	0	-
COMMON GASOLINE	211	“GASOLINE”, “COMMON GASOLINE”, “COMMON GASOLINE TYPE C”, “COMMON AUTOMOTIVE GASOLINE”	59	“common gasoline”, “gasoline”, “common fuel gasoline”, “common gasoline without additives”	1	“gasoline”
GASOLINE WITH ADDITIVES	42	“GASOLINE WITH ADDITIVES”, “GASOLINE WITH ADDITIVES.”, “GASOLINE WITH ADDITIVES GASOLINE WITH ADDITIVISE”	31	“gasoline with additives”, “common gasoline with additives”, “common gasoline with additives”	0	-
HYDROUS ETHANOL	27	”ETHANOL”, “HYDROUS ETHANOL”, “ETHANOL ETHANOL (FUEL)”, “REGULAR HYDROUS ETHANOL”	4	“ethanol”, “ethanol hydrate”	1	“ethanol”

moving special characters and stopwords and creating structured attributes.

Three experiments were conducted on real-world public bidding data to evaluate both approaches. Such experiments involved grouping items based on different strategies: (i) grouping with the original, untreated item descriptions, (ii) grouping using all terms from the *Words* structure category, and (iii) grouping using only the first token from the *Words* category. Our results revealed that the *Words* strategy, characterized by minimal information loss and lower prices, reduced data variance and successfully identified a substantial portion of overpricing instances.

Furthermore, we assessed both of our proposed methodologies—the statistical approach and the textual standardization process—utilizing a ground-truth dataset with a dual focus on quantitative and qualitative analyses. In the quantitative analysis, we compared the maximum reference prices established by the ground-truth data and the overpricing thresholds determined via the three different grouping strategies. Simultaneously, our qualitative evaluation investigated the textual descriptions within

each grouping strategy, aligning them with their respective ground-truth products. Overall, the *Words* strategy proved to be the most effective in aligning with the ground-truth data, offering a better balance between sensitivity and specificity, making it a suitable choice for real-world applications.

While our work presents promising strategies for detecting overpricing and anomalies in public bidding, we acknowledge certain limitations in our approach. First, future research could explore alternative strategies for identifying groups, such as applying clustering algorithms and their integration with word embeddings. This could enhance identifying outliers within content groups, providing a more nuanced understanding of irregularities. Additionally, our methodology’s application to real-world scenarios warrants further exploration, with a particular emphasis on considering temporal and geographical variations in prices. Assessing the impact of these factors on the effectiveness of the IQR technique would contribute to a more comprehensive understanding of the practical implications of our proposed approach.

Finally, although there are challenges, such as the lack of standardization in the descriptions of the bid items, our

findings demonstrate that the proposed approaches can help identify possible irregularities. However, it is essential to highlight that these methods are not a definitive solution, requiring continuous monitoring and improvement of the techniques used. Future efforts involve submitting the flagged bids as suspicious cases for evaluation by domain experts, offering a more rigorous validation of the proposed statistical approach. This iterative process will help refine and enhance the techniques, contributing to more effective overpricing detection in public bidding.

Declarations

Acknowledgements

The authors thank this work's collaborator, Henrique R. Hott.

Funding

This work was funded by the Prosecution Service of the State of Minas Gerais (in Portuguese, *Ministério Público do Estado de Minas Gerais*, or simply MPMG) through the Analytical Capabilities Project (in Portuguese, *Programa de Capacidades Analíticas*) and by CNPq, CAPES, and FAPEMIG.

Competing interests

The authors declare that they have no competing interests.

References

- Brandão, M. A. *et al.* (2023). Plus: A semi-automated pipeline for fraud detection in public bids. *Digit. Gov.: Res. Pract.* Just Accepted. DOI: 10.1145/3616396.
- Brandão, M. A. *et al.* (2023). Impacto do pré-processamento e representação textual na classificação de documentos de licitações. In *Proceedings of the 38th Brazilian Symposium on Databases, SBBD 2023, Belo Horizonte, MG, Brazil, September 25-29, 2023*, pages 102–114. SBC. DOI: 10.5753/sbbd.2023.231658.
- Brasil (2018). Law no. 13.709 of august 14, 2018: Brazilian general data protection law (lgpd). http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/113709.htm Accessed: 16 January 2024.
- Coelho, G. M. C. *et al.* (2022). Text classification in the brazilian legal domain. In Filipe, J., Smialek, M., Brodsky, A., and Hammoudi, S., editors, *Proceedings of the 24th International Conference on Enterprise Information Systems, ICEIS 2022, Online Streaming, April 25-27, 2022, Volume 1*, pages 355–363. SCITEPRESS. DOI: 10.5220/0011062000003179.
- Constantino, K. *et al.* (2022). Segmentação e classificação semântica de trechos de diários oficiais usando aprendizado ativo. In *SBBD*, pages 304–316, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2022.224656.
- Correa, M. A. O. S. and Leal, A. G. (2018). Identification of overpricing in the purchase of medication by the federal government of brazil, using text mining and clustering based on ontology. In *Proceedings of the 2018 2nd International Conference on Cloud and Big Data Computing, ICCBDC 2018, Barcelona, Spain, August 03-05, 2018*, pages 66–70. ACM. DOI: 10.1145/3264560.3264569.
- Costa, L. L. *et al.* (2022). Alertas de fraude em licitações: Uma abordagem baseada em redes sociais. In *Proceedings of the 11th Brazilian Workshop on Social Network Analysis and Mining*, pages 37–48, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/brasnam.2022.223175.
- Gabardo, A. C. and Lopes, H. S. (2014). Using social network analysis to unveil cartels in public bids. In *2014 European Network Intelligence Conference*, pages 17–21. DOI: 10.1109/ENIC.2014.11.
- Green, R., Zimmerer, T., and Steadman, M. (1994). The role of buyer sophistication in competitive bidding. *Journal of Business & Industrial Marketing*, 9:51–59. DOI: 10.1108/08858629410053489.
- Hott, H. R. *et al.* (2023). Evaluating contextualized embeddings for topic modeling in public bidding domain. In Naldi, M. C. and Bianchi, R. A. C., editors, *Intelligent Systems - 12th Brazilian Conference, BRACIS 2023, Belo Horizonte, Brazil, September 25-29, 2023, Proceedings, Part III*, volume 14197 of *Lecture Notes in Computer Science*, pages 410–426. Springer. DOI: 10.1007/978-3-031-45392-2_27.
- Lima, M. C. *et al.* (2020). Inferring about fraudulent collusion risk on brazilian public works contracts in official texts using a bi-lstm approach. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1580–1588. Association for Computational Linguistics. DOI: 10.18653/V1/2020.FINDINGS-EMNLP.143.
- Luna, R. and Figueiredo, D. (2022). Caracterização das licitações públicas no estado do rio de janeiro: Diversidade, licitantes Únicos e redes. In *WCGE*, pages 145–156, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/wge.2022.222675.
- Matschak, T., Trang, S., and Prinz, C. (2022). A taxonomy of machine learning-based fraud detection systems. In Beck, R., Petcu, D., Fotache, M., Matook, S., Helms, R., Wiener, M., Rusu, L., and Tuunanen, T., editors, *30th European Conference on Information Systems - New Horizons in Digitally United Societies, ECIS 2022, Timisoara, Romania, June 18-24, 2022*.
- Monteiro, M. d. S., Batista, G. O. d. S., and Salgado, L. C. d. C. (2023). Investigating usability pitfalls in brazilian and foreign governmental chatbots. *Journal on Interactive Systems*, 14(1):331–340. DOI: 10.5753/jis.2023.3104.
- Oliveira, G. P. *et al.* (2022). Detecting inconsistencies in public bids: An automated and data-based approach. In *WebMedia*, pages 193–201, Porto Alegre, RS, Brasil. SBC. DOI: 10.1145/3539637.3558230.
- Oliveira, G. P. *et al.* (2023). Assessing data quality inconsistencies in brazilian governmental data. *Journal of Information and Data Management*, 14(1). DOI: 10.5753/jidm.2023.3220.
- Pereira, A. *et al.* (2022). Usando redes complexas na identi-

- cação de empresas fraudulentas em licitações públicas. In *WCGE*, pages 13–24, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/wcge.2022.222704.
- Pereira, G. *et al.* (2021). Classificação taxonômica de categorias de serviços públicos para aplicações digitais. In *WCGE*, pages 119–130, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/wcge.2021.15982.
- Reis, V. Q. d., Rabello, M. E. R., Lima, A. C., Jardim, G. P. S., Fernandes, E. R., and Brefeld, U. (2023). Data practices in apps from brazil: What do privacy policies inform us about? *Journal on Interactive Systems*, 14(1):1–8. DOI: 10.5753/jis.2023.2954.
- Silva, M. *et al.* (2023). Análise de sobrepreço em itens de licitações públicas. In *Anais do XI Workshop de Computação Aplicada em Governo Eletrônico*, pages 118–129, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/wcge.2023.230608.
- Silva, M. O. *et al.* (2022). Lipset: Um conjunto de dados com documentos rotulados de licitações públicas. In *SBB DSW*, pages 13–24, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/dsw.2022.224925.
- Xiao, Z. and Jiao, J. (2021). Explainable fraud detection for few labeled time series data. *Secur. Commun. Networks*, 2021:9941464:1–9941464:9. DOI: 10.1155/2021/9941464.